From Theoretical Promise to Observational Reality: Calibration, Foreground Subtraction, and Signal Extraction in Hydrogen Cosmology

by

Adrian Chi-Yan Liu

A.B., Princeton University (2006)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author .....
Department of Physics
May 25, 2012

Certified by .....
Max Erik Tegmark
Professor of Physics
Thesis Supervisor

Accepted by .....
Krishna Rajagopal
Professor of Physics
Associate Department Head for Education

From Theoretical Promise to Observational Reality: Calibration, Foreground Subtraction, and Signal Extraction in Hydrogen Cosmology

by

Adrian Chi-Yan Liu

Submitted to the Department of Physics
on May 25, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

By using the hyperfine 21 cm transition to map out the distribution of neutral hydrogen at high redshifts, hydrogen cosmology has the potential to place exquisite constraints on fundamental cosmological parameters, as well as to provide direct observations of our Universe prior to the formation of the first luminous objects. However, this theoretical promise has yet to become observational reality. Chief amongst the observational obstacles are a need for extremely well-calibrated instruments and methods for dealing with foreground contaminants such as Galactic synchrotron radiation.

In this thesis we explore a number of these challenges by proposing and testing a variety of techniques for calibration, foreground subtraction, and signal extraction in hydrogen cosmology. For tomographic hydrogen cosmology experiments, we explore a calibration algorithm known as redundant baseline calibration, extending treatments found in the existing literature to include rigorous calculations of uncertainties and extensions to not-quite-redundant baselines. We use a principal component analysis to model foregrounds, and take advantage of the resulting sparseness of foreground spectra to propose various foreground subtraction algorithms. These include fitting low-order polynomials to spectra (either in image space or Fourier space) and inverse variance weighting. The latter method is described in a unified mathematical framework that includes power spectrum estimation. Foreground subtraction is also explored in the context of global signal experiments, and data analysis methods that incorporate angular information are presented. Finally, we apply many of the aforementioned methods to data from the Murchison Widefield Array, placing an upper limit on the Epoch of Reionization power spectrum at redshift $z = 9.1$.

Thesis Supervisor: Max Erik Tegmark
Title: Professor of Physics

Acknowledgments

What a thrilling ride this has been! To be in a position to experience all this has been a wonderful privilege, and the delightful company along the way has made everything even better. It goes without saying, then, that I have a lot of people to thank[1].

I'd like to start by thanking my advisor, Max Tegmark. Max was one of the main reasons I came to MIT, and I could not have made a better choice in deciding to work for him. Working for a world-class scientific mind has been eye-opening to say the least. But in addition, Max has been—to a crude approximation—an infinitely wise advisor on pretty much anything. Our conversations on topics as wide-ranging as economics, politics, careers, and the state of the human species have been fascinating, and I am forever grateful to him for the experience.

I'd also like to thank my other committee members, Jacqueline Hewitt and Scott Hughes. While she has probably forgotten about this by now, Jackie was the first person to explain radio interferometry to me in a way that I was able to understand. Since then we've collaborated on a number of other projects, and I have also benefitted from her thoughtful advice. Scott has been a great person to go to for advice, and it has been a real privilege taking his classes, teaching with him, and doing service work with him. All that we have left to do is to write a paper together, and I look forward to the prospect! (Primordial gravitational wave background signatures in hydrogen cosmology, perhaps?)

One of the wonderful things about the astrophysics division at MIT is the opportunity to interact with other faculty members on an informal basis. I thank Deepto Chakrabarty for his wise advice on navigating graduate school and beyond, Saul Rappaport for his "simple" questions, Rob Simcoe for our fun, informal interactions, and both Paul Schechter and Ed Bertschinger for pushing all of us to do the best work possible.

Research in astrophysics and cosmology is naturally a collaborative venture, and I thank my collaborators. Judd Bowman, Matias Zaldarriaga, Jonathan Pritchard, and Avi Loeb had a particularly strong influence on me. I would also like to thank my "academic siblings"—the older ones (Molly Swanson, Yi Mao, Mark Hertzberg, and Courtney Peterson) for acting as role models that I could look up to, a "twin" (Andy Lutomirski) for using his razor-sharp intellect to make me a better physicist, and the younger ones (Josh Dillon and Jeff Zheng) for forcing me to "grow up" and

[1]This is not meant to be an exhaustive list. Naturally there are many people who should be here as well, but did not fit in the page and a half. You know who you are, and I hope you realize what a difference you made.

be more of a leader. It has also been an enormous honor to work with our bright UROPs, Michael Valdez, Nevada Sanchez, Jon Losh, Scott Morrison, Katelin Schutz, Hrant Gharibyan, Ioana Zelko, and Victor Buza.

Thank you, also, to all my friends. Phil Zukin for experiencing the ups and downs of being an aspiring cosmologist since we met during the various grad school open houses. Chris Williams for being a fantastic apartment-mate as well as a brilliant research collaborator. Leo Stein for his greetings and phonetic hiatuses. Leslie Rogers for showing us all how a truly nice person behaves. Sarah Vigeland for geeking out with me over the West Wing. Mike Matejek for the Princeton pride. Nicolas Smith for his humor. Beyond the astrograd crowd, I would like to additionally thank Jan Hsi Lui for the filet o' fishes over the years, Weston Powell for his hahas, and Angela Wu for her wisdom. I would also like to thank Emily Berger for believing in me (and for providing strawberry ice cream at the appropriate critical juncture), Diann Benti for her incredible support and her making me more aware of the world beyond the ivory tower, and Kaitlin Reilly for completely understanding me.

Finally, I would like to express my deep, deep gratitude for my family. My mother taught me my compassion, my father inspired my interest in science, and my sister has been a wonderful partner in crime. This paragraph is disproportionately short because there is only one thing that I have to thank them for—everything.

"Don't promise too much, Mr. Theorist, don't promise too much!" snapped the detective. "You may find it a harder matter than you think."
—Detective Athelney Jones, in "The Sign of the Four" by Sir Arthur Conan Doyle

"Technology of that kinda can be worth an untold fortune. Imagine being able to control a radio device simply by sending a command via radio waves"
—Sherlock Holmes, in the 2009 movie "Sherlock Holmes".

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Theoretical Promise of Hydrogen Cosmology

In recent decades, cosmology has emerged as a mature experimental science. A wide range of observational probes, including the cosmic microwave background, galaxy surveys, gravitational lensing, supernova studies, cluster counts, chemical abundance studies, and others have complemented theoretical models to provide a concordance ΛCDM model of our Universe (Peebles, 2012; Reid et al., 2010; Komatsu et al., 2011). In this model, primordial fluctuations in the density field were amplified by gravitational clustering to produce the rich hierarchy of structures that we see today, all in an expanding space that switched from decelerated to accelerated expansion when our Universe was about half its current age.

Much of the success of observational cosmology comes from measurements of the cosmic microwave background (the CMB) and galaxy surveys. Together, these probes have (for instance) provided measurements of spatial curvature to within 1% and provided independent confirmation of dark energy, complementing measurements of supernova recession velocities (Tegmark et al., 2006). Remarkably, the CMB and galaxy surveys are now sensitive enough to provide limits on fundamental physics that have nothing to do with astrophysics. For instance, our best-fit models of the CMB now weakly prefer a non-zero neutrino mass, although the data have yet to provide a statistical significant detection (Komatsu et al., 2011).

Despite this success, there is much more that can be done. In particular, CMB experiments and galaxy surveys map out only a tiny fraction of our Universe's comoving volume, as we can see in Figure 1-1. The thick black line denotes the CMB,

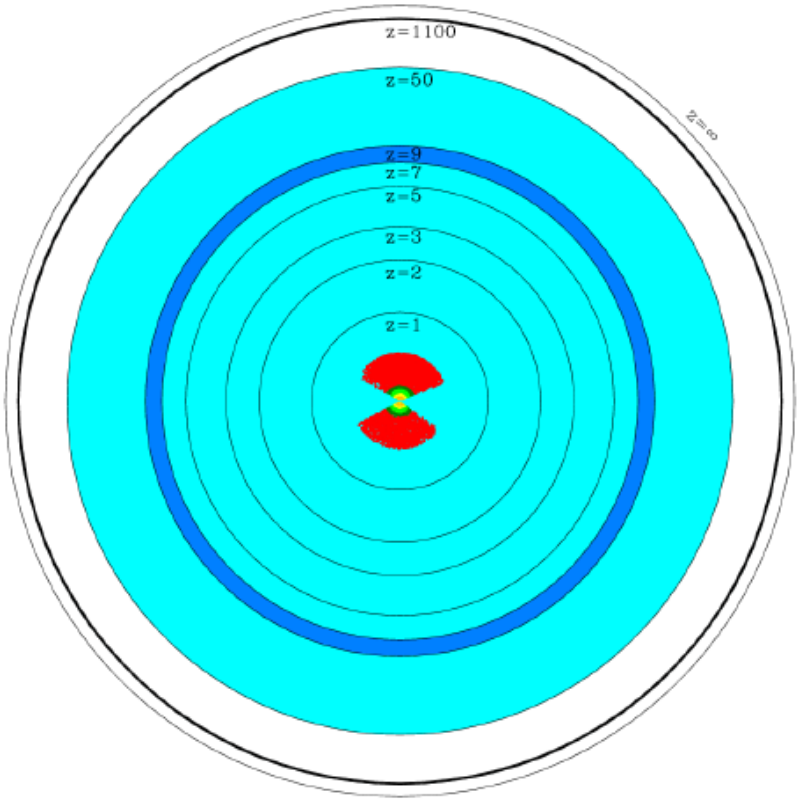

Figure 1-1: A scale diagram of the comoving volume of our observable Universe, reproduced from Mao et al. (2008). We are located at the center of the diagram, the red and yellow regions denote the reach of galaxy surveys, and the cosmic microwave background is shown in the thick black line. Hydrogen cosmology using the 21 cm line can potentially map out the entire region in light blue, while the dark blue strip shows the range of redshifts probed by current generation experiments.

which, as a *surface* of last scattering, provides information mostly from just a thin shell at high redshift[1]. Galaxy surveys probe a three-dimensional volume, but limitations on sensitivity and the abundance of galaxies at high redshifts limit their radial extent to low redshifts. Since error bars on theoretical parameters typically scale as $1/\sqrt{N}$, where N is the number of independent cosmological modes measured, we would clearly do much better if we could map out a larger fraction of our observable Universe.

Hydrogen cosmology (also known as 21 cm cosmology) has the ability fill the unexplored volume between galaxy surveys and the CMB. The basic idea is to use the hyperfine 21 cm transition of the hydrogen atom to map the spatial distribution of hydrogen. Since one is observing an optically thin spectral line, the resulting maps are fully three-dimensional, because the observed wavelength of the signal can be compared to the rest wavelength of 21 cm to determine the redshift. Moreover, the technique only requires diffuse concentrations of hydrogen and not collapsed astronomical objects such as galaxies, allowing it to probe epochs of our cosmic history that come before galaxy formation.

In hydrogen cosmology, neutral hydrogen atoms in the inter-galactic medium are backlit by the CMB, causing observable emission or absorption in the redshifted 21 cm line. Whether the signal appears as an emission feature or an absorption feature (or does not appear at all) depends on the interplay between the CMB temperature T_γ and the spin temperature T_s, which is defined by

$$\frac{n_1}{n_0} = \frac{g_1}{g_0} e^{-E_*/k_B T_s}, \tag{1.1}$$

where n_1 and n_0 are the relative number densities of neutral hydrogen atoms in the excited and ground hyperfine states respectively, $g_1 = 3$ and $g_0 = 1$ are the statistical weights, k_B is Boltzmann's constant, and $E_* = 5.9 \times 10^{-6}\,\mathrm{eV}$ is the hyperfine energy splitting. The spin temperature quantifies the relative number of hydrogen in the excited versus the ground hyperfine state, and translates into an observed 21 cm brightness temperature T_b via (Furlanetto et al., 2006)

$$T_b(\hat{\mathbf{r}}, \nu) = (27\,\mathrm{mK}) x_H \left(\frac{T_s - T_\gamma}{T_s}\right) \left(\frac{1+z}{10}\right)^{1/2} (1+\delta_b)(1+\delta_x) \left[\frac{\partial v_r/\partial r}{(1+z)H(z)}\right]^{-1}, \tag{1.2}$$

[1]There are some exceptions to this, such as the Integrated Sachs-Wolfe effect, the Sunyaez-Zeldovich effect, and gravitational lensing of the CMB. These are all modifications to the primary anisotropies of the CMB that are imprinted on the signal as CMB photons travel from the surface of last scattering to our telescopes.

where $\widehat{\mathbf{r}}$ is a unit vector specifying a direction in the sky, ν is the measured frequency, $\bar{x}_H$ is the mean neutral hydrogen fraction, z is the redshift, $H(z)$ is the Hubble parameter, v_r is the velocity along the radial (line-of-sight) direction r (including both the peculiar velocity as well as the Hubble flow), δ_b is the fractional baryon overdensity, and δ_x is the neutral fraction overdensity. Since x_H, δ_b, δ_x, and T_s all vary with position, the 21 cm signal is anisotropic. Mapping the signal at multiple redshifts also provides information on the evolution of the underlying physics.

In Figure 1-2 we show theoretical predictions for the spatially averaged 21 cm brightness temperature. One sees that the redshift range over which the brightness temperature is non-zero is quite considerable, leading to the vast light blue region in Figure 1-1, which shows the potential reach of hydrogen cosmology. Given this large range of redshifts, the 21 cm line can in principle probe a wide variety of different phenomena, and in what follows we will split our discussion into three epochs. First, a high redshift epoch from $z \sim 200$ to $z \sim 40$ provides a view of our Universe prior to the formation of the first luminous objects, allowing "clean" studies of the underlying cosmology. A low redshift epoch from $z \sim 6$ to $z \sim 2$ comes after the formation of the first luminous objects, allowing "late" effects such as dark energy to be studied. In between these two epochs is an intermediate period from $z \sim 40$ to $z \sim 6$, enabling studies of the formation of the first luminous objects. For each epoch we will consider two main types of experiments: tomographic measurements seek to measure the spatial fluctuation of the brightness temperature field, encapsulating the information in statistical measures such as the power spectrum, while global signal measurements constrain the evolution of the spatially averaged signal with redshift.

1.1.1 High redshifts ($z \sim 200$ to $z \sim 40$)

After recombination, the intergalactic medium (IGM) is mostly composed of neutral hydrogen. However, a small density of residual electrons remain, and these can Compton scatter off CMB photons and subsequently collide with hydrogen atoms. This causes a tight coupling between the spin temperature and the CMB temperature, equilibrating them and causing there to be no net absorption or emission of 21 cm photons. The result is a zero brightness temperature, and hydrogen cosmology is impossible until the coupling becomes weak at $z \sim 200$ (see Pritchard & Loeb (2010) or Chapter 7 for more details).

Between $z \sim 200$ to $z \sim 40$, the hydrogen atoms trace the dark matter distribution, and so a measurement of the 21 cm power spectrum amounts to a measurement of the

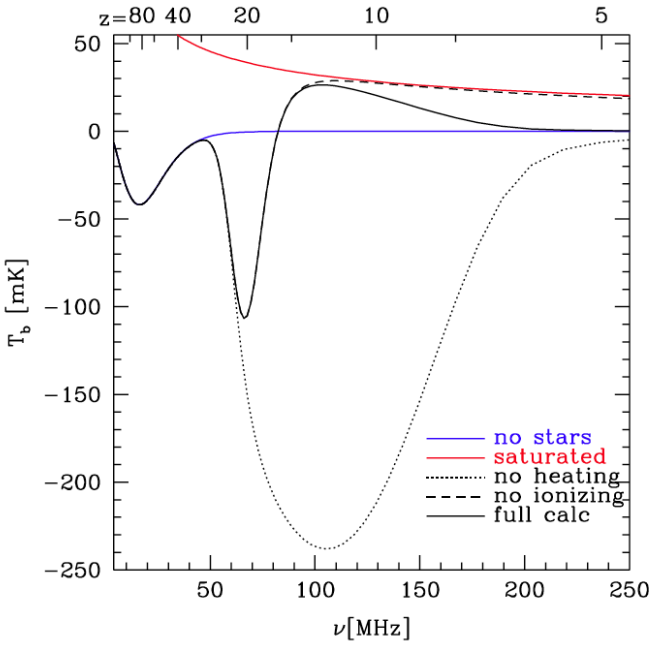

Figure 1-2: The theoretically expected global (*i.e.* spatially averaged) 21 cm signal. The solid black curve represents a fiducial model, and includes various effects such as X-ray heating of the inter-galactic medium (IGM), Lyman-α interactions between the first luminous objects and the IGM, and the reionization of the IGM. The other curves represent extreme, pedagogical models that do not represent physically possible scenarios, but are used to illustrate the various physical processes involved. In the model given by the blue curve, star formation does not occur, and the resulting lack of Lyman-α photons removes a coupling between the 21 cm transition and the gas temperature of the IGM (known as the Wouthysen-Field effect Wouthuysen (1952)). This has the effect of removing the absorption trough around 70 MHz. The red curve is a saturated model, obtained by taking the limit $T_s \gg T_\gamma$ and $x_H = 1$ in Equation 1.2. The dotted curve represents a model where there is negligible reheating of the IGM, and the dashed curve is one where the IGM is never ionized. Please see Chapter 7 for more details on the underlying physics of this figure. Reproduced from Pritchard & Loeb (2010).

matter power spectrum. Given the enormous volume that such a measurement would cover, the error bars on such a power spectrum would be exquisite. Moreover, being a higher redshift probe of the matter power spectrum than galaxy surveys, more cosmological modes are in the linear regime, and the theoretical modeling of the signal becomes simpler. Forecasts suggest that this may result in large improvements in our measurements of cosmological parameters. The cosmological constraint on the sum of the neutrino masses, for instance, may improve by two orders of magnitude (Mao et al., 2008). The small-scale properties of dark matter would also be better understood (Tegmark & Zaldarriaga, 2009). Studies of non-Gaussianity would also improve, since at these high redshifts there are no luminous objects, so there is less "gastrophysics" to worry about (Cooray et al., 2008).

1.1.2 Low redshifts ($z \sim 6$ to $z \sim 2$)

At low redshifts, our Universe has already been reionized by the first stars and galaxies, and there is little neutral hydrogen left in the IGM. Naively, this would suggest that hydrogen cosmology should be impossible after reionization. However, studies have suggested that pockets of self-shielded hydrogen residing in galaxies can give rise to an aggregate signal that can be detected if averaged over a large number of galaxies.

One application of this would be to once again use hydrogen as a tracer of the underlying matter power spectrum (Visbal et al., 2009; Loeb & Wyithe, 2008). However, now that the hydrogen resides in non-linear astronomical objects, work must be done to model the bias factor, which can be scale and redshift dependent (Wyithe & Loeb, 2009). Another application is to look for the baryon acoustic oscillation (BAO) scale (Wyithe et al., 2008; Chang et al., 2008; Masui et al., 2010; Mao, 2012). Doing so using the redshifted 21 cm line rather than galaxy surveys has two advantages. First, the 21 cm measurement can be made over a rather wide range of redshifts ($z \sim 2$ to 4 for the first generation of experiments), providing a standard ruler over wide epoch during which there is substantial cosmological evolution. This can be used to place constraints on the time evolution of the dark energy equation of state w. The second advantage is that these 21 cm BAO experiments do not require high angular resolution, for they need to average over multiple galaxies by design. In contrast, a galaxy survey needs to distinguish between contaminant field stars and distant galaxies, so good angular resolution is needed even if one is interested only in measuring large scale features such as the BAO scale.

1.1.3 Intermediate redshifts ($z \sim 40$ to $z \sim 6$)

At intermediate redshifts, the first luminous sources form, and our Universe undergoes reionization, transitioning from being mostly neutral to mostly ionized during what is known as the Epoch of Reionization (EoR). Unsurprisingly, this epoch contains rich 21 cm signatures. For instance, suppose one were to measure the 21 cm brightness temperature distribution $T_b(\mathbf{r})$. From this, one can obtain the power spectrum $P_T(k)$, which is defined by the equation

$$\langle \widehat{T}_b(\mathbf{k})^* \widehat{T}_b(\mathbf{k}') \rangle = (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') P_T(k), \tag{1.3}$$

where $\widehat{T}_b$ is the three-dimensional spatial Fourier transform of T_b, and δ is the Dirac delta function. For convenience, it is sometimes useful to consider the quantity

$$\Delta_{21}(k) \equiv \sqrt{\frac{k^3}{2\pi^2} P_T(k)}, \tag{1.4}$$

whose square measures the contribution to the root-mean-square fluctuations in $T_b(\mathbf{r})$ per logarithmic bin in k. In Figure 1-3 we show some example plots of $\Delta_{21}(k)$, reproduced from Morales & Wyithe (2009). The different curves and percentages refer to the averaged ionized fraction of the inter-galactic medium. Clearly, the shape of the power spectrum changes with ionization fraction, and theoretical simulations suggest that the ionization fraction is much more important in determining the shape than the redshift z (McQuinn et al., 2006a) . Thus, by measuring the power spectrum as a function of redshift, the ionization history of our Universe can be constrained.

As reionization proceeds, power shifts from fine scales (high k) to larger scales (low k) before disappearing completely. Heuristically, this is because the fluctuations in 21 cm emission mostly trace fluctuations in the density field when the ionized fraction is low, and the density fluctuations have much power at large k (at least as far as the quantity $\Delta_{21}(k)$ is concerned). As reionization proceeds, the fluctuations are dominated by the topology of ionized "bubbles" from which there is no 21 cm emission. These bubbles are formed in the vicinity of the first luminous sources as these sources ionize the IGM around them, and imprint a "knee" in the power spectrum, which moves to increasingly large scales as these bubbles grow. By studying the power spectrum in detail, then, constraints can be placed on the process of reionization (Madau et al., 1997; Tozzi et al., 2000a,b; Iliev et al., 2002; Furlanetto et al., 2004a; Loeb & Zaldarriaga, 2004; Furlanetto et al., 2004b; Barkana & Loeb, 2005b; Furlanetto et al., 2009). For instance, a measurement of the EoR power spectrum will be a good probe

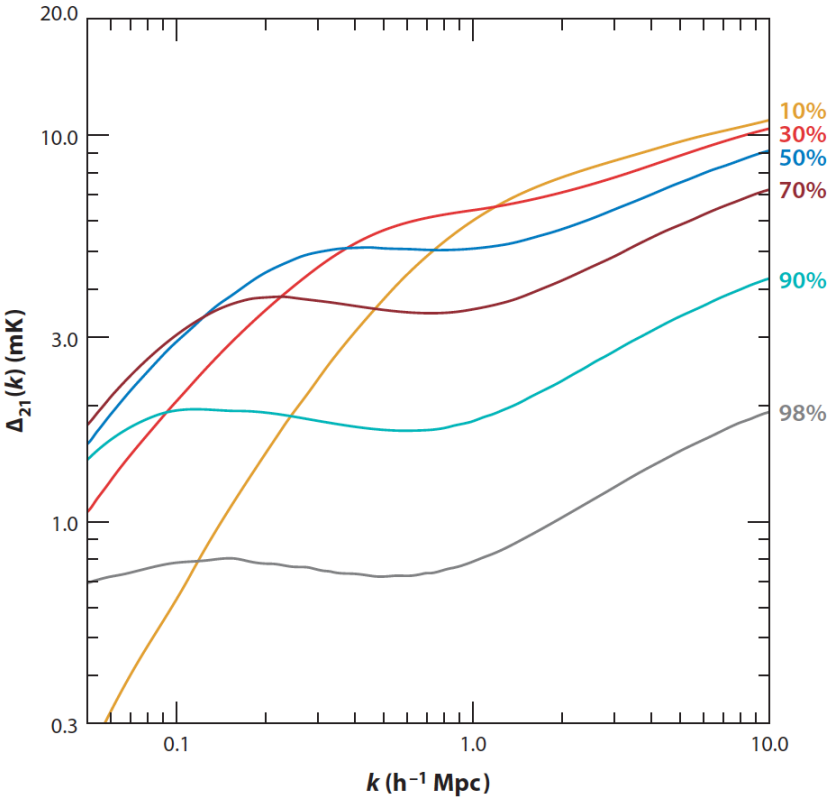

Figure 1-3: Theoretical models for the quantity $\Delta_{21}(k)$, with different curves signifying different ionization fractions of the IGM. Reproduced from Morales & Wyithe (2009).

of the nature of the first ionizing sources, since the characteristic masses of these sources affect the sizes of the ionized bubbles (McQuinn et al., 2006a). Reionization that is driven by larger mass objects can be thought of as reionization that is driven by objects that are more biased relative to the overall density field. This gives rise to larger bubbles, since the biased sources are more clustered and thus have greater ionizing power.

EoR measurements using 21 cm techniques will be particularly valuable because they are the only direct probe of the relevant redshifts. The CMB places constraints on reionization, but these constraints are indirect. For instance, free electrons in the post-reionization epoch will scatter CMB photons as they travel from the surface of last scattering to our telescopes, and this scattering is measured when one fits for the optical depth in the CMB data (Komatsu et al., 2011). However, the optical depth is an integrated quantity, and is therefore insensitive to changes in the reionization history as long as the column density of free electrons remains the same. The kinetic Sunyaev-Zeldovich (kSZ) signature in the CMB also depends on reionization, but

to go from the measured kSZ signal to reionization parameters requires elaborate reionization simulations (Zahn et al., 2011; Mesinger et al., 2012). The Gunn-Peterson effect limits our ability to directly probe all but the very end of reionization using traditional optical or IR measurements (Fan et al., 2006), and studies of individual high redshift galaxies using the Hubble Space Telescope (for example) are at best an indirect probe of the statistical properties of reionization (Robertson et al., 2010).

Hydrogen cosmology thus has the potential to play a unique role in our understanding of the EoR, via both the tomographic experiments and the global signal observations. These EoR measurements will be our focus for the rest of this thesis, although it should be noted that many of the techniques that we develop are also applicable to the high and low redshift regimes.

1.2 The Experimental Reality of Hydrogen Cosmology

While theoretically promising, hydrogen cosmology is observationally difficult. This is for two principal reasons. First, hydrogen cosmology requires radio telescopes of extraordinary sensitivity. From Equation 1.2 and Figure 1-2, we see that the cosmological signal is expected to be extremely faint, on the order of ~ 10's of mK. Thus, it is essential that the noise levels in one's instrument are comparable or below this. The second difficulty in hydrogen cosmology is presence of non-cosmological sources of radio flux *i.e.* foreground contaminants such as Galactic synchrotron radiation. In this section, we will discuss these problems and the limitations they impose.

Consider first the issue of sensitivity. For concreteness, let us examine the noise level in an imaging experiment[2], where one seeks to measure the brightness temperature of the redshifted 21 cm signal as a function of redshift and angle on the sky. Manipulating the standard radiometer equation yields an r.m.s noise of

$$\delta T_{noise} = \frac{D_{max}^2}{A_{tot}} \frac{T_{sys}}{\sqrt{\Delta\nu \Delta t}}, \tag{1.5}$$

where A_{tot} is the total effective collecting area of one's instrument, T_{sys} is the system temperature, $\Delta\nu$ is the frequency channel bandwidth, Δt is the integration time, and

[2]In what follows we quote expressions for the instrumental noise level that are appropriate only for imaging experiments. These expressions require modification to be directly applicable to tomographic measurements. However, the mathematics is provided only for concreteness, and the conceptual arguments carry over to tomographic measurements.

D_{max} is the diameter over which receiving elements are spread (Furlanetto et al., 2006). For instance, with a single dish radio telescope D_{max} is equal to the physical diameter of the dish, whereas with a radio interferometer array D_{max} is given by the distance between the most widely separated receiving elements. Naturally, one wishes to design an instrument that minimizes δT_{noise}, and in the following paragraphs we will briefly review the trade-offs to be considered.

Equation 1.5 clearly suggests that one should integrate for as long as possible, since $\delta T_{noise} \to 0$ as $\Delta t \to \infty$. However, given that one has finite resources, the problem is somewhat more subtle. For instance, with a limited field of view, should the total integration time be spread over multiple parts of the sky? Or should one concentrate on getting the longest integration over a single field? This problem has been solved in the context of CMB measurements (Tegmark, 1997a), and the answer depends on the signal-to-noise (S/N) ratio of the experiment. At low S/N, the measurement is limited by instrumental noise, so it is advantageous to focus on a single field on the sky. At high S/N, cosmological measurements are limited by sample variance, where a single field may contain cosmological fluctuations that are unrepresentative of the "true" underlying statistics of the observable Universe. In this regime it is therefore advantageous to sample multiple fields, reducing sample variance. Between these two extremes, it can be shown mathematically that one should integrate over a single field until one reaches S/N$\approx$ 1 before imaging multiple fields. Current instruments are firmly in the low S/N limit, and stand to gain from further integration.

Another way to increase sensitivity is to reduce D_{max}. For an interferometer array (say) with a fixed number of receiving elements (and thus fixed total collecting area A_{tot}), this corresponds to taking the receiving elements and placing them in as tightly packed a physical configuration as possible. This results in an interferometer with a large number of redundant baselines (pairs of antennas separated by some displacement vector). Since a given baseline of an interferometer array measures a particular Fourier mode on the sky, this provides us with an intuitive understanding of why tightly packed arrays have high sensitivity—with a large number of identical baselines, one measures the same Fourier mode multiple times, and the signal can be summed coherently. This results in a power sensitivity that scales with N_b, where N_b is the number of identical/redundant baselines that go into determining a particular $\mathbf{k}$ mode. This is in contrast with combining measurements with the same wavenumber k but with differing (non-identical) orientations, where the summations must be performed incoherently, yielding only a $\sqrt{N_b}$ gain (Parsons et al., 2011). Thus, in the low S/N regime discussed in the previous paragraph, one improves the measurement

by designing arrays that are maximally redundant. An example of such an array configuration would one where the receiving elements are arranged in a regularly spaced rectangular grid[3]. Such an arrangement has the added advantage that visibilities can be computed in $N \log N$ time rather than N^2 time, where N is the number of receiving elements. There are thus many reasons to build compact, redundant arrays.

Conventionally, however, there has been resistance to the notion of building compact arrays. Since the angular resolution of an instrument is roughly given by $\theta \sim \lambda / D_{max}$, the cost of building a compact array to increase sensitivity is decreased angular resolution. From a theoretical perspective, this is fine as the cosmological signal is expected to be isotropic, and thus one can compensate for poor angular resolution by having high spectral resolution. From a practical perspective, however, critics of compact arrays point out that most calibration algorithms rely on the ability to resolve known, bright point sources. They point out—correctly—that if an instrument is not correctly calibrated, then the considerations of Equation 1.5 are irrelevant, for the dominant source of error would not be noise, but rather, calibration offsets. The problem is worse than simply an overall, multiplicative flux scale, as calibration parameters may be off for individual receiving elements within a larger array, distorting any resulting images. Whether or not highly redundant arrays can be precisely calibrated without the luxury of high angular resolution is an open question. One promising technique for doing so is known as redundant baseline calibration, where the fact that identical baselines should yield identical measurements is used to self-consistently solve for calibration parameters. In Chapter 2 we examine this possibility, and recently the technique has met with some experimental success (Noorishad et al., 2012).

Unlike the parameters of Equation 1.5 that have been considered so far, the system temperature T_{sys} is not one that can be easily improved upon by better instrumental design, which reduces only the *receiver noise temperature* from the electronics of one's receiving elements. Receiver noise is not the dominant contribution to T_{sys} at the low radio frequencies relevant to hydrogen cosmology. Instead, instruments tend to be *sky-noise dominated* (Morales & Wyithe, 2009), where $T_{sys} \approx T_{sky}$, the sky temperature. This temperature is typically large, as it is dominated not by the cosmological signal but instead by non-cosmological foregrounds such as Galactic synchrotron emission. These foregrounds are extremely bright at low frequencies, reaching temperatures of 100's to 1000's of K (Wang et al., 2006; de Oliveira-Costa et al., 2008; Shaver et al.,

[3]The discussion that follows can in fact be generalized to any *hierarchy* of regular grids, not just single rectangular grids. Please see Tegmark & Zaldarriaga (2010) for details.

1999), giving rise to a large T_{sky}. Since receiver noise temperatures can be easily kept to a magnitude of $\sim$ 10's of K (Morales & Wyithe, 2009), the dominant source is noise is sky noise.

While our discussion thus far has been guided by the mathematics of an imaging experiment, the qualitative conclusions that we have reached carry over to tomographic experiments, where one seeks to measure a power spectrum of the 21 cm brightness fluctuations $P(k)$. In the noise-limited (as opposed to sample variance-limited) regime, the errors in a power spectrum estimated by an interferometric measurement can be approximated by the scaling relation

$$\delta P(k) \propto A_e^{-3/2} B^{-1/2} k^{-3/2} n(k_\perp) \frac{T_{sys}^2}{\Delta t}, \tag{1.6}$$

where B is the *total* bandwidth of the entire observation, k is the spatial wavenumber of temperature fluctuations, and $n(k_\perp)$ is the number of interferometer baselines sensitive to angular Fourier modes with wavenumber $k_\perp$ (Furlanetto et al., 2006). It is normalized so that the integral over the half-plane of angular Fourier space $\mathbf{k}_\perp$ is equal to the number of baselines. Since all quantities that appear in the numerator of Equation 1.5 appear in the numerator of Equation 1.6 and similarly for quantities in the denominator, our qualitative arguments hold whether we are referring to imaging or power spectrum measurements. Put succinctly, the foreground sky is bright at low frequencies, giving rise to a high sky-noise dominated system temperature T_{sys}, which necessitates compact arrays with large collecting area and long integration times to sufficiently suppress noise fluctuations for a sensitive measurement of the cosmological signal.

Besides being a source of noise in our measurement, foregrounds provide yet another headache in our quest to detect the cosmological signal. Since they are truly a sky signal (albeit not the signal we are after), foregrounds will appear in our images of the sky, and need to be removed before the cosmological power becomes apparent. Thus, even in the limit of an ideal, noiseless experiment, it is necessary to have a robust way to remove foreground contamination from the data. The challenge is daunting at frequencies relevant to hydrogen cosmology. At such frequencies (100 to 200 MHz, and potentially even lower for next-generation instruments), Galactic synchrotron radiation, bright radio point sources, and a continuum of unresolved radio sources all contribute to the radio sky. As mentioned above, Galactic synchrotron radiation has a brightness temperature that can reach $\sim$ 100's of K, and the unresolved radio sources (mostly due to galactic synchrotron from distant unresolvable galaxies)

will typically contribute another few ~ 10's of K (Wang et al., 2006; de Oliveira-Costa et al., 2008; Shaver et al., 1999). The resolved point sources (being point sources) are not directly comparable in temperature units, but do contribute significantly, and are a significant contaminant that must be dealt with.

The aforementioned foreground sources must be subtracted in a way that not only mitigates their effects on the data, but also leaves the cosmological signal untouched as far as possible. While the cosmological community has had much success removing foregrounds from CMB measurements in the past, the foreground problem in hydrogen cosmology is more difficult to deal with, for several reasons. First, CMB measurements are typically conducted at much higher frequencies ($\sim 100\,\mathrm{GHz}$). Most foregrounds decrease in amplitude towards higher frequencies, and thus are much dimmer at CMB experiment-appropriate frequencies. In fact, the foregrounds are so much dimmer that the cosmological signal is the dominant signal in a CMB experiment that is pointed away from the Galactic plane. In addition, the CMB spectrum is almost precisely that of a blackbody (although extremely minor perturbations may arise due to effects such as the thermal Sunyaev-Zeldovich effect). The formal brightness temperature of the cosmological CMB signal is thus frequency-independent, which is not the case for foregrounds. By measuring the CMB at different frequencies, then, one has a set of redundant measurements that can be used to distinguish signal from foregrounds. In hydrogen cosmology, we do not have this luxury, for as we mentioned above, different observed frequencies of the 21 cm line correspond to different redshifts, and therefore contain unique information. Designing a foreground subtraction algorithm that can clean effectively while being lossless (*i.e.* does not destroy information about the cosmological signal) is therefore a non-trivial task, and we discuss this in Chapter 6 after having built up to it in Chapters 3, 4, and 5.

1.3 Current and Near-Future Hydrogen Cosmology Experiments

Despite the challenges highlighted in the previous section, the lure of rich scientific rewards have led to a large number of hydrogen cosmology experiments in recent years. A small number make use of existing telescope facilities [such as the Giant Metrewave Radio Telescope (GMRT)]. The remainder represent new instrumentation development, and the resulting experiments are in various stages of the scientific cycle, with some currently in active data acquisition modes while others exist only as

promising proposals. We shall now give a brief overview of the relevant intermediate redshift experimental efforts.

1.3.1 Global Signal Experiments

Currently, the only global signal experiment with real data is the Experiment to Detect the Global EoR Signal (EDGES) (Bowman et al., 2008; Bowman & Rogers, 2010). EDGES consists of a single dipole antenna deployed in Western Australia, operating between 100 and 200 MHz. After three months of (non-continuous) observation, the instrument has produced a spectrum averaged over most of the southern celestial sphere—a good approximation to the global spectrum.

Unfortunately, the global spectrum that has been measured is not simply that of the cosmological signal. The thermal noise is quite well-suppressed (about 6 mK), but the measured spectrum is dominated by strong foregrounds. Because foregrounds are expected to be spectrally smooth (to a good approximation, they are well-fit by power laws with small corrections (Wang et al., 2006)), EDGES is able to rule out abrupt changes in the cosmological global signal. Such a signal would arise if, for instance, our Universe experienced a rapid and abrupt epoch of reionization. This would mean a sudden change in ionization fraction at some redshift, thus giving a sudden drop-off of redshifted 21 cm emission. Based on such reasoning, EDGES has been able to place a lower limit (at the 95% confidence level) of $\Delta z > 0.06$ for the duration of reionization.

Beyond EDGES, a number of other global signal experiments have been proposed. The Dark Ages Radio Explorer (DARE) is a proposed satellite experiment that will target a lower frequency range (40 to 120 MHz), making it a powerful probe of early reionization physics, as discussed earlier in this chapter. If funded, DARE will be placed in a lunar orbit, using the Moon as a shield for terrestrial sources of radio frequency interference. The receiving elements on the satellite consist of a pair of dipoles, with a ground plane that narrows the beam pattern to a full-width-half-maximum power of 57° at 75 MHz (Burns et al., 2012). Among other factors, this narrower beam pattern (compared to an experiment like EDGES) will help DARE better separate foregrounds from the cosmological signal, and early simulations suggest that the satellite will be able to easily detect signatures of reionization (Harker et al., 2012). However, these simulations may be overly optimistic as far as foreground models go, and in Chapter 7 we improve on the assumptions made in previous studies and propose new data analysis algorithms.

Targeting approximately the same frequency range as DARE but from the ground is the Large-aperture Experiment to Detect the dark Ages (LEDA) experiment (Greenhill & Bernardi, 2012). The instrument will be located at a Long Wavelength Array (LWA) station, and uses four dipoles to make a global signal measurement at 30 to 88 MHz. These dipoles will also be correlated with 256 LWA receiving elements for calibration and foreground subtraction purposes. Detailed projections of the LEDA's performance have yet to be published.

1.3.2 Tomographic Experiments

Tomographic hydrogen cosmology experiments are those that seek to either image the redshifted 21 cm emission or to measure the power spectrum. They are typically targeted at higher frequencies than global signal experiments. They therefore probe late reionization scenarios, when ionized bubbles around the first ionizing sources imprint strong spatial features into the data that are amenable to characterization by power spectra. At higher frequencies, foregrounds are also fainter, lessening their contaminating effect.

The GMRT-Epoch of Reionization (GMRT-EoR) experiment currently has the best upper limit on the EoR power spectrum (Paciga et al., 2010). The experiment consists of 30 antennas, each with a diameter of 45 m (thus giving a large collecting area). The antennas are operated as an interferometer array, with 14 antennas arranged in a tight central core of less than 1 km. The rest of the antennas are spaced far apart, with the longest baseline (separation between antennas) being 25 km. Operating with a bandwidth of 16.7 MHz centered around 150 MHz, GMRT-EoR targets a redshift range of $8.1 < z < 9.2$. Their limits on the fluctuation power per logarithmic bin in k given by $\Delta^2(k) = \frac{k^3}{2\pi^2} P(k)$ are $\sim 10^{-1}\,\mathrm{K}^2$ and $\sim 10^{-3}\,\mathrm{K}^2$ at $k = 0.2h\mathrm{Mpc}^{-1}$ and $k = 0.65h\mathrm{Mpc}^{-1}$ respectively. These limits are based on 50 hours of observation, and more time has been granted for further integration.

Unlike GMRT-EoR, the Murchison Widefield Array (MWA) is a custom-built instrument for EoR measurements (Lonsdale et al., 2009). The MWA consists of groups of 16 dipoles that are analog beamformed into a single receiving element (a "tile"). The eventual goal is to form an interferometer array with 512 such tiles, with the outputs from the tiles individually digitized and correlated as an interferometer array. A prototype 32-tile system has recently completed a series of data runs, and in Chapter 8 we present preliminary results from a power spectrum measurement using data centered around 140 MHz.

The Precision Array for Probing the Epoch of Reionization (PAPER), like the MWA, is custom-built for the EoR (Parsons et al., 2010). The two experiments, however, differ in their design philosophies. The MWA's antenna tiles are relatively cheap but numerous, producing a well-controlled synthesized beam. The plan is then to employ a real-time system to simultaneously fit for calibration parameters using a large number of calibration sources. PAPER, on the other hand, currently uses 64 extremely stable and well-characterized antenna elements, each of which are dual-polarization dipole antennas mounted on a ground-screen structure. The stability of their antenna elements aids calibration, and all-sky foreground maps have been published using data between 139 MHz and 174 MHz from the experiment (Parsons et al., 2010). PAPER is currently experimenting with redundant baseline interferometer layouts, the hope being that redundant calibration algorithms will be sufficient to control calibration systematics, allowing the array to take advantage of the sensitivity gains mentioned above.

The Low Frequency Array (LOFAR) is designed to be a general purpose low frequency radio observatory, spanning a frequency range of 10 to 250 MHz (Jelic et al., 2008). The observatory is comprised of a large interferometer array spread over northwestern Europe, but the EoR measurement will primarily use 864 antennas concentrated in a central core in the Netherlands. The 864 antennas are each formed from 4×4 collections of dipoles which are analog beamformed. These antennas/collections of dipoles are further sorted into 36 stations with digital beamforming (and therefore multiple-beamforming) capability. LOFAR intends to take advantage of supercomputing resources to fit for calibration and nuisance parameters using a maximum likelihood formalism. EoR results are expected from LOFAR next year.

In summary, the experimental reality of hydrogen cosmology is one that is best described as a work-in-progress. Sensitivity and foreground subtraction requirements are formidable, though hopefully not insurmountable, and multiple approaches to the problem are being pursued simultaneously by different groups.

1.4 Roadmap

The goal of this thesis is to tackle many of the observational obstacles highlighted in the previous section. The vast majority of the following chapters should be considered "theoretical experiment", and involve forecasting, simulation, and algorithm development:

- Chapter 2, titled *"Precision Calibration of Radio Interferometers Using Redun-*

dant Baselines" deals with the problem of sensitivity in hydrogen cosmology. Specifically, we consider whether the technique of *redundant baseline calibration* can be used to calibrate radio interferometers to a sufficient precision to meet their design sensitivities for hydrogen cosmology. The chapter is based on a published paper (Liu et al., 2010), written in collaboration with Max Tegmark, Scott Morrison, Andrew Lutomirski, and Matias Zaldarriaga. The idea of applying redundant calibration to arrays designed for hydrogen cosmology originated from Max Tegmark, and Andrew Lutomirski and Scott Morrison did some preliminary numerical explorations. I produced all the numerical results presented in the paper and developed the generalized formalism for calibrating near-redundant arrays.

- Chapter 3, titled *"How well can we measure and understand foregrounds with* 21 *cm experiments?"* moves onto the issue of foreground contamination. While guided by the eventual need to remove foregrounds from the data, the focus is on our ability to model foregrounds. Using a principal component analysis, we show that foregrounds have "simple" spectra, in the sense that they can be parameterized by a small handful of parameters. The chapter is based on a published paper (Liu & Tegmark, 2012), written in collaboration with Max Tegmark, who suggested the use of the Wiener filter. I performed all calculations.

- Chapter 4, titled *"Will point sources spoil 21cm tomography?"* deals with the actual subtraction of foregrounds. We simulate a popular proposal for foreground subtraction in hydrogen cosmology, where low-order polynomials are subtracted from spectra along individual lines of sight in the data. The chapter is based on the published paper (Liu et al., 2009b), written in collaboration with Max Tegmark and Matias Zaldarriaga. Max Tegmark suggested the project, Matias Zaldarriaga provided critical feedback, and I performed all calculations.

- Chapter 5, titled *"An Improved Method for 21cm Foreground Removal"* is a direct sequel to Chapter 4. We show that the method presented in Chapter 4 can be improved at virtually no extra computational cost by cleaning foregrounds Fourier-mode-by-Fourier-mode, instead of line-of-sight-by-line-of-sight. The chapter is based on a published paper (Liu et al., 2009a), written in collaboration with Max Tegmark, Judd Bowman, Jacqueline Hewitt, and Matias Zaldarriaga. I performed most of the calculations in this paper, while Judd

Bowman simulated an independent pipeline to verify the results. Max Tegmark, Jacqueline Hewitt, and Matias Zaldarriaga provided vital feedback.

- Chapter 6, titled *"A Method for* 21 *cm Power Spectrum Estimation in the Presence of Foregrounds"* generalizes the foreground subtraction methods presented in previous chapters, and places foreground subtraction in a unified framework with power spectrum estimation. The chapter is based on a published paper (Liu & Tegmark, 2011), written in collaboration with Max Tegmark, who conceived of this project and developed many of the relevant mathematical tools in the context of previous CMB and galaxy survey work. I adapted these tools for hydrogen cosmology and performed all the calculations.

- Chapter 7, titled *"Optimizing global* 21 *cm signal measurements with spectral and spatial information"* moves away from tomographic measurements and focuses on measurements of the global 21 cm signal. An optimal algorithm for extracting the cosmological signal is derived and simulated. The chapter is based on research done in collaboration with Jonathan Pritchard, Max Tegmark, and Abraham Loeb. Abraham Loeb suggested the project, Jonathan Pritchard provided simulations of the theoretical signal, Max Tegmark provided critical feedback, and I developed the formalism and performed all the calculations. This chapter has yet to be published.

Following these chapters, we conclude with an attempted power spectrum measurement using the Murchison Widefield Array:

- Chapter 8, titled *"An Empirical Upper Limit on the Redshifted* 21 *cm Power Spectrum with the Murchison Widefield Array"* uses the 32-tile MWA system to place an upper limit on the redshifted 21 cm power spectrum. The chapter is based on research done in collaboration with Christopher Williams, Max Tegmark, and Jacqueline Hewitt. Christopher Williams reduced the data from its raw output to maps of the sky, while I took the maps of the sky and estimated power spectra and error statistics from them using the quadratic estimator formalism. Jacqueline Hewitt and Max Tegmark provided critical feedback and experience. This chapter has yet to be published.

It goes without saying that in all of the projects listed above, Max Tegmark served as my research advisor, providing feedback and guidance in every aspect of the research.

Chapter 2

Precision Calibration of Radio Interferometers Using Redundant Baselines

2.1 Introduction

Technological advances of recent years have enabled the design and construction of radio telescopes with broadband frequency coverage, low instrumental noise, moderate angular resolution, high spectral resolution, and wide fields of view. As mentioned in Chapter 1, these instruments have generated considerable scientific interest due to their unique potential to do 21 cm cosmology, mapping the high-redshift universe using redshifted radio emission from neutral hydrogen.

From a practical standpoint, however, none of the aforementioned applications will be feasible without reliable algorithms for instrumental calibration. Compared to traditional radio astronomy experiments, instruments designed for 21 cm tomography will require particularly precise calibration because of the immense foreground subtraction challenges involved. Epoch of Reionization experiments, for example, will be attempting to extract cosmological signals from beneath foregrounds that are of order 10^4 times brighter (de Oliveira-Costa et al., 2008). Such procedures, where one must subtract strong foregrounds from a bright sky to uncover weak cosmological signals, are particularly susceptible to calibration errors. A prime example of this is the effect that a miscalibration has on the subtraction of bright, resolved point sources. Datta et al. (2009), for instance, argued that bright point sources must be localized to extremely high positional accuracy in order for their subtraction to be successful,

and Liu et al. (2009b) showed that failure to do so could comprise one's subsequent ability to subtract any *unresolved* point sources, making the detection of the 21 cm cosmological signal difficult. This places stringent requirements on calibration accuracy.

While standard algorithms are available for calibrating radio interferometers, the unprecedented challenges faced by 21 cm tomography and the advent of a new class of large radio arrays (such as PAPER (Parsons et al., 2010), LOFAR (Jelic et al., 2008), MWA (Lonsdale et al., 2009), CRT (Peterson et al., 2006), the Omniscope (Tegmark & Zaldarriaga, 2010), and the SKA (Carilli & Rawlings, 2004)) mean that it is timely to consider the relative merits of various algorithms. For example, several authors (Bhatnagar et al., 2008; Morales & Matejek, 2009) have pointed out that these new high precision radio arrays require close attention to be paid to heterogeneities in the primary beams of different antenna elements. In this chapter, we focus on the technique of *redundant calibration*, which takes advantage of the fact that these new telescopes generically possess a large number of redundant or near-redundant baselines. This property of exact or near-redundancy is not an accident, and arises from the science requirements of 21 cm tomography[1], where the demanding need for high sensitivity on large spatial scales calls for arrays with a large number of short baselines (Morales, 2005; Lidz et al., 2008). Since these arrays typically also have very wide fields-of-view and in many cases operate at frequencies where all-sky survey data is scarce, the fewer assumptions one needs to make about the structure of the sky signal, the better. This is an advantage that redundant calibration has over other calibration methods — no *a priori* assumptions need to be made about calibrator sources.

The use of redundant baselines for calibration goes far back. For example, Noordam & de Bruyn (1982) used this approach to produce high dynamic range images of $3C84$, and similar techniques have been used in Solar imaging (Ishiguro, 1974; Ramesh et al., 1999). The methods used are described in detail in Wieringa (1992); Yang (1988); Pearson & Readhead (1984), and closely resemble the logarithmic implementation we review in Section 2.2.3. In this chapter, we build on this previous work and extend it in several ways:

1. We provide a detailed examination of the errors and biases involved in existing redundant baseline methods.

[1]Note that we are *not* advocating the redesign of arrays to achieve redundancy for the sake of redundant baseline calibration. The design of an array should be driven by its science requirements, and the ability to use redundant calibration should simply be viewed as an option to be explored if there happens to be redundancy.

2. We introduce methods to mitigate (or, in the case of biases, eliminate) these problems.

3. We show how slight deviations from prefect redundancy and coplanarity can be incorporated into the analysis,

with the view that it may be necessary to correct for all these effects in order to achieve the precision calibration needed for precision cosmology.

The rest of the chapter is organized as follows. After defining our notation, we describe the simplest possible calibration algorithm (*i.e.* point source calibration) in Section 2.2.1. In Section 2.2.3 we review the redundant baseline calibration proposed by Wieringa (1992). Readers familiar with radio interferometer calibration techniques are advised to skip to Section 2.2.4, where we examine Wieringa's scheme more closely by deriving the optimal weightings for its fitting procedure. In Section 2.2.5 we focus on simulations of redundant calibration for monolithic square arrays, which show that existing redundant calibration algorithms have certain undesirable error properties. These properties suggest a linearized variation that we propose in Section 2.2.7. In Section 2.3 we consider redundant calibration algorithms and point source calibration schemes as extreme cases of a generalized calibration formalism, which allows us to treat the case of near-redundant baseline calibration. We summarize our conclusions in Section 2.4.

2.2 Calibration Using Redundant Baselines

We begin by defining some basic notation. Suppose an antenna i measures a signal s_i at a given instant[2]. This signal can be written in terms of a complex gain factor g_i, the antenna's instrumental noise contribution n_i, and the true sky signal x_i that would be measured in the limit of perfect gain and no noise:

$$s_i = g_i x_i + n_i. \tag{2.1}$$

[2]More precisely, s_i refers to frequency domain data at a given instant. In other words, it is the result of dividing the time-ordered data into several blocks, each of which is then Fourier transformed in the time direction to give a signal s_i that is both a function of frequency and time (albeit at a coarser time resolution than the original time-ordered data).

Each baseline measures the correlation between the two signals from the two participating antennas:

$$
\begin{aligned}
c_{ij} &\equiv \langle s_i^* s_j \rangle && \text{(2.2a)} \\
&= g_i^* g_j \langle x_i^* x_j \rangle + \langle n_i^* n_j \rangle + g_i^* \langle x_i^* n_j \rangle + g_j \langle n_i^* x_j \rangle && \text{(2.2b)} \\
&= g_i^* g_j \langle x_i^* x_j \rangle + n_{ij}^{res} \equiv g_i^* g_j y_{i-j} + n_{ij}^{res}, && \text{(2.2c)}
\end{aligned}
$$

where we have denoted the true correlation $\langle x_i^* x_j \rangle$ by y_{i-j}, and the angled brackets $\langle \dots \rangle$ denote time averages. The index $i-j$ is not to be taken literally but is intended to signify the fact that any given correlation should depend only on the *relative* positions of the two antennas. As is usual in radio interferometry, we have assumed that noise contributions from different antennas are uncorrelated, and that instrumental noise is uncorrelated with the sky signal[3]. This means that $\langle n_i^* x_j \rangle$ reduces to $\langle n_i^* \rangle \langle x_j \rangle$, and since $\langle n_i \rangle \to 0$ in the limit of long integration times, we can assume that the last three terms in Equation 2.2b average to some residual noise level n^{res}. The size of n^{res} will of course depend on the details of the instrument, but in general will be on the order of $T_{sys}/\sqrt{\tau \Delta\nu}$, where T_{sys} is some fiducial system temperature, τ is the total integration time, and $\Delta\nu$ is the bandwidth. For the purposes of this chapter, the details of the noise term do not matter, and for notational convenience we will drop the superscript "res" from now on. The key qualitative results of this chapter (such as the fact that many existing redundant baseline calibration schemes are biased) do not depend on the noise level.

It should be noted that the equation for c_{ij} presented above does not represent the most general form possible for the measured correlations. So-called non-closing errors, for example, can result in gain factors that do not factor neatly on an antenna-by-antenna basis. In this chapter we will forgo a discussion of these errors, with the view that a good number of these contributions can be mitigated (although not eliminated completely) by good hardware design. Alternatively, our results may be interpreted as best-case scenario *upper limits* on calibration quality.

Since the gain factors in Equation 2.2c are in general complex, we can parameterize them by an amplitude gain η and a phase φ by letting $g_i \equiv e^{\eta_i + i\varphi_i}$. Our equation

[3]This assumption is expected to be valid so long as the antennas are far enough apart that mutual coupling effects do not significantly affect the data. Noorishad et al. (2012) found that such effects mainly caused sidelobe beam patterns to be non-identical, which is in principle bad news for redundant calibration. In practice, however, they found that it is still possible to use redundant calibration to produce calibration results that are competitive with more traditional techniques.

then becomes

$$c_{ij} = \exp\left[(\eta_i + \eta_j) + i(\varphi_j - \varphi_i)\right] y_{i-j} + n_{ij}. \tag{2.3}$$

The goal of any radio interferometer is to extract the true correlations y_{i-j} (from which one can construct a sky map using Fourier transforms) from the measured correlations c_{ij}. Formally, however, this is an unsolvable problem for the generic case, as one can see from Equation 2.3 — with N antennas, one has $N(N-1)/2$ measured correlations (c_{ij}'s), which are insufficient to solve for the $N(N-1)/2$ true correlations (y_{i-j}'s) in addition to the $2N$ unknown η's and φ's. (The n_{ij} noise terms are also unknown quantities, but are generally not treated as ones that need to be solved for, but as noise to be minimized in a least-squares fit). In abstract terms, all calibration schemes can be thought of as methods for reducing the number of unknowns on the right hand side of Equation 2.3 so that the system of equations becomes overdetermined and the η's and φ's can be solved (or fit) for.

2.2.1 Basic Point Source Calibration

The most straightforward way to calibrate a radio telescope is to image a single bright point source. A point source has constant amplitude in uv-space, and thus y_{i-j} is a complex number whose complex amplitude is independent of baseline. Moreover, the phase of y_{i-j} must also be zero, since by definition we are only "looking" at the bright point source if our radio array is phased so that the phase center lies on the source. Thus, $y_{i-j} = S$, and Equation 2.3 reduces to

$$c_{ij} = \exp\left[(\eta_i + \eta_j) + i(\varphi_j - \varphi_i)\right] S + n_{ij}, \tag{2.4}$$

which is an overdetermined set of equations for all but the smallest arrays, since we have $N(N-1)/2$ complex inputs on the left hand side but only $2N$ real numbers (η's and φ's) to solve for on the right hand side. The system is therefore solvable, up to two intrinsic degeneracies: the overall gain $\sum_i \eta_i$ is indeterminate, as is the overall phase $\sum_i \varphi_i$. These degeneracies, however, can be easily dealt with by having some knowledge of our instrument and our calibration source. For instance, the overall gain can be computed if the calibrator source is a catalogued source with known flux density.

It is important to note that the point source calibration method we have presented in this section represents only the simplest, most straightforward way in which one

could calibrate a radio telescope. There exist far more sophisticated methods[4], such as the many self-calibration schemes that are capable of calibrating a radio telescope off most reasonable sky signals. In practice, one almost never needs to use basic point source calibration, for it is just as easy to use a self-calibration algorithm. The scheme described here should therefore be considered as no more than a "toy model" for calibration.

2.2.2 Redundant Baseline Calibration

The main drawback of the approach described in the previous section is that it requires the existence of an isolated bright point source. This requirement becomes increasingly difficult to fulfill as cosmological science drivers push radio telescopes towards regimes of higher sensitivity and wider fields of view. While self calibration methods have been shown to produce high dynamic range maps over wide fields of view and do *not* require the existence of isolated bright point sources, they are reliant on having a reasonable *a priori* model of the sky. Thus, if possible it may be preferable to opt for redundant calibration, which is completely model independent. With the emergence of telescopes with high levels of redundancy (such as the Cylinder Radio Telescope — see Peterson et al. (2006) — or any omniscopes[5] like those described in Tegmark & Zaldarriaga (2009, 2010)) redundant baseline calibration becomes a competitive possibility.

Mathematically, if an array has a large number of redundant baselines, then one can reduce the number of unknowns on the right hand side of Equation 2.3 by demanding that the true visibilities y_{i-j} for a set of identical baselines be the same. Since the number of *measured* visibilities c_{ij}'s stays the same, our system of equations becomes overdetermined provided there are a sufficient number of redundant baselines, and it becomes possible to fit for the η and φ parameters.

As an example, consider a one-dimensional array of five radio antennas where the antennas are equally spaced. In this case, one has four unique baselines (of lengths 1, 2, 3, and 4), and 10 measured correlations. Thus, one must fit for 18 real numbers (four complex numbers from the true visibilities of the four unique baselines, plus five

[4]See Cornwell & Fomalont (1999) or Rau et al. (2009), for example, for nice reviews.

[5]We define an omniscope as any instrument where antenna elements are arranged on a regular grid, and full digitized data is collected on a per element basis without any beamforming.

η's and five φ's) from 20 numbers (10 complex measured correlations):

$$\begin{aligned}
c_{1,2} &= \exp\left[(\eta_1+\eta_2)+i(\varphi_2-\varphi_1)\right] y_1 + n_{1,2} \\
c_{2,3} &= \exp\left[(\eta_2+\eta_3)+i(\varphi_3-\varphi_2)\right] y_1 + n_{2,3} \\
&\vdots \quad \text{(all four baselines of length 1)} \\
c_{1,3} &= \exp\left[(\eta_1+\eta_3)+i(\varphi_3-\varphi_1)\right] y_2 + n_{1,3} \\
c_{2,4} &= \exp\left[(\eta_2+\eta_4)+i(\varphi_4-\varphi_2)\right] y_2 + n_{2,4} \\
&\vdots \quad \text{(all three baselines of length 2)} \\
c_{1,4} &= \exp\left[(\eta_1+\eta_4)+i(\varphi_4-\varphi_1)\right] y_3 + n_{1,4} \\
&\vdots \quad \text{(both baselines of length 3)} \\
c_{1,5} &= \exp\left[(\eta_1+\eta_5)+i(\varphi_5-\varphi_1)\right] y_4 + n_{1,5}.
\end{aligned} \tag{2.5}$$

As is usual in radio interferometry, we have omitted equations corresponding to baselines of length zero, *i.e.* autocorrelations of the data from a single antenna. This is because autocorrelation measurements carry with them correlated noise terms, and thus are likely to be much noiser than other correlated measurements.

In the following sections we will present two methods for explicitly solving Equation 2.5. We begin in Section 2.2.3 with a logarithmic method, which is very similar to one described by Wieringa (1992). This method is the simplest way to implement a redundant baseline calibration scheme, but suffers from several problems which we detail in Section 2.2.5 and 2.2.6. In Section 2.2.7, we show how these problems can be solved with an alternative method based on linearization.

2.2.3 The Logarithmic Implementation

We proceed by rewriting our equations in the following way:

$$c_{ij} = g_i^* g_j y_{i-j} \left(1 + \frac{n_{ij}}{g_i^* g_j y_{i-j}}\right). \tag{2.6}$$

Taking the logarithm gives

$$\ln c_{ij} = \eta_i + \eta_j + i(\varphi_j - \varphi_i) + \ln y_{i-j} + \underbrace{\ln\left(1 + \frac{n_{ij}}{g_i^* g_j y_{i-j}}\right)}_{\equiv w_{ij}}, \tag{2.7}$$

and requiring that the real and imaginary parts be separately equal gives linear equations of the form

$$\ln|c_{i,j}| = \eta_i + \eta_j + \ln|y_{i-j}| + \mathrm{Re}\, w_{ij} \tag{2.8a}$$

$$\arg|c_{i,j}| = \varphi_j - \varphi_i + \arg|y_{i-j}| + \mathrm{Im}\, w_{ij}. \tag{2.8b}$$

Note that the amplitude and phase calibrations have now decoupled[6], and can thus be written as two separate matrix systems. The amplitude calibration, for example, takes the form

$$\underbrace{\begin{pmatrix} \ln|c_{1,2}| \\ \ln|c_{2,3}| \\ \ln|c_{3,4}| \\ \ln|c_{4,5}| \\ \ln|c_{1,3}| \\ \ln|c_{2,4}| \\ \ln|c_{3,5}| \\ \vdots \\ \ln|c_{2,5}| \\ \ln|c_{1,5}| \end{pmatrix}}_{\equiv \mathbf{d}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ & \vdots & & & \vdots & & & \vdots & \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\equiv \mathbf{A}} \underbrace{\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \ln|y_1| \\ \ln|y_2| \\ \ln|y_3| \\ \ln|y_4| \end{pmatrix}}_{\equiv \mathbf{x}} + \underbrace{\begin{pmatrix} s_{1,2} \\ s_{2,3} \\ s_{3,4} \\ s_{4,5} \\ s_{1,3} \\ s_{2,4} \\ s_{3,5} \\ \vdots \\ s_{2,5} \\ s_{1,5} \end{pmatrix}}_{\equiv \mathrm{Re}\,\mathbf{w}}, \tag{2.9}$$

for the one-dimensional, five-element array considered in Section 2.2.2. The vector $\mathbf{d}$ stores the measured correlations, the matrix $\mathbf{A}$ is completely determined by the array configuration, and the $\mathbf{x}$ is what we hope to solve for.

As it currently stands, however, this set of linear equations has no unique solution even in the absence of noise (*i.e.* even when $\mathbf{w} = 0$). This can be seen from the fact that the vector $\mathbf{x}_{null} = (1, 1, 1, 1, 1, -2, -2, -2, -2)$ (written here as a row vector for convenience) lies in the null space of $\mathbf{A}$ (*i.e.* $\mathbf{A}\mathbf{x}_{null} = 0$), so for any solution $\mathbf{x}_0$, one can form a new solution $\mathbf{x}_0 + \mathbf{x}_{null}$. The new solution corresponds to adding a constant to all η's, which — since $g \equiv \exp(\eta + i\varphi)$ — is equivalent to multiplying all gains by a constant factor and simultaneously dividing all the sky signals by the same factor.

[6]From a rigorous standpoint, it is not strictly true that the amplitude and phase calibrations of an interferometer decouple. Indeed, this is evident even in our simple model, where $\mathrm{Re}\, w_{ij}$ and $\mathrm{Im}\, w_{ij}$ both contain factors of the complex gain $g \equiv e^{\eta + i\varphi}$, which means that Equations 2.8a and 2.8b are in principle coupled, even if only weakly so. However, one must remember that the noise parameters are treated simply as given numbers which are *not* solved for when solving these calibration equations. This means that in finding the best fit solution to Equations 2.8a and 2.8b, the η and φ calibrations *can be performed separately.* It is only in this sense that the systems are "decoupled".

Physically, this is an expression of the fact that internal calibration schemes (*i.e.* schemes like this one where the calibration is done by exploiting internal mathematical consistencies) are insensitive to the absolute gain of the radio interferometer as a whole, a problem which is present even performing a simple point source calibration, as we noted in Section 2.2.1.

This mathematical degeneracy in absolute amplitude calibration can be broken by arbitrarily specifying an absolute gain level. For example, one can add the equation

$$0 = \sum_i \eta_i \tag{2.10}$$

to the system of equations, thus ensuring that one cannot uniformly elevate the gain levels and still satisfy all constraints. The set of equations governing the phase calibration require the addition of a similar equation, since the absolute phase of a system is not a physical meaningful quantity and therefore must be arbitrarily specified:

$$0 = \sum_i \varphi_i. \tag{2.11}$$

With the phase calibration, however, there exist two additional degeneracies: the calibration is insensitive to tilts of the entire telescope in either the x or the y direction, since such tilts are equivalent to rotations of the sky and redundant algorithms are (by design) independent of the nature of the sky signal. To break these degeneracies, one adds the following equations:

$$0 = \sum_i r_{x,i}\varphi_i \tag{2.12a}$$

$$0 = \sum_i r_{y,i}\varphi_j, \tag{2.12b}$$

where $\mathbf{r}_i \equiv (r_{x,i}, r_{y,i})$ denotes the physical location of the ith antenna in the array. While these extra equations are somewhat arbitrary (the L.H.S. of Equation 2.10, for instance, can be any real number), the true gain, phase, and tilt parameters can always be fixed after-the-fact by referring to published flux densities and locations of known, catalogued bright point sources in the final sky maps.

With these four degeneracies broken, our equations can be solved using the familiar least-squares estimator $\hat{\mathbf{x}}$ for the parameter vector:

$$\hat{\mathbf{x}} = \left[\mathbf{A^t N^{-1} A}\right]^{-1} \mathbf{A^t N^{-1} d}. \tag{2.13}$$

Here, $\mathbf{N}$ is the noise covariance matrix, which takes the form $\langle \mathrm{Re}\,\mathbf{w}\,\mathrm{Re}\,\mathbf{w}^{\mathbf{t}} \rangle$ for the amplitude calibration and $\langle \mathrm{Im}\,\mathbf{w}\,\mathrm{Im}\,\mathbf{w}^{\mathbf{t}} \rangle$ for the phase calibration. In the following section we examine the form of this matrix in detail. Note that since the parameter vector $\mathbf{x}$ contains not only the calibration factors η and φ but also the true sky correlations y_{i-j}, one can solve for the true sky signal directly from the uncalibrated correlations $\mathbf{d}$, as one expects from any self or redundant calibration scheme.

This concludes our review of the Wieringa (1992) redundant calibration method, cast in notation conducive to the development of the new material that follows. It is worth noting that in Wieringa (1992), the noise covariance matrix $\mathbf{N}$ is set to the identity. In the next section, we show that this is *not* the optimal weighting to use for the calibration fit, and derive the noise covariance matrix that corresponds to inverse variance weighting.

2.2.4 The Noise Covariance Matrix

The form of $\mathbf{N}$ will depend on one's model for the instrumental noise. In what follows, we will for convenience assume that the noise is Gaussian, small compared to the signal, and uncorrelated between baselines; generalization to other noise models is straightforward[7]. Although these assumptions are not necessary for calibration (since Equation 2.13 can be implemented for any invertible choice of $\mathbf{N}$), making them will allow us to gain a more intuitive understanding of the errors involved. Recalling the form of the noise term after taking the logarithm of our equations, we can say

$$w_{ij} = \ln\left(1 + \frac{n_{ij}}{g_i^* g_j y_{i-j}}\right) \approx \frac{n_{ij}}{g_i^* g_j y_{i-j}}, \tag{2.14}$$

where we have invoked the assumption of high signal ($g_i^* g_j y_{i-j}$) to noise (n_{ij}) to expand the logarithm. Since $c_{ij} = g_i^* g_j y_{i-j} + n_{ij}$, we can rewrite this as

$$w_{ij} \approx \frac{n_{ij}}{c_{ij} - n_{ij}} \approx \frac{n_{ij}}{c_{ij}}, \tag{2.15}$$

where we have neglected higher order terms. Equation 2.15 is more convenient than Equation 2.14 because it is written in terms of the measured correlation c_{ij} instead of the noiseless $g_i^* g_j y_{i-j}$. Since the logarithmic scheme separates the real and imag-

[7]Typically, the electronic noise will be uncorrelated between different antennas/amplifier chains. This means that the noise correlation matrix $\mathbf{N}$ for the baseline will be highly sparse but not diagonal: away from its diagonal, it will have nonzero entries for precisely those pairs of baselines which have an antenna in common. This sparseness is very helpful, making the computations numerically feasible even for massive arrays as described below.

inary parts of the calibration, the crucial quantity is not w_{ij} but rather its real (or imaginary) part:

$$\begin{aligned} w_{ij} &\approx \frac{n_{ij}}{c_{ij}} = \frac{e^{-i\phi}}{|c_{ij}|}(n_x + in_y) && (2.16a) \\ &= \frac{1}{|c_{ij}|}(\cos\phi - i\sin\phi)(n_x + in_y) && (2.16b) \\ \Rightarrow \operatorname{Re} w_{ij} &= \frac{1}{|c_{ij}|}(n_x\cos\phi + n_y\sin\phi), && (2.16c) \end{aligned}$$

where $\phi \equiv \arg c_{ij}$. With this result, we can form the noise covariance matrix $\mathbf{N} \equiv \langle \operatorname{Re}\mathbf{w} \operatorname{Re}\mathbf{w}^t \rangle$. Since we are assuming that the noise is uncorrelated between different baselines, our matrix must be diagonal, and its components can be written in the form $\mathbf{N}_{\alpha\beta} = \langle (\operatorname{Re}\mathbf{w}_\alpha)^2 \rangle \delta_{\alpha\beta}$ where $\delta_{\alpha\beta}$ is the Kronecker delta, and α and β are Greek *baseline* indices formed from *pairs* of Latin *antenna* indices. For example, the baseline formed by antennas $i = 1$ and $j = 2$ might be labeled baseline $\alpha = 1$, whereas that formed by antennas $i = 1$ and $j = 3$ might be given baseline index $\alpha = 2$. The diagonal components are given by

$$\begin{aligned} \mathbf{N}_{\alpha\alpha} &= \langle n_\alpha n_\alpha \rangle && (2.17a) \\ &= \frac{1}{|c_\alpha|^2}(\underbrace{\langle n_x n_x \rangle}_{\equiv\sigma^2}\cos^2\phi + 2\underbrace{\langle n_x n_y \rangle}_{=0}\cos\phi\sin\phi + \underbrace{\langle n_y n_y \rangle}_{\equiv\sigma^2}\sin^2\phi) && (2.17b) \\ &= \frac{\sigma^2}{|c_\alpha|^2}, && (2.17c) \end{aligned}$$

where we have assumed that the real and imaginary parts of the residual noise n are independently Gaussian distributed with spread σ^2. This expression provides a simple prescription for the inverse variance weighting of Equation 2.13: one simply weights each measured correlation by the square modulus of itself (the σ^2 in the numerator is irrelevant in the computation of the estimator $\hat{\mathbf{x}}$ because the two factors of $\mathbf{N}$ in Equation 2.13 ensure that any constants of proportionality cancel out).

For the phase calibration, the fact that inverse variance weighting corresponds to weighting by $1/|c|^2$ has a simple interpretation. In Figure 2-1, we plot various correlations on the complex plane. The green dots represent a set of simulated noiseless visibilities, while the red triangles show what the baselines of a noisy 4 by 4 regular square antenna array would measure. From the plot, it is clear that for a given amount of noise, visibilities that are closer to the origin of the complex plane (*i.e.* those with smaller $|c|$) are more susceptible to loss of phase ($\arg c$) information from

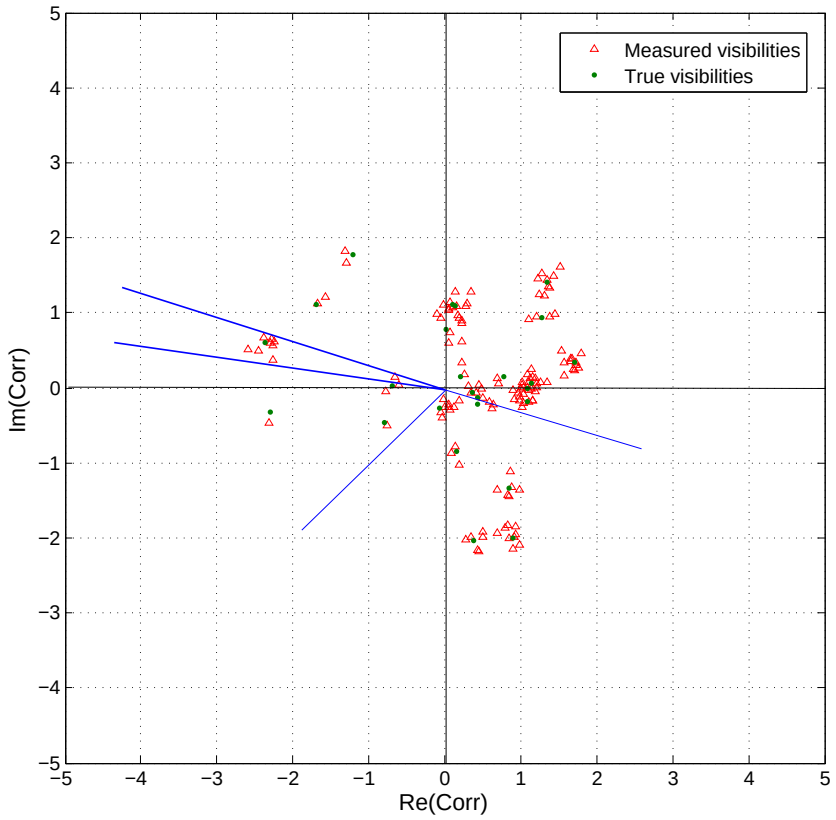

Figure 2-1: A plot of true visibilities (green dots) and measured noisy visibilities (red triangles) on the complex plane. The blue lines delineate the phase extent of the noisy visibilities from two sets of redundant baselines. One sees that in the presence of noise, visibilities closer to the origin are more susceptible to loss of phase information.

noise. The gain ($\log|c|$) errors are similarly large for such visibilities. These visibilities should be given less weight in the overall fit, which is what the inverse variance weighting achieves.

2.2.5 Simulation Results and Error Properties

In Figures 2-2 and 2-3, we show simulated calibration results of a 16 antenna element interferometer. The sky signal used in the simulations was that of a Gaussian random field. That such a sky is somewhat unrealistic is irrelevant, since redundant calibration schemes are sky-independent[8] by construction.

Like in Figure 2-1, the antennas in these simulations were arranged in a 4 by 4 square grid, with the antenna spacings assumed to be completely regular. The pre-

[8]That is, provided we exclude unreasonable mathematical exceptions such as a sky devoid of any sources.

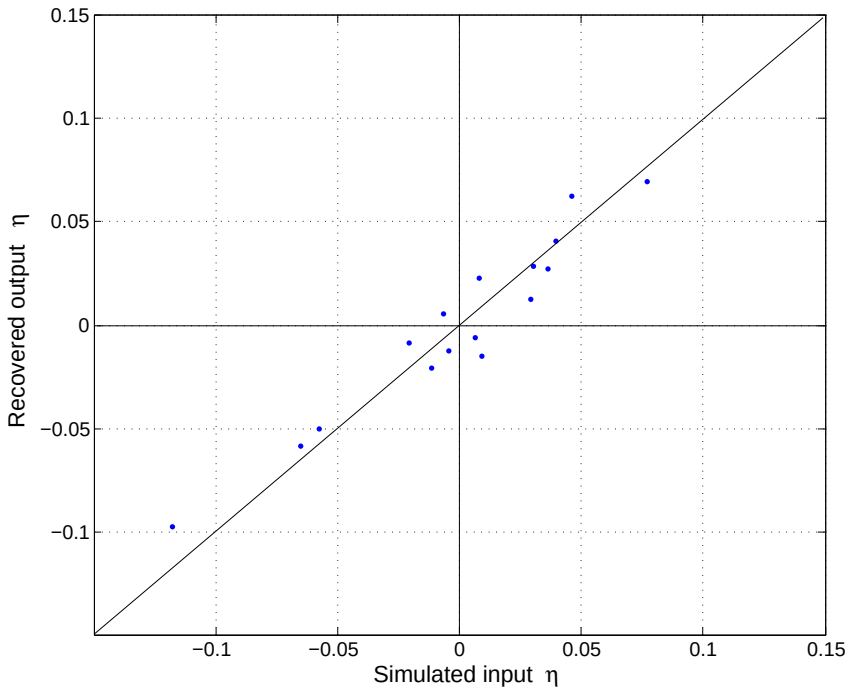

Figure 2-2: Scatter plot of simulated vs. recovered antenna gain parameter η for a 4 by 4 square array with a signal-to-noise ratio of 10.

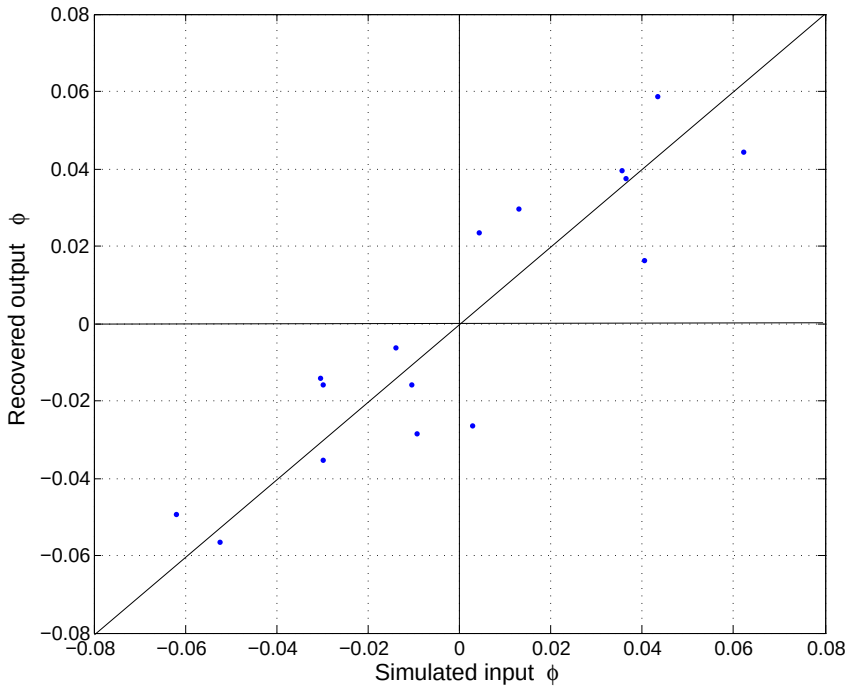

Figure 2-3: Scatter plot of simulated vs. recovered antenna phase parameter φ for a 4 by 4 square array with a signal-to-noise ratio of 10.

cise physical separation between adjacent antennas need not be specified, because a dilation of the entire array layout changes only the sky signal measured by the interferometer, and not the *pattern* of baseline redundancies. Repeating the simulations with rescaled versions of the same array is therefore unnecessary, again because of the fact that we can calibrate from any sky signal. The same reasoning makes it unnecessary to specify an observing frequency for the simulations.

In the plots, it is clear that there is some scatter in the recovered antenna parameters. This is due to non-zero instrumental noise, which was simulated to be Gaussian *i.e.* n_{ij} in Equation 2.2c was taken to be a Gaussian random variable. Rather than using a specific noise model to fix the amplitude of the noise, we simply parameterize the problem in terms of the signal-to-noise ratio (SNR). Figures 2-2 and 2-3 show the results for an SNR of 10. (Such an SNR would be typical for an interferometer with bandwidth $\Delta\nu \approx 10\,\mathrm{kHz}$ and integration time $\tau \approx 10\,\mathrm{s}$ observing Galactic synchrotron emission at 150 MHz). The scatter seen in these figures can be viewed as error bars in the final calibration parameters outputted from the calibration. To quantify these errors, we can compute the deviation $\varepsilon \equiv \mathbf{x} - \hat{\mathbf{x}}$ and form the quantity

$$\mathbf{\Sigma} \equiv \langle \varepsilon \varepsilon^t \rangle = \left[\mathbf{A^t N^{-1} A} \right]^{-1}, \tag{2.18}$$

where the last equality can be proved with a little algebra and Equation 2.13 (Tegmark, 1997b). The square root of the diagonal elements of Σ then give us the expected error bars on the corresponding elements in $\hat{\mathbf{x}}$. For example, σ_1, the expected error in the calibration of the first antenna's η, is given by $(\mathbf{\Sigma}_{11})^{1/2}$ if one arranges the elements of the parameter vector $\mathbf{x}$ in the manner shown in Equation 2.9.

The fact that our system is linear means that to understand the errors in our calibration, we need only compute error bars from simulations of systems with an SNR of 1; the errors from systems of arbitrary SNR can be scaled accordingly. To see this, consider Equation 2.17c, which we can rewrite by scaling out some fiducial mean value $|c_0|$ for the magnitude of the visibilities:

$$\mathbf{N}_{\alpha\beta} = \frac{\sigma^2}{|c_\alpha|^2}\delta_{\alpha\beta} = \frac{\sigma^2}{|c_0|^2|c'_\alpha|^2}\delta_{\alpha\beta} = \frac{1}{(\mathrm{SNR})^2}\frac{\delta_{\alpha\beta}}{|c'_\alpha|^2} \tag{2.19a}$$

$$\equiv \frac{\mathbf{N}'_{\alpha\beta}}{(\mathrm{SNR})^2}, \tag{2.19b}$$

where we have defined normalized visibilities c'_α and a normalized noise covariance

matrix $\mathbf{N'} \equiv \frac{\delta_{\alpha\beta}}{|c'_\alpha|^2}$. Plugging this into Equation 2.18 gives

$$\mathbf{\Sigma} = \left[\mathbf{A^t N^{-1} A}\right]^{-1} = \frac{1}{(\mathrm{SNR})^2}\left[\mathbf{A^t N'^{-1} A}\right]^{-1}. \tag{2.20}$$

We therefore see that in what follows, general conclusions for any SNR can be scaled from our simulations of a system with an SNR of 1.

Although redundant calibration schemes make it *possible* to calibrate an interferometer using any reasonable sky signal, it is important to point out that the *quality* of the calibration, or in other words, the error bars predicted by Equation 2.20, *are* sky-dependent. To see this, consider a situation where the sky signal is dominated by a small number of Fourier modes whose wavevectors are of precisely the right magnitude and orientation for its power to be picked up by only the longest baselines of an array. Since our noise covariance matrix is proportional to $1/|c|$, using such a sky to calibrate our array will give small error bars for the antennas that are part of the longest baselines and relatively large error bars for all other antennas. On the other hand, a sky that has comparable power across all scales is likely to yield calibration error bars that are more consistent across different antennas. Given this dependence of calibration quality on sky signal, in what follows we compute ensemble-averaged error bars based on random realizations of our Gaussian skies to give a sense of the typically attainable accuracy.

From our simulations, we find that the dependence of calibration quality on antenna location is quite weak. This can be seen from the scale on Figure 2-4, where we show the expected error bars (estimated by taking an average of 30 random realizations) in the recovered η for the antennas that comprise a regularly spaced 18 by 18 square interferometer. Far more important than antenna location in determining the calibration errors is the *total* number of antennas in an array. In Figure 2-5, we show the average[9] error in η as a function of the number of antennas in a square array. The error $\Sigma^{1/2}$ roughly asymptotes to a $1/\sqrt{N}$ dependence, where N is the total number of antennas in the array, as would be expected if the calibration algorithm is thought of as combining independent information from different redundant baselines. Precisely the same trend is seen in the φ errors.

In running the simulations, we find that minor variations in the baseline distribution have little effect on the final calibration errors[10]. For instance, varying the

[9] Averaging both over different antennas in an array and over different simulations.

[10] *Major* variations, of course, can drastically affect the errors. An extreme example of this would be a change in baseline distribution that completely destroys all redundancy, which would render the algorithm unusable and formally give rise to infinite error bars.

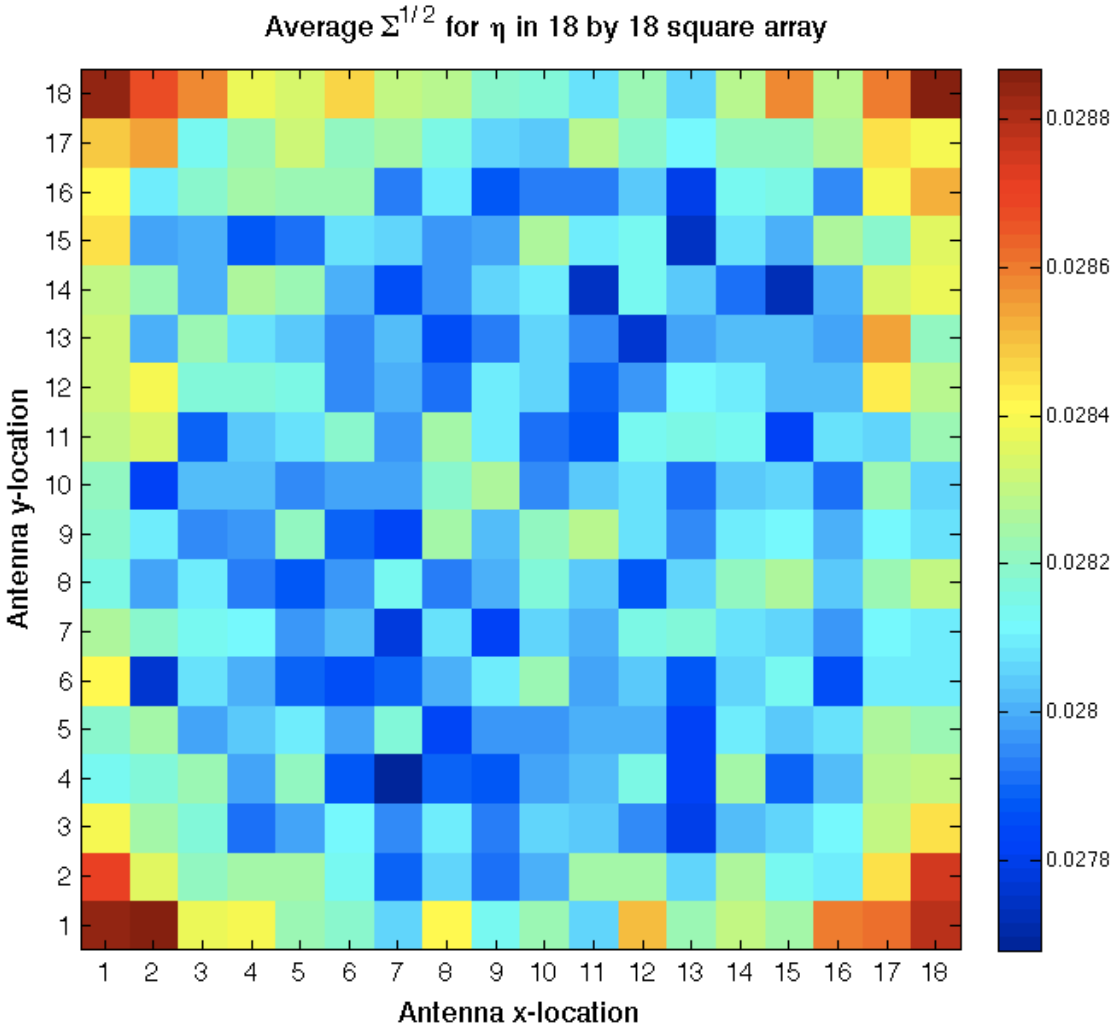

Figure 2-4: Expected error bars in recovered η for different antennas in an 18 by 18 square array with a signal-to-noise (SNR) ratio of 1, although in a realistic situation one would expect a more favorable SNR. By examining the color bar in this figure, one sees that the errors do not vary very much throughout the array. However, there is a systematic effect where the errors are typically larger towards the edges and corners, because antennas there participate in fewer identical baselines than their counterparts near the center of the array.

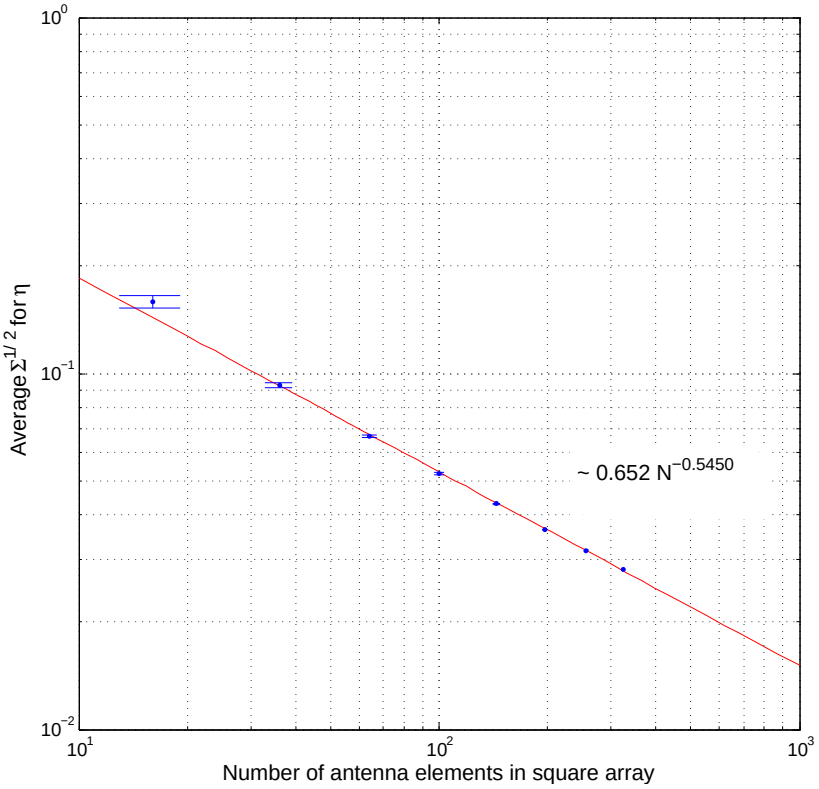

Figure 2-5: Expected average error bars in recovered η as a function of the size of the simulated square array. All simulations have a signal-to-noise (SNR) ratio of 1, although in a realistic situation one would expect a more favorable SNR.

aspect ratio of a 100 antenna system from a 10 by 10 array, to a 5 by 20, 4 by 25, and a 1 by 100 array has mostly minimal effects on the η and φ errors. Thus, for an array with baseline redundancies that are at roughly the same level as for a square array, we can put everything together and arrive at the following rule of thumb for the calibration errors:

$$\Delta\eta \equiv \Sigma_\eta^{1/2} \approx \Delta\varphi \equiv \Sigma_\varphi^{1/2} \approx \frac{0.5}{\mathrm{SNR}} \frac{1}{\sqrt{N}}, \tag{2.21}$$

where N is again the total number of antennas in the array.

Not captured by Equation 2.21 is the fact that antennas closer to the center of an array tend to have lower calibration errors than those that are peripherally located, as can be seen in Figure 2-4. Intuitively, this is due to the fact that while every antenna forms the same number of baselines (because each antenna forms a baseline with all other antennas in the array), antennas that are located in different parts of the array

form different *sets* of baselines. An antenna that is in the corner of a square array, for example, forms $N-1$ baselines that are all different from each other, whereas an antenna that is close to the center will form baselines in redundant pairs because of the symmetry of the square array. In the system of equations, the centrally located antenna will therefore be more tightly constrained by the data, and the resulting error bars in its calibration parameters will be smaller. These differences are, however, rather minimal and are unlikely to significantly affect the actual calibration of a real (non-simulated) array.

As mentioned above, in addition to the antenna calibration parameters η and φ, redundant calibration algorithms yield estimates y_{ij} for the true sky visibilities. In Figures 2-6, 2-7, and 2-8, we show the visibility results that fall out of the calibration. Figure 2-6 serves as a legend for the uv plane, with different colors and symbols signifying the different *unique* baselines found in a 4 by 4 square array. These symbols are used to plot the uncalibrated input correlations measured by the interferometer (*i.e.* the c_{ij}'s) on a complex plane in Figure 2-7. Each color and symbol combination may appear multiple times in Figure 2-7, because two baselines that are identical in principle may give different readings in practice because of instrumental noise and the fact that different antennas will have different calibration parameters. Indeed, one can see from Figure 2-7 that the uncalibrated correlations are rather spread out over the complex plane, and do not seem close to the simulated true sky visibilities, which are denoted in the figure by solid magenta dots.

After calibration, however, one has estimates for the antenna calibration parameters, and thus these uncalibrated correlations can be corrected to yield measurements that are close to the correct sky visibilities. One simply divides out the complex gain factors, since the measured correlations c_{ij} are related to the true sky visibilities y_{i-j} by $c_{ij} = g_i^* g_j y_{i-j}$. The results are shown in Figure 2-8, where it is clear from the color and symbol combinations that identical baselines now yield measurements that cluster around the simulated true values, which are still given by the solid magenta dots. The scatter that remains within a given set of identical baselines is due to instrumental noise (set at a level corresponding to an SNR of 10 for this simulation). It should be noted, however, that while this is a perfectly workable method for computing estimates of the sky visibilities, it is unnecessary. Redundant calibration outputs a parameter vector estimate $\hat{\mathbf{x}}$ (see Equation 2.13) that contains the true visibilities. These are shown in Figure 2-8 using magenta x's, and emerge from the calibration at no additional computational cost[11].

[11] In fact, it is not even strictly valid to consider any "extra" computational cost associated with

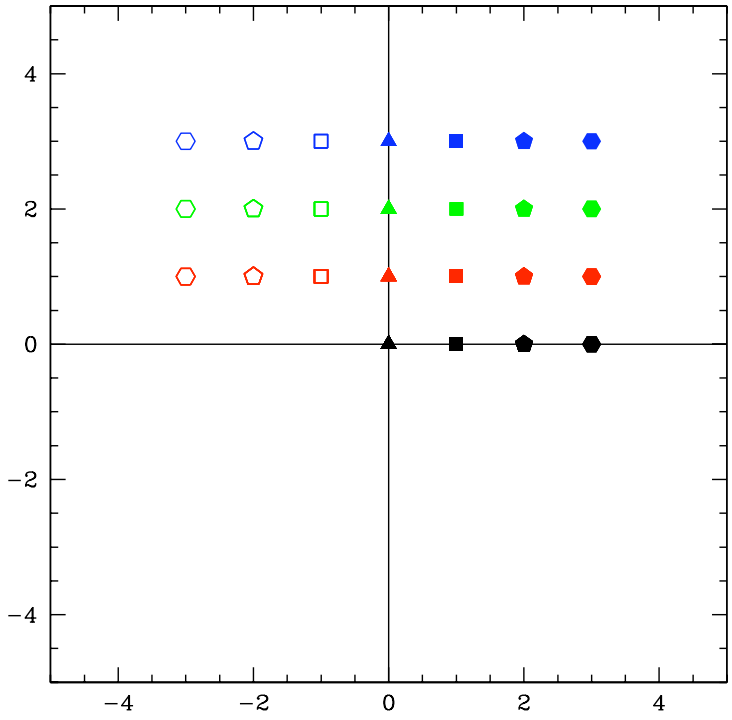

Figure 2-6: Legend of baselines for Figures 2-7 and 2-8.

Since the true visibilities (like the antenna gain parameters) are simply components of the parameter vector $\mathbf{x}$, we can again use the machinery of Equation 2.20 to compute the expected error bars in the estimated visibilities. The results are similar to what was found for the antenna calibration parameters, and are summarized in Figure 2-9. The errors are again found to asymptote to a $1/\sqrt{N}$ dependence, but here N refers to the *number of redundant baselines that go into determining a given visibility*, and not the total number of antennas in the array. That the error bars do not depend on the total number of antennas is clear from Figure 2-9, where data points from simulations of arrays of various sizes lie on top of each other. It is also a property that is intuitively expected, since adding extra antennas increases the mathematical constraints on a given visibility only if the extra antennas form new baselines that correspond to that visibility's location on the uv plane.

solving for the y_{ij}'s, for as one can see from Equation 2.9, the inclusion of the y_{ij}'s is *necessary* for redundant calibration to work.

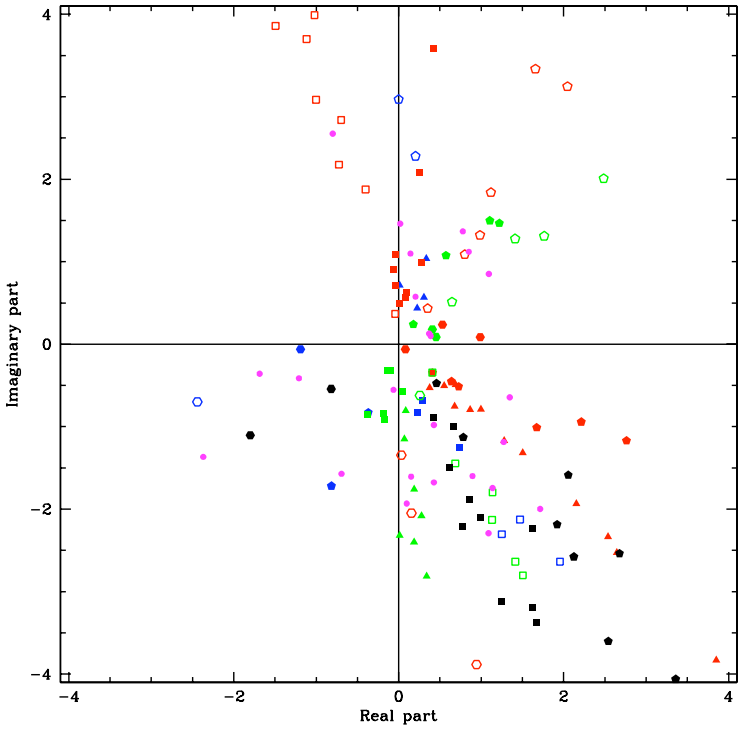

Figure 2-7: Uncalibrated visibility measurements plotted on the complex plane according to the *uv* plane baseline legend provided by Figure 2-6. The simulated (*i.e.* true) sky visibilities are given by the magenta dots.

2.2.6 Problems with the logarithmic implementation

While our simulations and real-world applications like those detailed in Noordam & de Bruyn (1982); Ishiguro (1974); Ramesh et al. (1999); Wieringa (1992); Yang (1988); Pearson & Readhead (1984) have shown above that taking logarithms of Equation 2.3 yields a linear system that can be successfully used to fit for calibration parameters and visibilities, this logarithmic implementation is not without its drawbacks. In particular, the implementation has two undesirable properties:

Phase Wrapping

For the logarithmic implementation to work, the phase calibration parameter φ needs to be close to zero for all antennas. This is because one must take the complex logarithm of $c_{ij} = g_i^* g_j y_{i-j}$, which is an operation that is determined only up to additions or subtractions of multiples of 2π in the imaginary part (*i.e.* in the phase of the original number). Thus, while the calibration solution may correctly recover

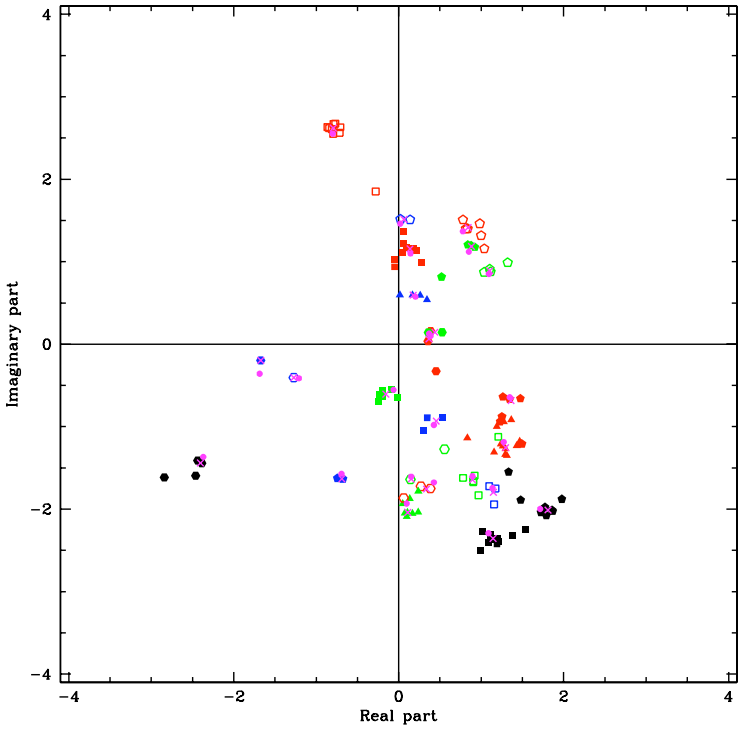

Figure 2-8: Same as Figure 2-7, except the visibilities have been calibrated. The simulated sky visibilities are again given by the magenta dots, while the best estimates from the calibration algorithm for these visibilities are given by the magenta x's.

the *measured* correlations (since one is insensitive to additions or subtractions of 2π in the phase when re-exponentiating to find $c_{ij} = g_i^* g_j y_{i-j}$), it does not give correct phases for g_i^*, g_j, and y_{i-j}, which are ultimately the quantities of interest. This is a problem even when only a small number of antennas have large phase calibration parameters, since to find the visibilities and antenna gain parameters redundant algorithms essentially performs a *global* fit, where all antennas are coupled to each other via the visibilities.

In Figure 2-10 we show the phase calibration results for a *noiseless* simulation of a 4 by 4 array with phases that are significantly greater than zero. The logarithmic method is shown using the open blue circles, and since a noiseless simulation should yield perfect parameter recovery, the scatter in the trend suggests a problem with the method. A detailed examination of the simulation's numerical results reveals that it is indeed the multi-valued property of the complex logarithm that causes the trouble, and from the plot one also sees that a small number of very badly fit antennas can give rise to inaccuracies for all antennas, as is expected from the global nature of the

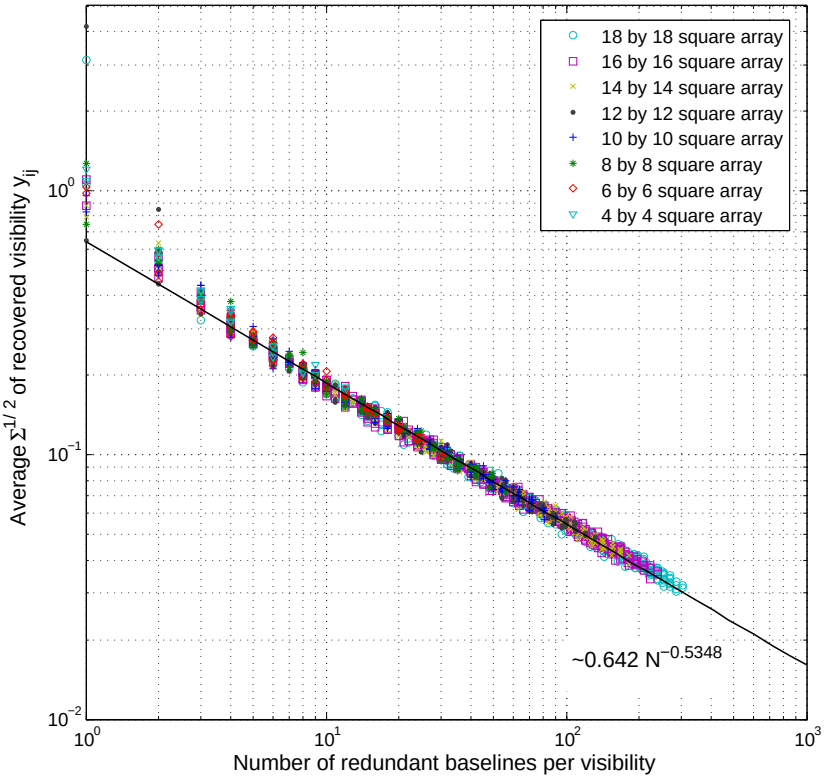

Figure 2-9: Expected error bars on the estimated visibilities, as a function of the number of redundant baselines that go into determining each visibility. A signal-to-noise (SNR) ratio of 1 is assumed, although in a realistic situation one would expect a more favorable SNR.

fit.

It is important to note that even when the phase calibration fails because of large phases, the *amplitude* calibration can still be implemented successfully. This is because the taking of the complex logarithm means that in our system of equations, the amplitudes and phases separate into the real and imaginary parts respectively, as we can see from Equations 2.8a and 2.8b. The two calibrations are therefore completely decoupled, and instabilities in one do not affect the other. The fact that an accurate gain calibration can be extracted regardless of the quality of one's phase calibration will be important in Section 2.2.8.

Bias

The logarithmic method is not unbiased, in the sense that ensemble averages of noisy simulations do not converge to the true simulated parameter values. While Equation

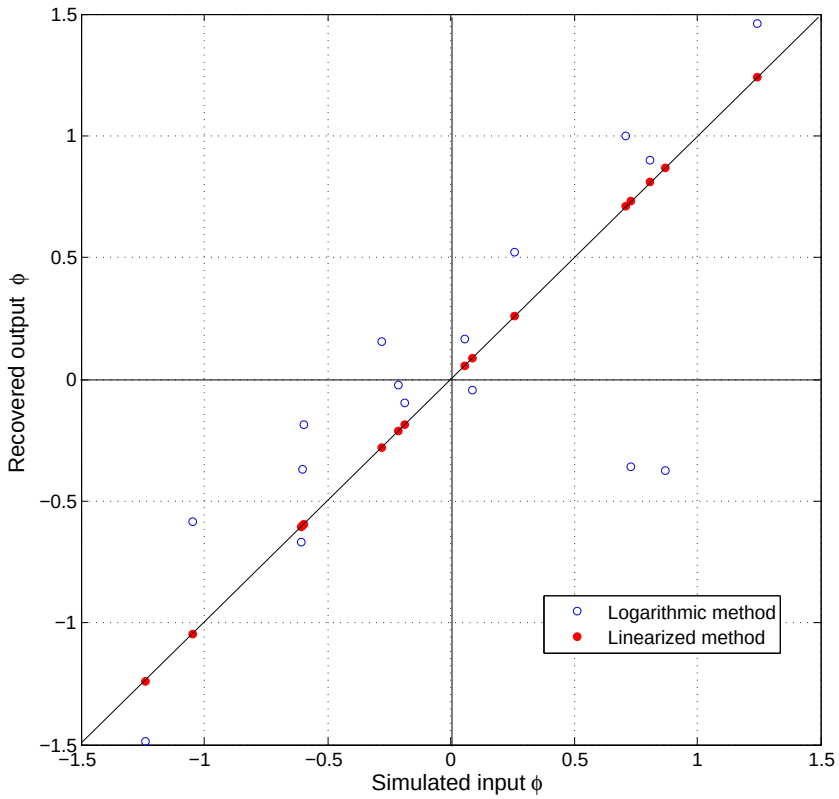

Figure 2-10: Scatter plots of simulated vs. recovered antenna phase parameter φ for a *noiseless* 4 by 4 square array with relatively large phases. The results from the logarithmic implementation described in Section 2.2.3 are shown using open blue circles, whereas the results from the linear implementation described in Section 2.2.7 are shown using solid red circles. Large phases are seen to affect the accuracy of the logarithmic method, but not the linear method.

2.13 can be shown to be unbiased (Tegmark, 1997b), the proof assumes that the noise covariance matrix $\mathbf{N}$ is Gaussian and that the weights used in the fits (encoded by $\mathbf{N}^{-1}$) are independent of the data. In our case both assumptions are violated. While we may assume that n_x and n_y in Equation 2.17b are Gaussian distributed, this is not the case in amplitude/phase angle space. In addition, since the diagonal entries of $\mathbf{N}$ are taken to be $1/|c|$, the measured data enter the least squares fit (Equation 2.13) not just through the data vector $\vec{d}$ but also via $\mathbf{N}^{-1}$, making the fitting process a nonlinear one.

The bias in the method can be easily seen in Figure 2-11, where the blue dots show the simulated results for η, averaged over 90 random ensemble realizations of a 4 by 4 array with an SNR of 2. There is clearly a systematic bias in the recovered output η's.

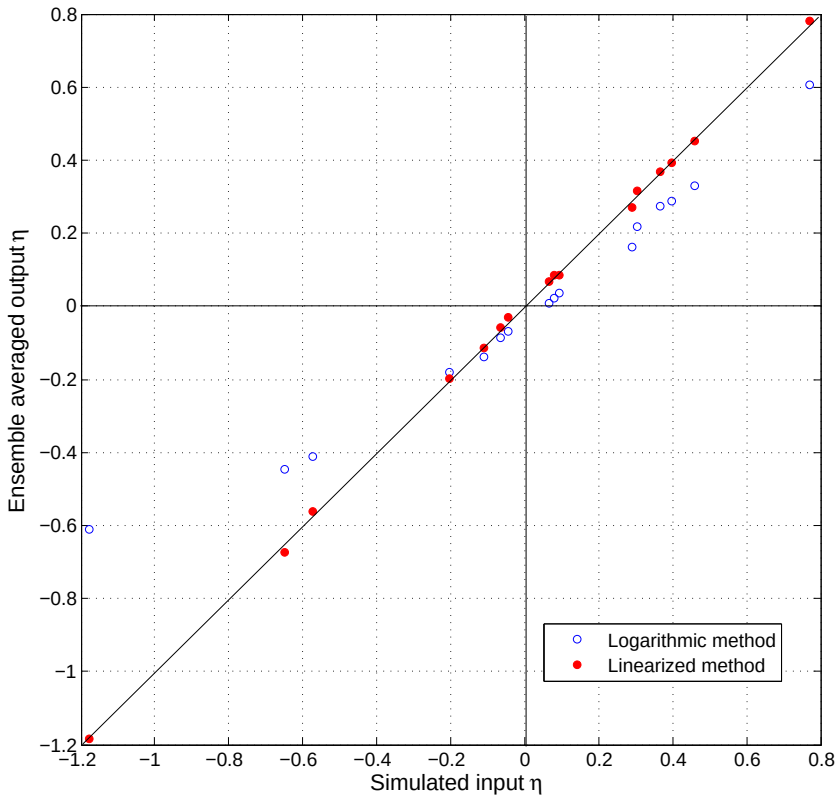

Figure 2-11: Scatter plots of simulated vs. ensemble averaged recovered antenna phase parameter φ for a 4 by 4 square array with a signal-to-noise ratio of 2. The results from the logarithmic implementation described in Section 2.2.3 are shown using open blue circles, whereas the results from the linear implementation described in Section 2.2.7 are shown using solid red circles. The logarithmic algorithm gives a biased result, whereas the linear implementation does not.

2.2.7 The Linearized Approach

Since both of the problems discussed in the previous section arise at least partly from taking the logarithm of Equation 2.6, we now propose a method that does not require that step. As an alternative to taking logarithms, one can instead linearize the equations in Section 2.2.2. We do so by Taylor expanding our expressions about some fiducial guesses η_0, φ_0, and y_0 for the parameters η, φ, and y_0. The gain of a particular antenna, for example, becomes

$$\begin{aligned} g &\equiv e^{\eta+i\varphi} && \text{(2.22a)} \\ &= g_0 + \left.\frac{\partial g}{\partial \eta}\right|_{\eta_0,\varphi_0} (\eta-\eta_0) + \left.\frac{\partial g}{\partial \varphi}\right|_{\eta_0,\varphi_0} (\varphi-\varphi_0) + \ldots && \text{(2.22b)} \\ &= \exp\left(\eta_0 + i\varphi_0\right)\left[1 + (\eta-\eta_0) + i(\varphi-\varphi_0)\right] + \ldots && \text{(2.22c)} \end{aligned}$$

Along with our fiducial guess for the true correlations (the y's), we define guesses for our measured correlations:

$$c_{ij}^0 \equiv y_{i-j}^0 \exp\left(\eta_i^0 - i\varphi_i^0\right) \exp\left(\eta_j^0 + i\varphi_j^0\right). \tag{2.23}$$

In addition, we define the *deviation* δ_{ij} of the guessed correlations from the measured correlations

$$\delta_{ij} \equiv c_{ij} - c_{ij}^0 \tag{2.24}$$

as well as an analogous quantity for the true sky correlations

$$y_{i-j}^1 \equiv y_{i-j} - y_{i-j}^0. \tag{2.25}$$

To calibrate our telescope, we take the measured deviations δ_{ij} as our inputs and compute the corrections to the guessed antenna parameters and guessed true sky correlations that are required to match these deviations. Plugging our definitions and expansions into $c_{ij} = g_i^* g_j y_{i-j}$ gives

$$\delta_{ij} \approx \exp\left(\eta_i^0 - i\varphi_i^0\right) \exp\left(\eta_j^0 + i\varphi_j^0\right) \left[y_{i-j}^1 + y_{i-j}^0 \left(\Delta\eta_i + \Delta\eta_j - i\Delta\varphi_i + i\Delta\varphi_j\right)\right] \tag{2.26}$$

where we have defined $\Delta\eta \equiv \eta - \eta_0$, $\Delta\varphi \equiv \varphi - \varphi_0$, and have discarded second order terms in the small quantities $\Delta\eta$, $\Delta\varphi$, and y^1. The result is a linear system in $\Delta\eta$, $\Delta\varphi$, and y^1, which means we can use the matrix formalism presented in the previous section, the only differences being that the η and φ equations are now coupled, and that the matrix analogous to the $\mathbf{A}$ used in the logarithmic method — which we will call $\mathbf{B}$ for the linearized method — now depends on the fiducial guesses as well as the array layout. A least-squares fit can once again be employed to solve for the $\Delta\eta$'s, the $\Delta\varphi$'s, and y_{i-j}^1. Once the fit has been obtained, it can be used to update our fiducial guesses, and if desired, one can continue to improve the quality of the calibration by repeating the algorithm iteratively — the updated calibration parameters are simply used as the fiducial guesses for the next cycle of fitting. Note that unlike before, it is unnecessary to use a weighted fit *i.e.* one can set $\mathbf{N} = \mathbf{I}$ in Equation 2.13. This is because the linearized algorithm works with visibilities directly (as opposed to, say, their logarithm or some other function of the visibilities), and since our assumption of uncorrelated baseline errors means that individual baselines have the same noise level, all correlations should be weighted the same in the fitting.

In Figure 2-12, we show plots of simulated and recovered visibilities on the complex plane as one steps through each iteration of the algorithm. The top left plane shows

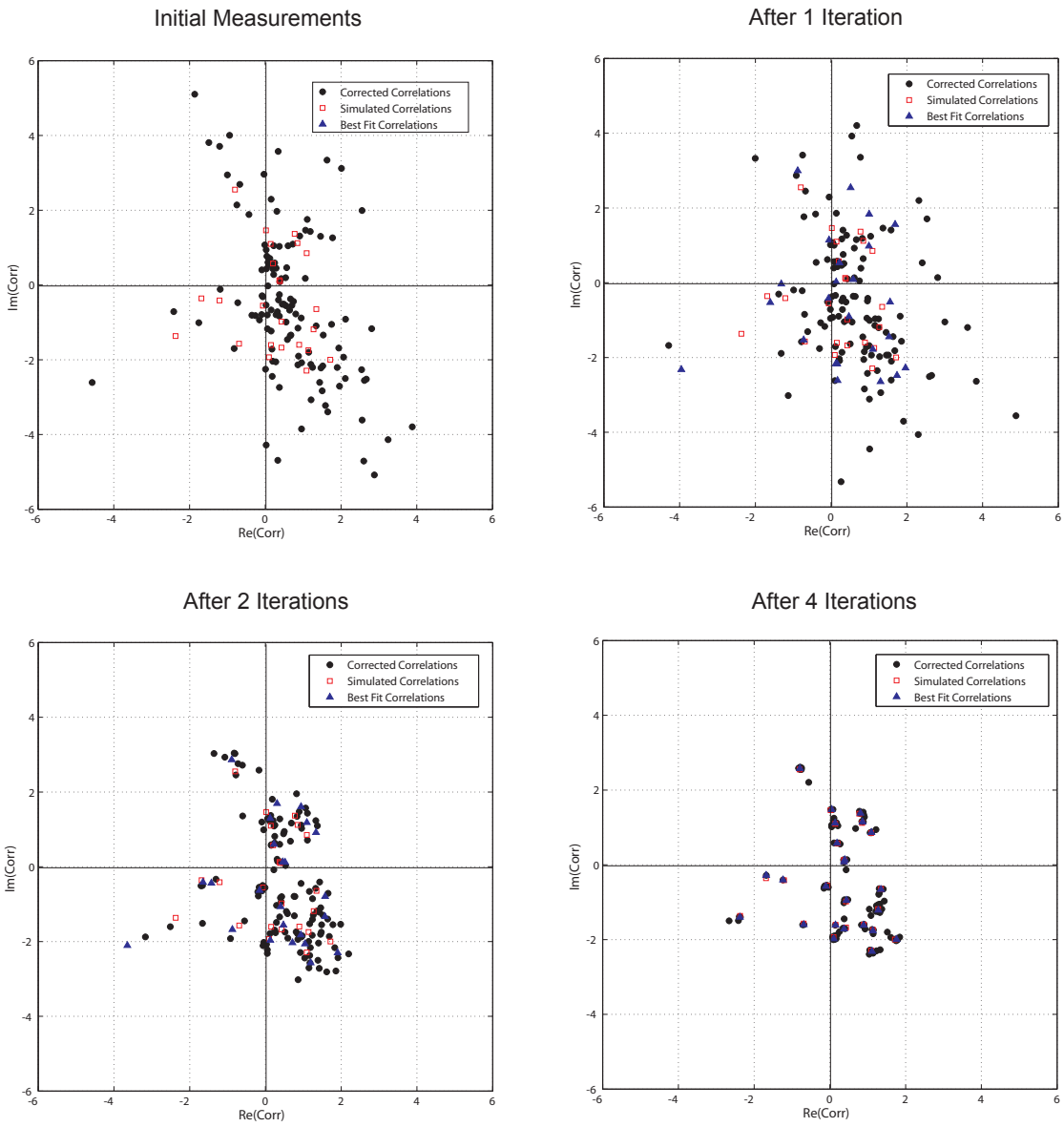

Figure 2-12: Visibility simulations, measurements, and estimated visibilities plotted on complex planes. The top left plane shows the simulated and measured visibilities before any calibration scheme has been applied. The top right plane shows the simulated visibilities, corrected visibilities, and best fit visibilities after one iteration. The bottom planes show the same quantities after two and four iterations respectively.

the simulated correlations as well as the measured correlations. After one iteration, we plot the corrected visibilities as well as the estimated true visibilities in the top right. The recovered visibilities are seen to be still quite different from the simulated ones. The bottom left and bottom right planes show the second and fourth iterations respectively, and one can see that the recovered visibilities converge toward their simulated values.

Scatter plots showing the correlation between the simulated and estimated antenna calibration parameters are qualitatively identical to Figures 2-2 and 2-3 for the logarithmic implementation, so we do not show them here. However, the linearized approach solves the two problems with the logarithmic algorithm discussed in Section 2.2.6. In Figure 2-10, the filled red circles representing the linearized method show a perfect recovery of simulated phases in the noiseless limit. This demonstrates that the linearized algorithm can tolerate arbitrarily large phases, which the logarithmic method could not. Figure 2-11 shows that the ensemble-averaged parameter estimates coming out of the linearized method converge to their true values, demonstrating that the linearized algorithm is unbiased. That this is the case is particularly important if one applies the algorithm to highly noisy systems, for then one can compensate for the low SNR by using massive arrays with large numbers of baselines and by averaging the results over long time periods[12].

2.2.8 Numerical issues for the linearized approach

Although the linearized algorithm evades the problems with the logarithmic approach, it is slightly more complicated to put into practice. Below we discuss two problems that may arise in the numerical implementation of the method and how to solve them.

Convergence

Bad fiducial guesses can result in misfits, although such fits can be easily identified and improved upon. Given that our basic measurement equation (Equation 2.3) is nonlinear in our parameters, the linearization that we have presented in this section can only be expected to stably yield good fits when our fiducial guesses are reasonably close to their true values. In a naive implementation of the algorithm, bad fiducial guesses can result in an inability to iterate out of a local maximum in goodness-of-fit. However, bad fits can be readily identified by simply computing the goodness-of-fit

[12]That is, provided any natural drift in the calibration parameters occur over longer timescales than the required averaging time.

parameter, and if necessary, one can repeat the fit with a better search algorithm such as simulated annealing to find the correct maximum.

Another way to avoid misfits is to use the results of the logarithmic algorithm as our initial guess for the linearized method. Even though the phase calibration has been shown to be unreliable when the phases are large, the amplitude calibration can still be trusted since it comes from solving a completely separate system of equations. The fact that the amplitudes estimated using the logarithmic method are biased is not an issue, since our fiducial guess serves merely as a starting point for our iterative linearized scheme.

An attractive approach to implementing our method in practice could be to update the calibration on the timescale over which the calibration parameters are expected to drift by non-negligible amounts (say once per second or once per minute), using the previously determined calibration parameters as the initial guess and performing merely a single iteration each time. Avoiding multiple iterations in this fashion has the advantage that the noise from any given observation affects the output only linearly, thus avoiding nonlinear mappings that can potentially bias the results. The optimal duration between calibrations should be determined to minize the expected errors: if we calibrate too frequently, the calibration parameters get unnecessarily noisy because so little data is used to determine them, while if we calibrate too rarely, the true calibration parameters can drift appreciably before we recalibrate. An alternative way to determine this tradeoff is to re-estimate the calibration parameters more frequently than needed and then smooth the time series, each time replacing the calibration parameter vector actually used by p times the new estimate and $(1-p)$ times the last estimate, and determining which choice for $p \in (0,1)$ minimizes the rms calibration errors.

Computational Cost

The linear method is more computationally intensive than the logarithmic one, especially if more than one iteration is required. However, the computational cost is still acceptable if one takes advantage of the sparseness of $\mathbf{B}$. At first sight, the setting of $\mathbf{N} = \mathbf{I}$ in Equation 2.13 for the linearized method but not for the logarithmic method would seem to make the linear method computationally quicker. However, the presence of a non-trivial $\mathbf{N}$ in the logarithmic method does not increase the computational cost, for in the absence of cross-talk $\mathbf{N}$ is diagonal. The computational cost of its inversion is therefore negligible, and its presence in Equation 2.13 simply changes the weighting of matrix elements when finding matrix products.

That the linear method is slightly more computationally intensive than the logarithmic method is due to the fact that the matrices for the linear method are larger than those in the logarithmic method. This is because the amplitude and phase calibrations do not decouple in the linearized scheme, and all the parameters must be solved for in one big system. In particular, all vectors are now twice as long as before, since a system where gains and phases are coupled is mathematically equivalent to one where the real and imaginary parts of all complex numbers are dealt with together. This has its greatest computational impact on the matrix inversion of $\mathbf{B^t B}$, which can scale as strongly as the vector length cubed if one simply implements the most straightforward matrix inversion routines. Since this inversion must be performed every time the fiducial values (which are part of $\mathbf{B}$) are updated, the computational time also scales linearly with the number of iterations, and so naively the linearized method seems rather expensive. However, the matrix $\mathbf{B}$ is by construction always sparse, which means the matrix inversion can be performed more efficiently. As an example, one may use the conjugate gradient method, where the sparseness of $\mathbf{B}$ means that the inversion scales linearly with vector length, improving the numerical cost scaling from $\mathcal{O}(N^3)$ to $\mathcal{O}(N)$. This is very similar to the mapmaking approach successfully implemented by the WMAP team in Hinshaw et al. (2003).

It should also be noted that if only one iteration is required and the gain parameters drift only by small amounts over time, then the linearized method can be extremely efficient over long integrations. This is because a *series* of calibrations can then be performed using the same fiducial values for the antenna parameters, and thus the matrix $[\mathbf{B^t B}]^{-1}\mathbf{B^t}$ never changes and can be computed once and for all[13]. One then simply stores this matrix, which is applied to the data whenever calibration is required. Note that this cannot be done with the logarithmic approach, since the relevant matrix there is $[\mathbf{A^t N^{-1} A}]^{-1}\mathbf{A^t N^{-1}}$ and from Equation 2.17c we see that $\mathbf{N}$ changes as the measured correlations change, so the combination $[\mathbf{A^t N^{-1} A}]^{-1}\mathbf{A^t N^{-1}}$ must be recomputed every time one would like to calibrate.

If an instrument with N antennas possesses a very large number of redundant baselines (like any reasonably large omniscope, say, where almost all of the $\sim N^2/2$ baselines are redundant), Figure 2-9 suggests that the errors will be extremely small, especially if the signal-to-noise ratio is high. With such an instrument, one will be able to calibrate accurately by utilizing merely a small subset of the redundant baselines;

[13]In principle, one also requires that transients are not present in the data, since transients may cause sudden changes in the visibilities y_{ij} and thus one cannot assume that the same fiducial guesses are good ones over all time. In practice, however, transients can be easily identified in the data, and one can simply ignore the data taken over time intervals where transients dominate.

for example, if each antenna is correlated with merely 20 others, the resulting system of equations will still be safely overdetermined while keeping the size of the **B**-matrix and the computational cost reasonable. This is important for any array utilizing fast Fourier transforms for the correlation process, since its attractive $N \log_2 N$ cost scaling would be lost if an $\mathcal{O}(N^2)$ correlator were needed for calibration.

In summary, our proposed calibration method can be performed rather cheaply, especially when one compares the computational cost to that of initially computing the interferometric correlations. To see this, note that one must compute correlations roughly once every 10^{-8} seconds for an interferometer operating at a frequency of $\sim 100\,\mathrm{MHz}$. This is much more expensive than the calibration step (especially if one incorporates the computational tricks suggested in this section), which only needs to be performed, say, once every second or minute. In other words, the correlation needs to be repeated on the timescale on which the electromagnetic waves oscillate whereas the calibration only needs to be repeated the time scale on which the calibration parameters drift.

2.3 A Generalized Calibration Formalism: Calibration as a Perturbative Expansion

In previous sections, the calibration algorithms that were discussed represented two idealized limits: with the bright point source calibration (Section 2.2.1), the existence of an *isolated* bright point source was key, while with redundant baseline algorithms (Section 2.2.2) a large number of baselines of equal length and orientation were required. Both algorithms were "zeroth order algorithms" in that they required exact adherence to these requirements. In this section, we perturb around these idealized assumptions, and show that point source calibration and redundant calibration are simply two extremes in a generalized continuum of calibration schemes.

Consider the sky response (*i.e.* the visibility) from a baseline $\mathbf{b}$ formed by two antennas at locations $\mathbf{r}_i$ and $\mathbf{r}_j$ (*i.e.* $\mathbf{b} \equiv \mathbf{r}_i - \mathbf{r}_j$). Assuming that our array is coplanar, we have

$$c(\mathbf{b}) = \int_{sky} g_i^* g_j \frac{J(\boldsymbol{\theta})B(\boldsymbol{\theta})}{\sqrt{1-\theta_x^2-\theta_y^2}} \exp\left(-\frac{i2\pi\mathbf{b}\cdot\boldsymbol{\theta}}{\lambda}\right) d^2\boldsymbol{\theta}, \tag{2.27}$$

where $J(\boldsymbol{\theta})$ is the intensity of the sky signal from the $\boldsymbol{\theta}$ direction, and $B(\boldsymbol{\theta})$ is the primary beam. Suppose we now define an effective sky signal $I(\boldsymbol{\theta}) \equiv \frac{J(\boldsymbol{\theta})B(\boldsymbol{\theta})}{\sqrt{1-\theta_x^2-\theta_y^2}}$.

Writing this in terms of its Fourier space description

$$I(\boldsymbol{\theta}) = \int \tilde{I}(\mathbf{k}) \exp(i2\pi\mathbf{k}\cdot\boldsymbol{\theta}) d^2\mathbf{k} \tag{2.28}$$

and inserting this into Equation 2.27, we get

$$c(\mathbf{b}) = g_i^* g_j \tilde{I}\left(\frac{\mathbf{b}}{\lambda}\right), \tag{2.29}$$

which is a well-known result: baselines of an interferometer probe different Fourier modes of the effective sky signal.

Now suppose that our interferometer possesses a number of nearly-redundant baselines. In other words, suppose that there exist a set of vectors $\mathbf{b}_0^\alpha = \{\mathbf{b}_0^1, \mathbf{b}_0^2, \dots\}$ on the uv plane that baselines tend to cluster around. Nearly-redundant baselines are considered part of the same cluster, and are associated with a single $\mathbf{b}_0^\alpha$. On the other hand, any baselines that are isolated on the uv plane are associated with "their own" $\mathbf{b}_0^\alpha$. Thus, for an array with N antennas the index α runs from 1 to $N(N-1)/2$ if there are zero near or exact redundancies in the baselines, and to some number substantially smaller than $N(N-1)/2$ if there are many near or exact redundancies. In other words, we have grouped the baselines into nearly-redundant clusters.

We now re-parameterize our baseline distribution in terms of $\delta\mathbf{b}_{ij} \equiv \mathbf{b}_{ij} - \mathbf{b}_0^\alpha$, which measures each baseline's deviation from perfect redundancy. Since $\delta\mathbf{b}_{ij}$ is a small quantity, we can Taylor expand Equation 2.29:

$$\begin{aligned} c_{ij} &= g_i^* g_j \tilde{I}\left(\frac{\mathbf{b}_0^\alpha + \delta\mathbf{b}_{ij}}{\lambda}\right) && \text{(2.30a)} \\ &\approx g_i^* g_j \left[\tilde{I}\left(\frac{\mathbf{b}_0^\alpha}{\lambda}\right) + \nabla_{\mathbf{u}}\tilde{I}\Big|_{\mathbf{b}=\mathbf{b}_0^\alpha} \cdot \frac{\delta\mathbf{b}_{ij}}{\lambda} + \dots\right], && \text{(2.30b)} \end{aligned}$$

where $\nabla_{\mathbf{u}}$ denotes a two-dimensional gradient in the uv plane. If we define $c_0^\alpha \equiv \tilde{I}\left(\frac{\mathbf{b}_0^\alpha}{\lambda}\right)$, we can rewrite our equation as

$$c_{ij} \approx g_i^* g_j c_0^\alpha \left(1 + \mathbf{h}_0^\alpha \cdot \frac{\delta\mathbf{b}_{ij}}{\lambda} + \dots\right), \tag{2.31}$$

where $\mathbf{h}_0^\alpha \equiv \nabla \ln \tilde{I}\Big|_{\mathbf{b}=\mathbf{b}_0^\alpha}$.

For algebraic brevity, we proceed using the logarithmic method of Section 2.2.3, the generalization to the linearized method being straightforward. Taking logarithms

of both sides yields

$$\begin{aligned} \ln c_{ij} &\approx (\eta_i + \eta_j) + i(\varphi_j - \varphi_i) + \ln c_0^\alpha + \ln\left(1 + \mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}\right) \\ &\approx (\eta_i + \eta_j) + i(\varphi_j - \varphi_i) + \ln c_0^\alpha + \mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}, \end{aligned} \tag{2.32}$$

where we have assumed that $\mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}$ is small in order to expand the last logarithm.

Before we proceed, it is worth examining the precise conditions under which our Taylor series expansions are valid. The crucial quantity is $\mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}$, which will be small if either $\mathbf{h}_0^\alpha$ or $\frac{\delta \mathbf{b}_{ij}}{\lambda}$ are small. Plugging Equation 2.28 into our definition of $\mathbf{h}$, we have

$$\mathbf{h} = \frac{\nabla \tilde{I}}{\tilde{I}} = -i2\pi \frac{\int \boldsymbol{\theta} I(\boldsymbol{\theta}) \exp(-i2\pi \mathbf{u} \cdot \boldsymbol{\theta}) d^2\boldsymbol{\theta}}{\int I(\boldsymbol{\theta}) \exp(-i2\pi \mathbf{u} \cdot \boldsymbol{\theta}) d^2\boldsymbol{\theta}}, \tag{2.33}$$

which is a quantity whose magnitude is at most the size of the primary beam in radians. This dependence on the primary beam means that $\mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}$ can be made to be small in different ways depending on the instrument one is considering:

1. For widefield instruments, the primary beam width is on the order of 1 radian, so one requires $|\delta \mathbf{b}_{ij}| \ll \lambda$.

2. For instruments with narrow primary beams such as the VLA, we have $|\mathbf{h}| \ll 1$, so the deviations $\delta \mathbf{b}_{ij}$ from perfect redundancies need not be small compared to the wavelength.

If the conditions listed above are satisfied, then calibrating our radio telescope is similar to the zeroth-order case, in the sense that it is once again tantamount to solving Equation 2.32. Here, however, we are *not* guaranteed to have enough equations to solve for all the relevant variables. In addition to the $2N$ antenna calibration parameters (η's and φ's), we now potentially have up to $3N(N-1)$ numbers to solve for — the $\ln c_0^\alpha$ term contributes up to $N(N-1)$ numbers, since as we discussed above α can run to $N(N-1)/2$ in a worst-case scenario, and specifying a single $\ln c_0^\alpha$ requires two numbers: a real part and an imaginary part; the $\mathbf{h}_0^\alpha$ term contributes up to $2N(N-1)$ numbers — twice that required to specify the $\ln c_0^\alpha$ terms, because $\mathbf{h}_0^\alpha$ is the logarithmic gradient of c_0^α in the uv plane, which is two dimensional. The $\delta \mathbf{b}_{ij}$'s do not need to be solved for since they can be computed from the layout of the array. Putting this all together, we see that in general there can be up to $3N(N-1)$ unknowns.

To properly solve Equation 2.32, one must therefore find ways to reduce the range

of the index α so that the number of equations exceeds the number of unknowns. In the next few subsections, we will consider different scenarios in which this is the case.

2.3.1 Zeroth Order: Perfect Point Sources or Perfect Redundancies

The zeroth order solution to Equation 2.32 involves setting up situations where the $\mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda}$ term can be set to zero. There are two ways to accomplish this:

1. **Design an array with perfect redundancies, so that $\delta \mathbf{b}_{ij} = 0$.** The form of $\mathbf{h}_0^\alpha$ is then irrelevant (which is another way of saying that we can calibrate with any sky signal), and the only requirement is for the number of *unique* baselines to be small enough that the maximum possible α is modest and there are only a small number of c_0^α terms to fit for. This is precisely the redundant calibration algorithms discussed in Sections 2.2.2 to 2.2.8.

2. **Have some knowledge of the sky signal used for calibration, so that c_0^α and $\mathbf{h}_0^\alpha$ are known.** For example, if one uses a phased array to image a point calibrator source, then c_0^α becomes a real constant independent of α, while the h_0^α's all vanish. This is the point source calibration discussed in Section 2.2.1.

2.3.2 First Order: Unknown Sources and Near Perfect Redundancies

If one is unable to satisfy the conditions listed in Section 2.3.1, it may still be possible to calibrate by solving for c_0^α and $\mathbf{h}_0^\alpha$ as well as all the calibration parameters simultaneously. For example, taking the real part of Equation 2.32 gives

$$\ln |c_{ij}| = (\eta_i + \eta_j) + \ln |c_0^\alpha| + \mathrm{Re}(h_{0u}^\alpha)\frac{\delta b_{ij}^u}{\lambda} + \mathrm{Re}(h_{0v}^\alpha)\frac{\delta b_{ij}^v}{\lambda}, \tag{2.34}$$

which can be written as a matrix system:

$$\begin{pmatrix} \ln|c_{12}| \\ \ln|c_{23}| \\ \ln|c_{34}| \\ \ln|c_{45}| \\ \ln|c_{56}| \\ \vdots \\ \ln|c_{29}| \\ \ln|c_{19}| \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 & 0 & \dots & \frac{\delta b^u_{12}}{\lambda} & 0 & \dots & \frac{\delta b^u_{12}}{\lambda} & 0 & \dots \\ 0 & 1 & \dots & 1 & 0 & \dots & \frac{\delta b^u_{23}}{\lambda} & 0 & \dots & \frac{\delta b^u_{23}}{\lambda} & 0 & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & \frac{\delta b^u_{34}}{\lambda} & 0 & \dots & \frac{\delta b^u_{34}}{\lambda} & 0 & \dots \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & \frac{\delta b^u_{45}}{\lambda} & \dots & 0 & \frac{\delta b^u_{45}}{\lambda} & \dots \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & \frac{\delta b^u_{56}}{\lambda} & \dots & 0 & \frac{\delta b^u_{56}}{\lambda} & \dots \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \\ 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \ln|c^1_0| \\ \ln|c^2_0| \\ \vdots \\ \mathrm{Re}(h^1_{0u}) \\ \mathrm{Re}(h^2_{0u}) \\ \vdots \\ \mathrm{Re}(h^1_{0v}) \\ \mathrm{Re}(h^2_{0v}) \\ \vdots \end{pmatrix}, \tag{2.35}$$

where for illustrative purposes we have chosen a nine-antenna system with the first three baselines almost-redundant. This set of equations can again be solved using Equation 2.13, provided the design of the interferometer is such that the baselines can be grouped into a relatively small number of near-redundant clusters. Fundamentally, the method here is the same as it has been in all previous sections — we have simply written the interferometer response as a linear function of the calibration parameters and the sky signal, and have found ways to reduce the number of parameters (whether by mathematical approximation, interferometer design, or careful selection of calibration source) so that the equations are solvable. In the first order expansion considered in this section, if r denotes the number of $\mathbf{b}^\alpha_0$'s with two or more baselines clustered around them and l denotes the number of completely isolated baselines, the calibration can be solved for provided our array satisfies

$$\frac{N(N-1)}{2} > N + 3r + l, \tag{2.36}$$

where as usual N is the number of antenna elements in the array. The left hand side is simply the number of measured complex correlations, while the right hand side counts the number of complex unknowns we need to solve for: N complex antenna gains, r complex visibilities at each cluster center, $2r$ complex numbers quantifying deviations in the complex visibilities due to departures from perfect redundancy in the two directions of the uv plane, and l complex visibilities for the isolated baselines.

For the $s \times s$ regular square arrays simulated in Section 2.2, the condition is satisfied if $s \geq 4$.

Like before, however, one must be cognizant of possible degeneracies in the matrix system. In particular, consider a cluster comprising of one or two baselines. In fitting for this cluster's visibility, we have one or two inputs (*i.e.* correlations) but need to solve for three parameters: c_0^α, h_{0u}^α, and h_{0v}^α. Our matrix system thus becomes singular. The same problem arises if either δb_{ij}^u or δb_{ij}^v are identical for a large number of baselines in a particular cluster. For instance, if all but one of the baselines in a cluster are perturbed in precisely the same way, then as far as the cluster goes we effectively only have two different baselines and cannot fit for three parameters. These degeneracies can be dealt with in a similar fashion as before: one simply imposes extra constraints on the system of equations, like requiring that h_{0u}^α and h_{0v}^α be zero for single-baseline clusters.

Simulation results for a 4×4 square array with Gaussian antenna position errors are shown in Figure 2-13, where we show the root mean square errors in the recovered gains $\Delta\eta$ as a function of the spread σ_b in the Gaussian from which the position errors are drawn. The simulations were run with precisely the same parameters as those in Section 2.2, except with zero instrumental noise, so any non-zero error is due purely to position errors (*i.e.* to the non-perfect redundancy of baselines). Plotted using the solid circles are the results that one obtains if one simply ignores the fact that the baselines are not perfectly redundant *i.e.* if one applies a zeroth order algorithm based on Equation 2.6. The errors are clearly linear in σ_b. Correcting for baseline errors to first order by implementing an algorithm based on Equation 2.31 gives errors that are smaller and quadratic in σ_b. It is also clear from the plot that the wider the primary beam of an instrument, the greater the errors in the calibration. Formally, this is because the beam width is directly proportional to the size of the linear baseline error correction term, as we noted above. This result can also be understood intuitively by considering the effect that a primary beam has on uv visibilities. A narrow primary beam acts as a wide convolution kernel on the uv plane, which effectively "averages" together visibilities over a wide neighborhood. This smoothes the visibilities on the uv plane, making the errors incurred by baseline errors more amenable to a perturbative correction.

In principle, there is no need to stop at the first order correction. One could, for instance, correct baseline errors to second order by using the following equation as

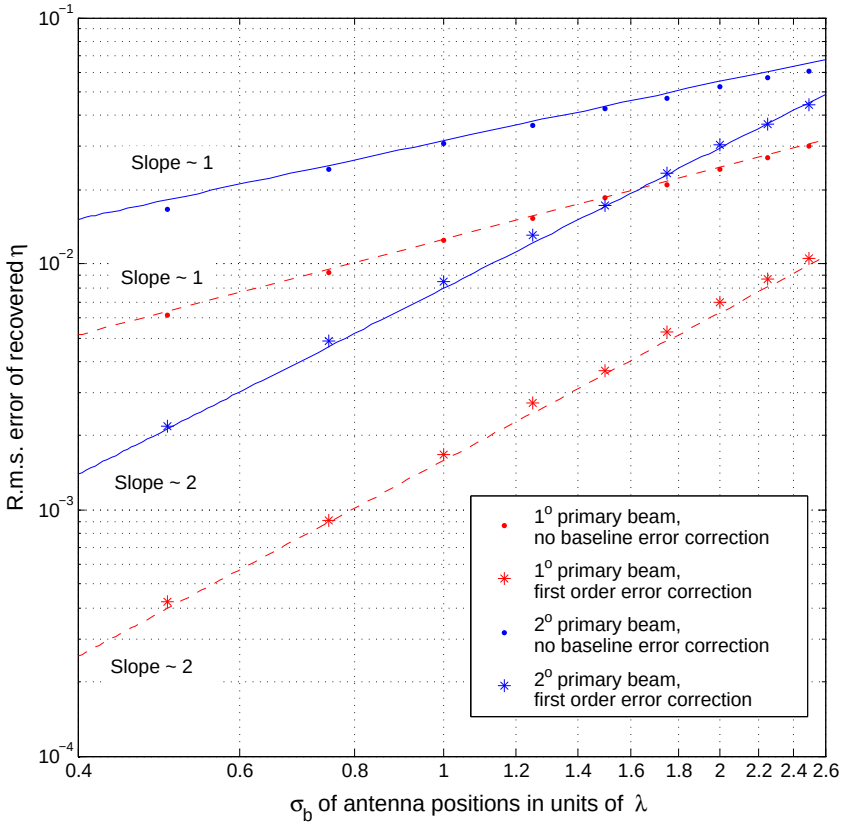

Figure 2-13: R.m.s. errors in recovered η for one antenna in a noiseless 4×4 square array as a function of antenna position spread σ_b. Shown in red dashed lines are the results for a primary beam with width 1°, while the results for a primary beam with width 2° are shown using the blue solid lines. The solid circles denote results from an algorithm that does not correct for baseline position errors, and the stars denote results from an algorithm that corrects these errors to first order.

the basic measurement equation:

$$c_{ij} \approx g_i^* g_j c_0^\alpha \left(1 + \mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda} + \frac{\delta \mathbf{b}_{ij}^t}{\lambda} \cdot \mathbf{j}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda} \ldots\right), \tag{2.37}$$

where $\mathbf{h}_0^\alpha \equiv \nabla \ln \tilde{I}\Big|_{\mathbf{b}=\mathbf{b}_0^\alpha}$ as before, and $\mathbf{j}_0^\alpha$ is a symmetric matrix of second derivatives of $\tilde{I}$ evaluated at $\mathbf{b}_0^\alpha$ and normalized by $\tilde{I}\left(\frac{\mathbf{b}_0^\alpha}{\lambda}\right)$. It is important to note, however, the somewhat counterintuitive fact that for a realistic array (*i.e.* one with instrumental noise), the errors in the recovered gain parameters may *increase* as one corrects to increasing order in baseline position error $\delta \mathbf{b}_{ij}$. This is because the higher the order that one corrects to, the greater the number of parameters that one must solve for. The number of correlation measurements, however, remains the same as before

(with $N(N-1)/2$ of them), so the system of equations becomes less constrained. This makes the calibration algorithm more susceptible to noise, which can negate the benefit that one obtains by correcting baseline position errors to higher order. Put another way, when we introduce a large number of parameters in an attempt to correct for baseline position errors, we run the risk of over-fitting the instrumental noise instead of averaging it down using redundant baselines.

In running the baseline error simulations, we find the errors in recovered φ to be much worse than those in the recovered η's. This is likely due to the fact that a small perturbation in antenna position can result in large changes in phases, which (as explained in Section 2.2.6) can cause problems in a logarithmic implementation.

The reader should also note that because quantities like $\mathbf{h}_0^\alpha$ are dependent on the sky signal, we find that the calibration errors induced by baseline position errors cannot be easily captured by simple formulae analogous to Equation 2.21. An extreme example of this was already discussed above in Section 2.3.1, where we saw that the correction terms for baseline errors approach zero as the sky becomes increasingly dominated by a single bright point source. The values on the vertical axis of Figure 2-13 are thus not generically applicable; in general, one must simulate the relevant array arrangement and sky signal to estimate the errors incurred by antenna position errors.

2.3.3 First Order: Non-coplanar Arrays

The methods described above can also be used to calibrate radio interferometers even if there are slight deviations from planarity. Suppose the antenna elements of our array are all almost (but not quite) located at a height of $z = 0$. The slight deviations from perfect coplanarity of the antenna elements result in slight deviations from coplanarity of the baselines, which we denote δb_z^{ij}. With these deviations, Equation 2.27 becomes

$$c(\mathbf{b}) = \int_{sky} g_i^* g_j \frac{J(\boldsymbol{\theta})B(\boldsymbol{\theta})}{\sqrt{1-\theta_x^2-\theta_y^2}} \exp\left(-\frac{i2\pi\mathbf{b}\cdot\boldsymbol{\theta}}{\lambda} - \frac{i2\pi}{\lambda}\delta b_z^{ij}\sqrt{1-\theta_x^2-\theta_y^2}\right) d^2\boldsymbol{\theta}. \quad (2.38)$$

Like before, this equation can be manipulated into a form conducive to calibration if one (or both) of the following conditions are met:

1. **Narrow primary beam**. With a narrow primary beam, one has $\sqrt{1-\theta_x^2-\theta_y^2} \approx 1-\frac{\theta_x^2}{2}-\frac{\theta_y^2}{2}\cdots \approx 1$. The part of the expression corresponding to the non-coplanar

correction thus factors out of the integral, giving

$$\begin{aligned} c(\mathbf{b}) &= \exp\left(-\frac{i2\pi}{\lambda}\delta b_z^{ij}\right)\int_{sky} g_i^* g_j \frac{J(\boldsymbol{\theta})B(\boldsymbol{\theta})}{\sqrt{1-\theta_x^2-\theta_y^2}}\exp\left(-\frac{i2\pi\mathbf{b}\cdot\boldsymbol{\theta}}{\lambda}\right)d^2\boldsymbol{\theta} \\ &= \exp\left(-\frac{i2\pi}{\lambda}\delta b_z^{ij}\right) g_i^* g_j \tilde{I}\left(\frac{\mathbf{b}}{\lambda}\right), \end{aligned} \tag{2.39}$$

where we have adopted the notation defined in Equation 2.28. This illustrates the fact that so long as the primary beam is narrow, deviations from coplanarity simply results in phase shifts in the measured correlations. Aside from these phase shifts, Equation 2.39 looks completely identical to Equation 2.29, and thus the calibration of a non-coplanar array is no harder than the calibration of a planar array when the primary beam is narrow. One simply defines adjusted correlations $\tilde{c} \equiv c\exp\left(-i\frac{2\pi}{\lambda}\delta b_z^{ij}\right)$ and the ensuing analysis proceeds like before, with no additional computational cost.

2. **Near coplanar array.** If the array is close to coplanar ($\delta b_z^{ij} \ll 1$), the calibration can still be performed, but this time at a slight increase in computational cost. With near coplanarity, Equation 2.38 can be Taylor expanded in δb_z:

$$\begin{aligned} c(\mathbf{b}) &= \int_{sky} g_i^* g_j \frac{J(\boldsymbol{\theta})B(\boldsymbol{\theta})}{\sqrt{1-\theta_x^2-\theta_y^2}}\exp\left(-\frac{i2\pi\mathbf{b}\cdot\boldsymbol{\theta}}{\lambda}\right)\left(1-\frac{i2\pi}{\lambda}\delta b_z^{ij}\sqrt{1-\theta_x^2-\theta_y^2}\right)d^2\boldsymbol{\theta} \\ &= g_i^* g_j \tilde{I}\left(\frac{\mathbf{b}}{\lambda}\right) - i g_i^* g_j \frac{2\pi}{\lambda}\delta b_z^{ij}\tilde{K}\left(\frac{\mathbf{b}}{\lambda}\right), \end{aligned} \tag{2.41}$$

where $\tilde{I}$ is defined as before in Equations 2.27 and 2.28 and

$$\tilde{K}(\mathbf{k}) \equiv \int J(\boldsymbol{\theta})B(\boldsymbol{\theta})\exp(-i2\pi\mathbf{k}\cdot\boldsymbol{\theta})d^2\boldsymbol{\theta}. \tag{2.42}$$

Perturbing this equation around perfectly redundant baselines using exactly the same methods as those employed between Equations 2.30a and 2.31, one obtains

$$c(\mathbf{b}) = g_i^* g_j\left[c_0^\alpha\left(1+\mathbf{h}_0^\alpha\cdot\frac{\delta\mathbf{b}_{ij}}{\lambda}\right) - i2\pi\frac{\delta b_z^{ij}}{\lambda}d_0^\alpha\right], \tag{2.43}$$

where $d_0^\alpha \equiv \tilde{K}\left(\frac{\mathbf{b}_0^\alpha}{\lambda}\right)$. Note that whereas $\tilde{I}$ is expanded to *first* order in $\delta\mathbf{b}$, we are only required to keep the zeroth order term for $\tilde{K}$ since any linear term in $\delta\mathbf{b}$ would be multiplied by δb_z and end up a second order term. Proceeding as

before, the analog of Equation 2.32 takes the form

$$\ln c_{ij} \approx (\eta_i + \eta_j) + i(\varphi_j - \varphi_i) + \ln c_0^\alpha + \mathbf{h}_0^\alpha \cdot \frac{\delta \mathbf{b}_{ij}}{\lambda} + i2\pi q_0^\alpha \frac{\delta b_z^{ij}}{\lambda}, \tag{2.44}$$

where $q_0^\alpha \equiv \ln(d_0^\alpha / c_0^\alpha)$. From here, we can once again form a linear system of equations that can be solved to yield the calibration parameters, the only difference being that we must solve for yet one more parameter per cluster of baselines.

2.4 Conclusions

Redundant calibration schemes calibrate radio interferometers by taking advantage of the fact that — once calibrated — redundant sets of baselines should yield identical visibility measurements regardless of what the sky looks like. In this chapter we have performed simulations to examine the error properties of redundant calibration, and have found that the usual logarithmic implementations are statistically biased. The linearized implementation introduced in Section 2.2.7, on the other hand, is both unbiased and computationally feasible, being far less computationally intensive than the standard initial step of computing the signal correlations.

Though for many of our simulations we deliberately chose unfavorable SNRs to exaggerate the effects of instrumental noise, the errors are seen to scale inversely with SNR, which means that Equation 2.21 and Figure 2-9 predict that redundant baseline calibration can yield high precision calibration parameters in a realistic application. Moreover, the fact that the linearized method of Section 2.2.7 is unbiased means that even if the SNR *were* unfavorable, one could simply build larger arrays and average over long time periods, in principle suppressing calibration errors to arbitrarily low levels.

We have also shown that both point source calibration and redundant calibration can be considered special cases within a generalized framework of algorithms, all of which find ways to reduce the number of calibration and visibility parameters so that there are few enough of them to be solved for using the $N(N-1)/2$ measurement equations. This can be done either by making *a priori* assumptions about the calibration sources, or by taking advantage of baseline redundancy. If one has many more constraints than unknowns, non-exact redundancy and non-coplanarity can also be taken into account, which may be essential as the precision requirements for calibration become more and more stringent. That redundant baseline calibration appears

able to satisfy such requirements while being computationally feasible is encouraging for large radio arrays, suggesting that calibration will not limit the vast scientific potential of 21 cm tomography.

Chapter 3

How well can we measure and understand foregrounds with 21 cm experiments?

3.1 Introduction

In this chapter, we turn to the issue of foreground contaminants. As we mentioned in previous chapters, foreground contamination (both from within our galaxy and from extragalactic sources) is expected to be at least four orders of magnitude brighter than the expected cosmological signal (de Oliveira-Costa et al., 2008), so any viable data analysis scheme must also contain a robust foreground subtraction algorithm.

Foreground subtraction is a problem that has been studied extensively in the literature, and many schemes have been proposed for tackling the issue. Early ideas focused on using the angular structure of the foregrounds to separate them from the cosmological signal (Di Matteo et al., 2002, 2004; Oh & Mack, 2003; Santos et al., 2005; Zaldarriaga et al., 2004), while most recent proposals focus on line-of-sight (*i.e.* spectral) information (Wang et al., 2006; Gleser et al., 2008; Jelic et al., 2008; Harker et al., 2009; Bowman et al., 2009; Liu et al., 2009a,b; Harker et al., 2010; Petrovic & Oh, 2011). By making use of spectral information, these proposals take advantage of the extremely high spectral resolution available in all 21 cm experiments, and indeed it was shown in Liu & Tegmark (2011) that because of the nature of the foregrounds and the instrumental parameters, an optimal estimation of the power spectrum should involve a foreground subtraction scheme that operates primarily using frequency information.

Most line-of-sight proposals so far have been *blind* schemes[1] in the sense that they do not require any prior foreground modeling. All such proposals take advantage of the smooth nature of the foreground spectra, and separate out the rapidly fluctuating cosmological signal by (for instance) subtracting off a predetermined set of low-order polynomials (Bowman et al., 2009; Liu et al., 2009a,b) or by imposing a predetermined filter in Fourier space (Paciga et al., 2010; Petrovic & Oh, 2011). The blind nature of these schemes may seem at first to be an advantage, because the low frequency (~ 50 to $300\,\mathrm{MHz}$) regime of radio foregrounds is as yet fairly unconstrained observationally and most models are based on extrapolations and interpolations from other frequencies, where the instruments are optimized for different science goals. However, even if our foreground models are not entirely accurate initially, a non-blind scheme can always be performed iteratively until the models converge to the true measured foregrounds. Moreover, blind schemes do not improve as one makes better and better measurements of the foregrounds, whereas non-blind schemes will continually improve as measurements place increasingly strong constraints on foreground models. In this chapter we therefore examine the foreground modeling process, and determine whether or not it will be possible to construct a foreground model that is good enough for foreground subtraction purposes if only a small number of independent parameters can be measured. In the spirit of "one scientist's noise is another scientist's signal", we also quantify the ability of 21 cm experiments to place constraints on phenomenological foreground parameters.

The rest of this chapter is organized as follows. In Section 3.2 we introduce our foreground model, and we show in Section 3.3 that the foregrounds can be described to an extremely high precision using just three parameters. This is good news for 21 cm cosmology, for it implies that by measuring just a small handful of empirical parameters, it is possible to construct a foreground model that is sufficiently accurate for high precision foreground subtraction. However, this also implies that the large number of (physically motivated) parameters present in a typical foreground parameterization are redundant and highly degenerate with each other, so the prospects for using 21 cm experiments to gain a detailed understanding of foreground physics are bleak. We show this in Section 3.4, and summarize our conclusions in Section 3.5.

[1] See Liu & Tegmark (2011) for an example that is *not* blind.

3.2 Foreground and Noise Model

For the purposes of this chapter, we limit our analyses to foregrounds in the frequency range 100 to 250 MHz, which roughly speaking covers the "sweet spots" of most 21 cm experiments that are designed to probe the Epoch of Reionization. At these frequencies, there are three dominant sources of foreground contamination (Shaver et al., 1999; Wang et al., 2006):

1. Extragalactic point sources.
2. Synchrotron emission from our Galaxy.
3. Free-free emission from our Galaxy.

Since foreground subtraction is best done using spectral information (Liu & Tegmark, 2011), we will ignore the spatial structure of these foregrounds for this chapter, as well as any polarization structure[2]. We now examine the spectra of each of these sources, and in particular compute their means and variances, which we define in the next section. With some minor modifications, our models are essentially those of Liu & Tegmark (2011).

3.2.1 Definition of the mean and the foreground covariance

Consider a multifrequency sky map of brightness temperature T from a typical 21 cm tomography experiment. The map $T(\hat{\mathbf{r}}, \nu)$ is a function of the direction in the sky $\hat{\mathbf{r}}$ as well as the frequency ν. We define a random function $x(\nu)$ such that

$$x(\nu) \equiv T(\hat{\mathbf{r}}, \nu) \tag{3.1}$$

is the spectrum measured at a randomly chosen pixel centered at $\hat{\mathbf{r}}$. Since the cosmological signal is expected to be dwarfed by the foregrounds, this spectrum will essentially be a spectrum of the foregrounds.

To quantify the statistical properties of this random function, we can calculate its mean and covariance. In principle, doing so requires taking the ensemble averages. In practice, we instead take averages of the pixels of the sky map. The mean is thus given by

$$m(\nu) = \langle x(\nu) \rangle \equiv \frac{1}{\Omega} \int T(\hat{\mathbf{r}}, \nu) d\Omega, \tag{3.2}$$

[2]See de Oliveira-Costa et al. (2008) for a discussion of the spatial structure of foregrounds, and Bernardi et al. (2010) and Geil et al. (2010) for discussions of polarized foregrounds and their removal.

where Ω is the total area covered by the sky map[3]. The statistical covariance can be similarly defined in the usual way:

$$C(\nu, \nu') \equiv \langle x(\nu)x(\nu')\rangle - m(\nu)m(\nu'), \tag{3.3}$$

where

$$\langle x(\nu)x(\nu')\rangle = \frac{1}{\Omega}\int T(\hat{\mathbf{r}}, \nu)T(\hat{\mathbf{r}}, \nu')d\Omega. \tag{3.4}$$

3.2.2 Extragalactic point sources

Extragalactic point sources can be thought of as consisting of two populations. The first consists of bright, isolated point sources that can be resolved by one's instrument. Following previous foreground studies, we assume that these bright point sources have already been removed, prior to the main foreground cleaning step that attempts to subtract off the rest of the foregrounds. Techniques such as forward modeling (Bernardi et al., 2011) and peeling have been explored for this purpose, and peeling simulations suggest that the bright sources can be removed down to $S_{max} \sim 10$ to 100 mJy (Pindor et al., 2010). To be conservative we will use $S_{max} = 100\,\mathrm{mJy}$ in all calculations that follow.

Below S_{max} is the second population of extragalactic point sources, consisting of a "confused" continuum of unresolved point sources. Along a given line-of-sight, we imagine the number of sources to be Poisson distributed with an average of $n\Omega_{pix}$ sources, where n is the number of sources per steradian and Ω_{pix} is the pixel size. We model the spectrum of each source as a power law with a random spectral index α drawn from a Gaussian distribution

$$p(\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left[-\frac{(\alpha-\alpha_0)^2}{2\sigma_\alpha^2}\right], \tag{3.5}$$

where α_0 is the mean spectral index (with its numerical value to be determined later) and $\sigma_\alpha = 0.5$ (Tegmark et al., 2000). If all of the unresolved point sources had the same flux S_* at some fiducial frequency $\nu_* \equiv 150\,\mathrm{MHz}$, the mean intensity would be

$$m^{ps}(\nu) = \left(\frac{A_\nu}{\Omega_{pix}}\right)(n\Omega_{pix}S_*)\int\left(\frac{\nu}{\nu_*}\right)^{-\alpha}p(\alpha)d\alpha, \tag{3.6}$$

[3]Note that this sky map need not cover all 4π steradians of the full sky. For instance, in an estimation of the power spectrum one may choose to use only the data from the cleanest parts of the sky.

where the quantity A_ν/Ω_{pix} converts our expression from having flux units to temperature units, and is given by

$$\left(\frac{A_\nu}{\Omega_{pix}}\right) = \frac{\lambda^2}{2k_B\Omega_{pix}} \approx 1.4 \times 10^{-6} \left(\frac{\nu}{\nu_*}\right)^{-2} \left(\frac{\Omega_{pix}}{1\,\mathrm{sr}}\right)^{-1} \mathrm{mJy}^{-1}\,\mathrm{K}, \tag{3.7}$$

where λ is the wavelength, k_B is Boltzmann's constant, and Ω_{pix} is the pixel solid angle. Of course, in reality the sources do not all have the same flux. To take this into account, we extrapolate a source count distribution from an empirical study done at higher flux levels:

$$\frac{dn}{dS_*} = B\left(\frac{S_*}{880\,\mathrm{mJy}}\right)^{-\gamma}, \tag{3.8}$$

where following Di Matteo et al. (2002) we take $B = 4.0\,\mathrm{mJy}^{-1}\mathrm{Sr}^{-1}$ and $\gamma = 1.75$ as our fiducial values[4]. Integrating over the distribution, Equation 3.6 becomes

$$\begin{aligned} m^{ps}(\nu) &= (17.4 x_{max}^{2-\gamma}\,\mathrm{K})\left(\frac{B}{4.0\,\mathrm{mJy}^{-1}\mathrm{Sr}^{-1}}\right) \times \\ &\quad \left(\frac{2-\gamma}{0.25}\right)^{-1} \left(\frac{\nu}{\nu_*}\right)^{-\alpha_{ps}+\frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu}{\nu_*}\right)}, \end{aligned} \tag{3.9}$$

where $x_{max} \equiv S_{max}/880\,\mathrm{mJy}$, and $\alpha_{ps} \equiv \alpha_0 + 2$ takes on a fiducial value of 2.5 to match measurements from the Cosmic Microwave Background (Tegmark et al., 2000). Note also that this implies $\alpha_0 \approx 0.5$, which is consistent with both Toffolatti et al. (1998) and Jackson (2005).

To compute the covariance of the distribution, we once again begin by considering a population of point sources of brightness S_* whose number density is determined by Poisson statistics (thus giving a result proportional to nS_*^2). We then integrate

[4]Despite the fact that the source count distribution was obtained by extrapolation, we expect that whatever the true distribution is, it should be well-approximated by a power law. This is because the key quantity is the integral of the source count, which is dominated by the brightest sources of the population. In that regime, we can linearize the distribution in log-log space, giving a power law in S_*.

over the source count distribution to get

$$\begin{aligned} C^{ps}(\nu,\nu') &= \frac{A_\nu A_{\nu'}}{\Omega_{pix}} \int_0^{S_{max}} \frac{dn}{dS_*} S_*^2 dS_* \left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_0+\frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)} \\ &= (4274 x_{max}^{3-\gamma}\,\mathrm{K}^2)\left(\frac{\Omega_{pix}}{10^{-6}\,\mathrm{Sr}}\right)^{-1}\left(\frac{B}{4.0\,\mathrm{mJy}^{-1}\mathrm{Sr}^{-1}}\right) \\ &\quad \times \left(\frac{3-\gamma}{1.25}\right)^{-1}\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_{ps}+\frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)}. \end{aligned} \tag{3.10}$$

Note that no further subtraction of the mean term is necessary, since it was implicitly accomplished when we invoked Poisson statistics.

3.2.3 Galactic Synchrotron Radiation

For Galactic synchrotron radiation, we imagine the foreground spectrum in each pixel to be well fit by a power law with spectral index α, but that the value of the spectral index may vary from pixel to pixel. In a given pixel, the spectrum is thus

$$x(\nu) = A_{sync}\left(\frac{\nu}{\nu_*}\right)^{-\alpha}, \tag{3.11}$$

where $A_{sync} = 335.4\,\mathrm{K}$ (Wang et al., 2006). Similar to the point sources, we assume that the indices in different pixels to be Gaussian distributed, only this time with a mean of $\alpha_{sync} = 2.8$ and a standard deviation of $\Delta\alpha_{sync} = 0.1$ (Wang et al., 2006). Performing the same integral as for the point sources, we obtain

$$m^{sync}(\nu) = A_{sync}\left(\frac{\nu}{\nu_*}\right)^{-\alpha_{sync}+\frac{\Delta\alpha_{sync}^2}{2}\ln\left(\frac{\nu}{\nu_*}\right)}. \tag{3.12}$$

Forming the foreground covariance using Equation 3.3, we obtain

$$\begin{aligned} C^{sync}(\nu,\nu') = A_{sync}^2 &\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_{sync}+\frac{\Delta\alpha_{sync}^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)} \\ &- m^{sync}(\nu) m^{sync}(\nu'). \end{aligned} \tag{3.13}$$

Parameter	Description	Fiducial Value
B	Source count normalization	$4.0\,\mathrm{mJy}^{-1}\,\mathrm{Sr}^{-1}$
γ	Source count power-law index	1.75
α_{ps}	Point source spectral index	2.5
σ_α	Point source index spread	0.5
A_{sync}	Synchrotron amplitude	335.4 K
α_{sync}	Synchrotron spectral index	2.8
$\Delta\alpha_{sync}$	Synchrotron index coherence	0.1
A_{ff}	Free-free amplitude	33.5 K
α_{ff}	Free-free spectral index	2.15
$\Delta\alpha_{ff}$	Free-free index coherence	0.01

Table 3.1: Free parameters in our foreground model and their fiducial values.

3.2.4 Free-free Emission

Free-free emission can be modeled in much the same way as the synchrotron radiation:

$$m^{ff}(\nu) = A_{ff}\left(\frac{\nu}{\nu_*}\right)^{-\alpha_{ff}+\frac{\Delta\alpha_{ff}^2}{2}\ln\left(\frac{\nu}{\nu_*}\right)} \tag{3.14}$$

$$C^{ff}(\nu,\nu') = A_{ff}^2\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_{ff}+\frac{\Delta\alpha_{ff}^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)} - m^{ff}(\nu)m^{ff}(\nu'). \tag{3.15}$$

but with $A_{ff} = 33.5\,\mathrm{K}$, $\alpha_{ff} = 2.15$, and $\Delta\alpha_{ff} = 0.01$ (Wang et al., 2006).

3.2.5 Total Foreground Contribution

To obtain total contribution to the mean signal, we simply sum the means of the various components:

$$m(\nu) = m^{ps}(\nu) + m^{sync}(\nu) + m^{ff}(\nu). \tag{3.16}$$

For the total covariance we sum the individual covariances:

$$C(\nu,\nu') = C^{ps}(\nu,\nu') + C^{sync}(\nu,\nu') + C^{ff}(\nu,\nu'). \tag{3.17}$$

In total, then, our foreground model consists of 10 free parameters, which are listed in Table 3.2.5 along with their fiducial values.

3.2.6 Noise model

In general, the noise level in a given pixel of a sky map produced by a radio telescope/interferometer scales as

$$\sigma_{noise} \propto \frac{\lambda^2 T_{sys}}{A_e \sqrt{\Delta\nu \Delta t}}, \tag{3.18}$$

where $\Delta\nu$ is the channel width of a frequency bin, Δt is the integration time, A_e is the collecting area of an antenna element, T_{sys} is the system temperature, and λ is the wavelength. To keep the discussion in this chapter as general as possible, we do not explicitly model the effective area and the system temperature, since both depend on one's specific instrument in complicated ways. Instead, we take a minimalistic approach and simply anchor our noise to levels that are expected for current generation 21 cm tomography experiments. From Bowman et al. (2009), we know that with 360 hours of integration at a channel width of 40 KHz, the MWA has a single pixel noise level at 158 MHz that is approximately 330 mK. Scaling this to our fiducial values, we obtain

$$\sigma_{noise} \sim (39\,\mathrm{mK}) \left(\frac{1\,\mathrm{MHz}}{\Delta\nu}\right)^{\frac{1}{2}} \left(\frac{1000\,\mathrm{hrs}}{\Delta t}\right)^{\frac{1}{2}}. \tag{3.19}$$

In the numerical studies conducted in this chapter, we will always be keeping Δt fixed at 1000 hrs. The frequency bin width $\Delta\nu$ will also be fixed at its fiducial value of 1 MHz, except when explicitly stated (for instance when we investigate the dependence of our results on $\Delta\nu$).

In the next section, we will be working almost exclusively in dimensionless units where the foreground power (quantified by the diagonal $\nu = \nu'$ elements of the foreground covariance) is equal to unity across all frequencies (see Equation 3.22). In such units, the noise can be taken to be approximately white *i.e.* frequency independent. To see this, note that the effective area A_e of an antenna scales as λ^2, which means that the quantity λ^2/A_e has no frequency dependence, and σ_{noise} depends on frequency only because of T_{sys}. Typical 21 cm experiments are sky noise dominated (Morales & Wyithe, 2009), which means that T_{sys} should be dominated by the foreground temperature, and therefore should have roughly the same frequency dependence as the foregrounds do. However, this frequency dependence was precisely the dependence that our choice of units was designed to null out. The noise thus becomes white to a good approximation, and is effectively a noise-to-signal ratio (henceforth denoted by

κ). Note that while we adopt this approximation for the rest of the chapter, it is by no means crucial, and in general one should always use units where the foreground covariance is white, even if in such units the noise is chromatic. For further discussion of this point, please see Section 3.3.2.

3.3 The Ease of Characterizing Foregrounds

While the foreground parameters described in Table 3.2.5 are certainly conventional choices, they are not the most economical, in the sense that many of them are redundant. That this is the case is not surprising, since so many different foreground sources can be accurately described by spectra that deviate only slightly from power laws. In Section 3.3.1 we will reparametrize our foreground model using a principal component analysis. In Section 3.3.4 we will determine the effective number of principal components that need to be measured to accurately describe the foreground spectra, and find the number to typically be three or four.

3.3.1 Eigenforeground Modes

Since a measured foreground spectrum will necessarily be discrete, we begin by discretizing the mean and the covariance of our foreground model from Section 3.2, so that

$$\mathbf{m}_\alpha \equiv m(\nu_\alpha) \tag{3.20}$$

and

$$\mathbf{C}_{\alpha\beta} \equiv C(\nu_\alpha, \nu_\beta), \tag{3.21}$$

where the indices run from 1 to N_c, the total number of frequency channels in one's instrument. We define a correlation matrix

$$\mathbf{R}_{\alpha\beta} \equiv \frac{\mathbf{C}_{\alpha\beta}}{\sqrt{\mathbf{C}_{\alpha\alpha}\mathbf{C}_{\beta\beta}}}, \tag{3.22}$$

and work with it instead of the covariance matrix.

We now perform a principal component analysis on the foregrounds[5]. That is, we

[5]We perform the principal component analysis using the correlation matrix $\mathbf{R}$ instead of the covariance matrix $\mathbf{C}$ because the challenge with foreground modeling is to successfully describe the fine relative perturbations about the smooth predictable power law spectrum. Since the relative fluctuations are quantified by $\mathbf{R}$, not $\mathbf{C}$, we should correspondingly use $\mathbf{R}$ for our principal component analysis.

rewrite the correlation matrix in a basis of "eigenforeground" vectors[6], where each eigenforeground vector $\mathbf{v}_n$ satisfies the eigenvalue equation

$$\mathbf{R}\mathbf{v}_n = \lambda_i \mathbf{v}_n. \tag{3.23}$$

Normalizing the eigenforeground vectors to unity and forming a matrix $\mathbf{V}$ where the columns of the matrix are the normalized eigenvectors, the correlation matrix can be expressed as

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V^t} = \sum_{n=1}^{N_c} \lambda_n \mathbf{v}_n \mathbf{v}_n^t, \tag{3.24}$$

where $\mathbf{\Lambda} \equiv \mathrm{diag}\{\lambda_1, \lambda_2, \lambda_3, \dots\}$. Note that since $\mathbf{R}$ is real and symmetric, $\mathbf{V^t} = \mathbf{V}^{-1}$, so $\mathbf{V}\mathbf{V^t} = \mathbf{V^t}\mathbf{V} = \mathbf{I}$. In the last equality of Equation 3.24, we see that each eigenvalue λ_n measures the contribution of its corresponding eigenforeground $\mathbf{v}_n$ to the total foreground variance.

In Section 3.3.4, we will seek to describe measured foreground spectra in terms of eigenforeground components. In other words, we wish to find the weight vector $\mathbf{a}$ (with component a_k for the k^{th} eigenforeground) in the equation

$$\mathbf{y} \equiv \mathbf{x} + \mathbf{n} = \sum_{k=1}^{N_c} a_k \mathbf{v}_k + \mathbf{n} = \mathbf{V}\mathbf{a} + \mathbf{n}, \tag{3.25}$$

where $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{n}$ are all vectors of length equal to the number of channels N_c, containing the measured spectrum of foregrounds, the true foregrounds, and the noise respectively. The vector $\mathbf{x}$, for instance, is simply a discretized and whitened version of $x(\nu)$ in Equation 3.1. Once an estimate $\hat{\mathbf{a}}$ of $\mathbf{a}$ has been determined, one can multiply by $\mathbf{V}$ to obtain an estimator $\hat{\mathbf{x}}$ of the foreground spectrum. We will find that by measuring a small number of parameters, we can characterize foreground spectra to a very high precision, thanks to the special properties of the eigenforegrounds, which we describe in Section 3.3.2.

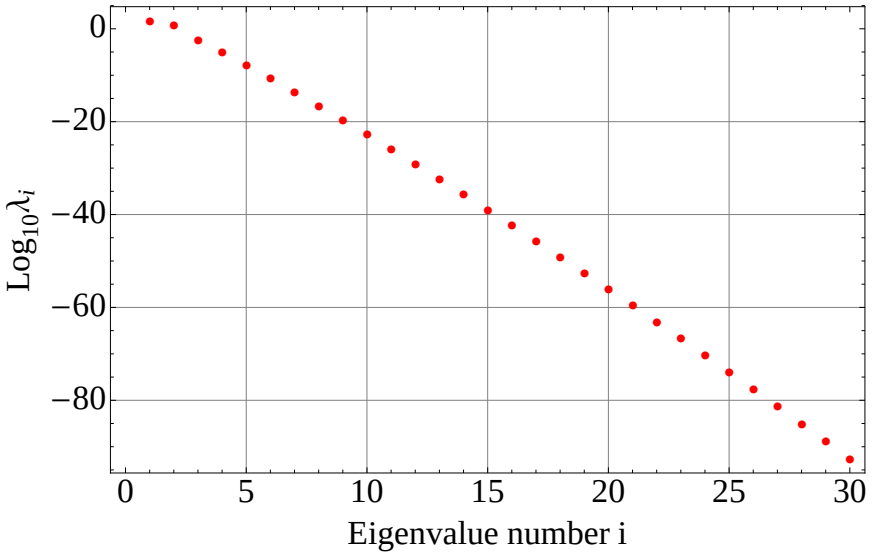

Figure 3-1: Eigenvalues of **R** with no noise for an experiment with 50 frequency channels, equally spaced from 100 MHz to 200 MHz.

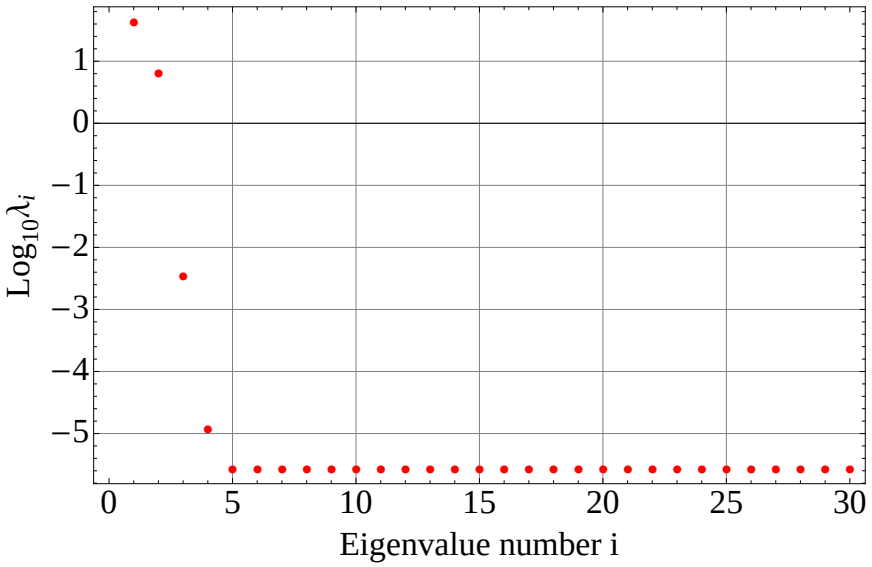

Figure 3-2: Eigenvalues of **R** for an experiment with 50 frequency channels, equally spaced from 100 MHz to 200 MHz, and noise levels given by the fiducial model of Section 3.2.6.

3.3.2 Features of the Eigenforegrounds

Since the foreground spectra do not contain any sharp features[7], we expect their rather featureless frequency dependences to be well-described by a small number of eigenforegrounds, *i.e.* we expect the eigenvalues λ_n to be large only for the first few values of n. Performing the analysis using the foreground model of Section 3.2 for a noiseless experiment with 50 equally spaced frequency channels going from 100 MHz to 200 MHz, we see in Figure 3-1 that the eigenvalues do fall off rapidly with mode number; indeed, the fall-off is exponential (a fact that we will explain later in this section), and in going from the first eigenvalue to the third there is a drop of more than three orders of magnitude. This means that if we select as our foreground parameters the expansion coefficients a_k of Equation 3.25, we can account for almost all of the foreground signal by measuring just a few numbers. In fact, in a realistic experiment it is impossible to measure more than the first few eigenvalues, as the foreground signal quickly becomes subdominant to the instrumental noise. This can be seen in Figure 3-2, where we once again show the eigenvalues, but this time including the noise model of Section 3.2.6. After the fourth eigenforeground, we see that the eigenvalues hit a noise floor because by then the eigenmodes are essentially measuring the noise. Note also that the large drop in magnitude of successive eigenvalues makes the qualitative results of this chapter robust to our assumption that the noise is white in our non-dimensionalized units — because the various eigenmodes contribute to the total foreground at such different levels, a slight chromaticity in the noise will not change the mode number at which the eigenmodes hit the noise floor in Figure 3-2. Violating our assumption will thus have very little effect on our results.

It should also be pointed out that in an analysis where the cosmological signal is included and is larger than the instrumental noise (as would be the case for a long integration time experiment), one would expect to hit a "signal floor" rather than a noise floor. The break in the eigenvalue spectrum in going from the exponential decay of the foreground dominated region to a flatter signal region can potentially be used as a diagnostic for separating foregrounds from signal. To properly investigate the robustness of this method, however, requires detailed simulations of the signal. Since the focus of this chapter is foreground modeling (and not signal extraction), we

[6]Throughout this chapter, we will use Greek indices to denote different frequencies and lowercase Latin indices to denote different foreground components/modes. In Section 3.4 we will use uppercase Latin indices to denote the different foreground parameters listed in Table 3.2.5.

[7]Following Morales et al. (2006); Gleser et al. (2008), we assume that narrowband contaminants such as terrestrial radio stations and radio recombination lines have been excised from the data prior to this point in the analysis.

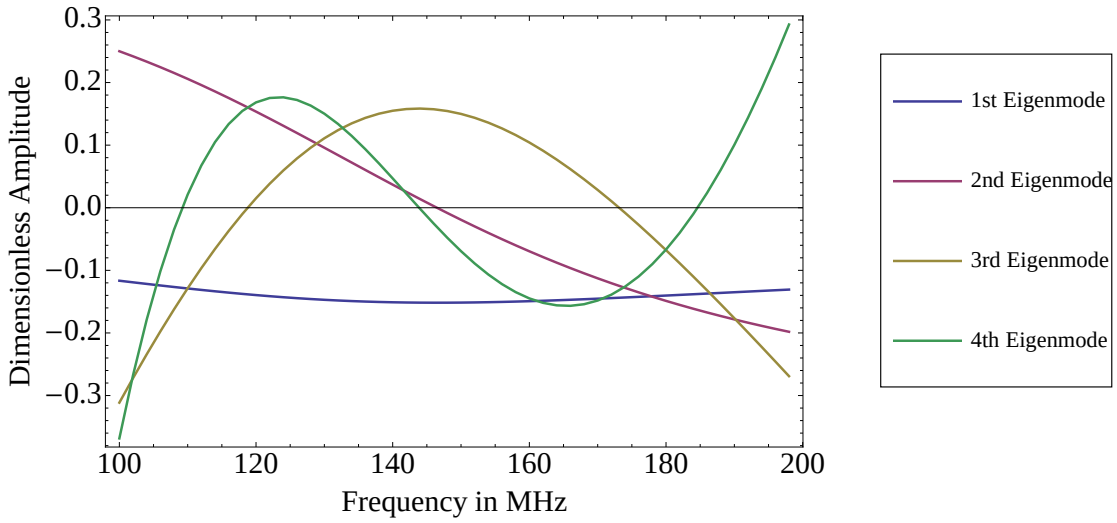

Figure 3-3: First few eigenvectors of $\mathbf{R}$ ("eigenforegrounds") for an experiment spanning a frequency range from 100 MHz to 200 MHz.

defer such an investigation to future work.

In Figure 3-3, we show the first few foreground eigenvectors, and in Figure 3-4 we restore the frequency-dependent normalization factors that were divided out in Equations 3.22. In both cases the eigenvectors are the ones defined by Equation 3.23 *i.e.* for the noiseless case, and henceforth we will only be using these noiseless eigenvectors. Even though we will include noise in our subsequent analysis, the use of the noiseless eigenvectors represents no loss of generality because the eigenvectors simply form a convenient set of basis vectors that span the space. In addition, with our assumption of white noise (in the units defined by Equation 3.22), the inclusion of noise alters only the eigenvalues, not the eigenvectors.

Several eigenforeground features are immediately apparent from the plots. First, the n^{th} eigenforeground is seen to have $n-1$ nodes. This is due to the fact that correlation matrix $\mathbf{R}$ is Hermitian (since it is real and symmetric), so Equation 3.23 takes the mathematical form of a time-independent Schrödinger equation for a one-dimensional system, and the node theorem of quantum mechanics applies. As a consequence, each successive foreground eigenmode probes a more rapidly fluctuating component of the spectrum, which explains the rapid fall in eigenvalues seen in Figure 3-1 — since foregrounds are such smooth functions of frequency, the more rapidly oscillating eigenmodes are simply not required.

Another characteristic feature of Figure 3-3 is that the n^{th} eigenforeground appears to look like a polynomial of order $(n-1)$, albeit with some slight deviations

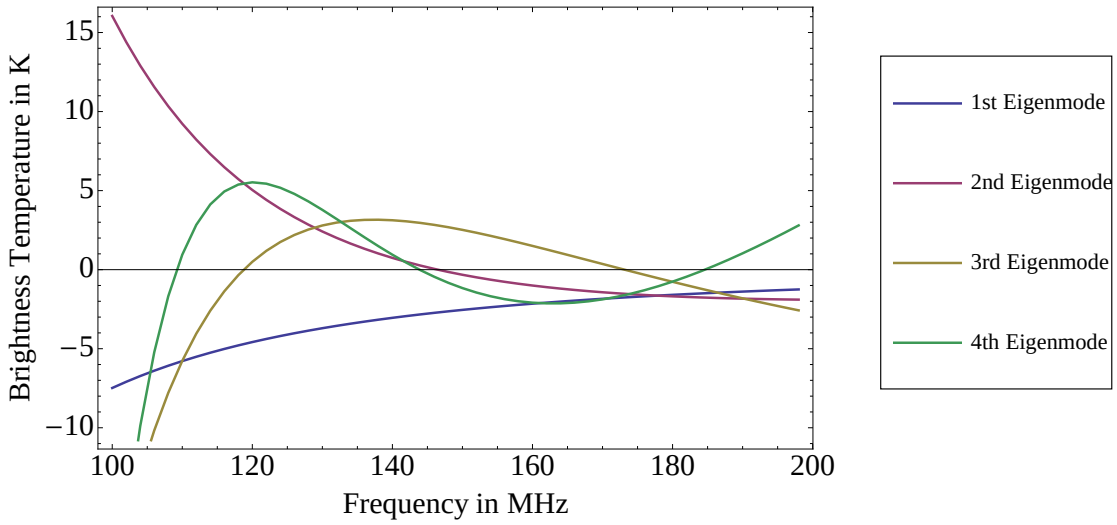

Figure 3-4: First few eigenvectors of $\mathbf{R}$ ("eigenforegrounds") for an experiment spanning a frequency range from 100 MHz to 200 MHz, but with the $\mathbf{C}_{\alpha\alpha}$ normalization factor restored so that the eigenforegrounds have units of temperature.

(the "linear" 1st eigenmode, for instance, has a small curvature to it). This approximate polynomial behavior explains the success of line-of-sight foreground subtraction schemes (Wang et al., 2006; Bowman et al., 2009; Liu et al., 2009a,b) that subtract low-order polynomials from foreground spectra — by subtracting out low-order polynomials, one is subtracting the modes with the largest eigenvalues, and since the eigenvalues fall so quickly (indeed, exponentially) with mode number, the result is that most of the foregrounds are cleaned out by the process. Together, Figures 3-3 and 3-4 also explain why it was found in Liu et al. (2009b) that polynomial foreground cleaning performs well only over narrow (~ 1 MHz) frequency ranges. In Figure 3-4, it is seen that over large frequency ranges the eigenforegrounds do not behave like polynomials, and that the polynomial behavior is only evident after dividing out a power-law-like normalization. In other words, the eigenforegrounds in Figure 3-4 still have a rough power law dependence to their spectrum, and thus are not well-fit by polynomials. Only if the frequency range is narrow will the power law dependence have a negligible effect, and so only in such a scenario would polynomial subtraction do well. Fortunately, there is a simple solution to this problem — if one wishes to perform polynomial subtraction over a wide frequency range, one can still do so successfully by first dividing out a fiducial spectrum (in the same way as we did in writing down Equation 3.22) before performing the fits[8]. Put another way, we simply

[8]Alternatively, one can fit foregrounds over a large frequency range by using polynomials in *logarithmic* frequency space. As noted in Bowman et al. (2009), however, one may want to avoid

have to perform the fits in Figure 3-3 (where the modes *always* look like polynomials) rather than in Figure 3-4.

3.3.3 Understanding the Eigenforegrounds

Despite the fact that it can be useful (in our whitened units) to think of the eigenforegrounds as polynomials of successively higher order, we caution that this interpretation is not exact. For instance, the first eigenmode in Figure 3-3 is clearly not precisely constant, and contains a small quadratic correction. Indeed, we will now show that it is more useful to think of the whitened eigenmodes as functions that would be sinusoids in ν were it not for an instrument's finite frequency range and spectral resolution.

Suppose that unresolved point sources were the sole contributor to our measured foregrounds. Working in a continuous description (to avoid finite bandwidth and resolution), the analog of Equation 3.22 takes the form[9]

$$R(\nu, \nu') = \frac{C(\nu, \nu')}{\sigma(\nu)\sigma(\nu')}, \tag{3.26}$$

where $C(\nu, \nu')$ is given by Equation 3.10 and

$$\sigma(\nu) \equiv [C(\nu, \nu)]^{\frac{1}{2}} = \left(\frac{\nu}{\nu_*}\right)^{-\alpha_{ps} + \sigma_\alpha^2 \ln(\nu/\nu_*)}. \tag{3.27}$$

The resulting correlation function $R(\nu, \nu')$ is given by

$$\begin{aligned} R(\nu, \nu') &= \exp\left[-\frac{\sigma_\alpha^2}{2}(\ln\nu - \ln\nu')^2\right] \\ &\approx \exp\left[-\frac{(\nu - \nu')^2}{2\nu_c^2}\right], \end{aligned} \tag{3.28}$$

where we have Taylor expanded about ν_* to obtain an unnormalized Gaussian with a *coherence length* of $\nu_c \equiv \nu_*/\sigma_\alpha \ln\nu_*$. Despite the fact that this expression was derived by only considering point sources, we find that it is an excellent approximation even when the full foreground model is used, provided one uses a longer coherence length

doing this because of the interferometric nature of most 21 cm experiments — since interferometers are not sensitive to the mean emission, negative values will be measured in certain pixels, making it problematic to take the logarithm.

[9] Aside from the omission of the noise term here, what follows is precisely the toy model introduced in Liu & Tegmark (2011).

to reflect the smoother Galactic synchrotron and free-free components ($\sim 64.8\,\mathrm{MHz}$ gives a good fit to the foreground model of Section 3.2).

Since $R(\nu, \nu')$ now just depends on the difference $\nu - \nu'$, the continuous analog of Equation 3.23 takes the form of a convolution:

$$\int R(\nu - \nu') f_n(\nu') d\nu' = \lambda_n f_n(\nu) \tag{3.29}$$

where f_n is the n^{th} eigenforeground spectrum. Because convolution kernels act multiplicatively in Fourier space, the f_i's are a family of sinusoids. Indeed, plugging the ansatz $f_n(\nu) = \sin(\gamma_n \nu + \phi)$ into Equation 3.29 yields

$$\int_{-\infty}^{\infty} \exp\left[-\frac{(\nu - \nu')^2}{2\nu_c}\right] \sin(\gamma_n \nu' + \phi) d\nu' = \lambda_n \sin(\gamma_n \nu + \phi), \tag{3.30}$$

where the eigenvalue is given by

$$\lambda_n = \sqrt{2\pi\nu_c^2} \exp\left(-2\nu_c^2 \gamma_n^2\right). \tag{3.31}$$

The eigenforegrounds shown in Figure 3-3 should therefore be thought of as a series of sinusoid-like functions that deviate from perfect sinusoidal behavior only because of "edge effects" arising from the finite frequency range of an experiment. As expected from the exponential form of Equation 3.31, the modes quickly decrease in importance with increasing wavenumber γ_n, and the more coherent the foregrounds (the larger ν_c is), the fewer eigenfunctions are needed to describe the foreground spectra to high accuracy.

One discrepancy between the eigenvalue behavior in our analytic treatment and the numerical results earlier on in the section is its dependence on wavenumber γ_n. Analytically, Equation 3.31 predicts that γ_n should appear quadratically in the exponent, whereas numerically Figure 3-1 suggests something closer to a linear relationship (provided one imagines that the mode number is roughly proportional to γ_n, which should be a good approximation at high mode numbers). We find from numerical experimentation that part of this discrepancy is due to the finite frequency range edge effects, and that as one goes from a very small range to a very large range, the dependence on γ_n steepens from being linear in the exponential to being a power law in the exponential, with the index of the power law never exceeding 2. In any case, the role of Equation 3.31 is simply to provide an intuitive understanding of the origin of the exponential fall drop in eigenvalues, and in the frequency ranges applicable to

21 cm tomography, we have verified numerically that a linear exponential fall-off fits the foreground model well. In particular, we can parametrize the eigenvalues with two parameters A and B, such that

$$\lambda_n \approx B \exp\left(-An\right), \tag{3.32}$$

and find that $A = 6.44$ and $B = 5.98 \times 10^5$ fits the foreground model well for an instrument with a frequency range of 100 to 200 MHz and frequency bins of $\Delta\nu = 2$ MHz. We caution, however, that these parameters vary with the frequency range, frequency bin width, and foreground model, and in general one must fit for A and B for each specific set of parameters. This is what we do to generate the numerical results of the following subsection. Generically, though, A will always be somewhat larger than unity, so the foreground spectra will always be dominated by the first few eigenmodes.

3.3.4 Eigenforeground Measurements

So far, we have established that foreground spectra should be describable to a very high accuracy using only a small number of principal foreground components, with the importance of the k^{th} foreground component quantified by the a_k coefficient in $\mathbf{a}$, which was defined in Equation 3.25. In a practical measurement, however, the presence of noise means that one does not know the true value of $\mathbf{a}$, but instead must work with an estimator $\hat{\mathbf{a}}$ that is formed from the data. There are many different ways to form this estimator, and one possibility would be to minimize the quantity

$$\chi^2 \equiv (\mathbf{y} - \mathbf{V}\hat{\mathbf{a}})^{\mathbf{t}}\mathbf{N}^{-1}(\mathbf{y} - \mathbf{V}\hat{\mathbf{a}}), \tag{3.33}$$

where $\mathbf{N}$ is defined as $\langle \mathbf{n}\mathbf{n}^{\mathbf{t}} \rangle$ using the noise vector $\mathbf{n}$ of Equation 3.25. This least-squares minimization (which is optimal if the noise is Gaussian) yields

$$\widehat{\mathbf{a}}^{LS} = [\mathbf{V}^{\mathbf{t}}\mathbf{N}^{-\mathbf{1}}\mathbf{V}]^{-\mathbf{1}}\mathbf{V}^{\mathbf{t}}\mathbf{N}^{-\mathbf{1}}\mathbf{y} = \mathbf{V}^{-\mathbf{1}}\mathbf{y} = \mathbf{V}^{\mathbf{t}}\mathbf{y}, \tag{3.34}$$

since the number of eigenmodes (*i.e.* the length of $\widehat{\mathbf{a}}^{LS}$) is equal to the number of frequency channels (*i.e.* the length of $\mathbf{y}$), so that $\mathbf{V}$ is a square orthogonal matrix and $\mathbf{V}^t = \mathbf{V}^{-1}$. Equation 3.34 states that even in the presence of noise, the least squares prescription calls for us to follow the same procedure as we would if there were no noise, namely, to take the dot product of both sides of Equation 3.25 with

each eigenforeground vector.

Defining an error vector $\varepsilon \equiv \delta\mathbf{a} = \hat{\mathbf{a}} - \mathbf{a}$ as the difference between the true $\mathbf{a}$ and the estimator $\hat{\mathbf{a}}$, we could instead choose to minimize the diagonal quantities $\langle|\boldsymbol{\varepsilon}_i|^2\rangle$. This corresponds to so-called Wiener filtering (Tegmark, 1997b), and the estimator is given by

$$\hat{\mathbf{a}}^{Wiener} = \mathbf{S}\mathbf{V}^\mathbf{t}\left[\mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t} + \mathbf{N}\right]^{-1}\mathbf{y}, \tag{3.35}$$

where $\mathbf{S} \equiv \langle\mathbf{a}\mathbf{a}^\mathbf{t}\rangle$ is the signal covariance matrix for the vector $\mathbf{a}$.

In Sections 3.3.4 and 3.3.4, we examine Wiener filtering and least-squares minimization. We will find that the estimated foreground spectrum $\hat{\mathbf{x}} = \mathbf{V}\hat{\mathbf{a}}$ from Wiener filtering is expected to be closer to the true foreground spectrum, and that the least-squares method can be adapted to give similar results, although with errors larger than for Wiener filtering.

Wiener Filtering

With the Wiener filtering technique, our estimate of the foreground spectrum is given by

$$\hat{\mathbf{x}}^{Wiener} = \mathbf{V}\hat{\mathbf{a}}^{Wiener} = \mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t}\left[\mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t} + \mathbf{N}\right]^{-1}\mathbf{y} \equiv \mathbf{W}\mathbf{y} \tag{3.36}$$

Understanding Wiener filtering thus comes down to understanding the filter $\mathbf{W}$. We first rewrite the expression slightly:

$$\mathbf{W} \equiv \mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t}\left[\mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t} + \mathbf{N}\right]^{-1} = \mathbf{V}\mathbf{S}\left[\mathbf{S} + \mathbf{V}^\mathbf{t}\mathbf{N}\mathbf{V}\right]^{-1}\mathbf{V}^\mathbf{t}, \tag{3.37}$$

where we have made liberal use of the fact that $\mathbf{V}^\mathbf{t} = \mathbf{V}^{-1}$. From Section 3.2.6, we know that $\mathbf{N}$ is proportional to the identity if we use whitened units. In the notation of that section, we have $\mathbf{N} = \kappa^2\mathbf{I}$, so

$$\mathbf{W} = \mathbf{V}\mathbf{S}\left[\mathbf{S} + \kappa^2\mathbf{I}\right]^{-1}\mathbf{V}^\mathbf{t}. \tag{3.38}$$

Now consider $\mathbf{S}$:

$$\begin{aligned}\mathbf{S} &= \mathbf{V}^\mathbf{t}\mathbf{V}\mathbf{S}\mathbf{V}^\mathbf{t}\mathbf{V} = \mathbf{V}^\mathbf{t}\langle(\mathbf{V}\mathbf{a})(\mathbf{V}\mathbf{a})^\mathbf{t}\rangle\mathbf{V} \\ &= \mathbf{V}^\mathbf{t}\langle\mathbf{x}\mathbf{x}^\mathbf{t}\rangle\mathbf{V} = \mathbf{V}^\mathbf{t}\mathbf{R}\mathbf{V} = \boldsymbol{\Lambda},\end{aligned} \tag{3.39}$$

where $\mathbf{x}$ denotes the true whitened foreground spectrum, as it did in Equation 3.25. In the penultimate step we used the fact that $\langle\mathbf{x}\mathbf{x}^\mathbf{t}\rangle$ is precisely our whitened foreground covariance, *i.e.* $\mathbf{R}$, and in the last step we used Equation 3.24. Inserting this result

into our Wiener filter gives

$$\mathbf{W}_{ij} = \left(\mathbf{V}\mathbf{\Lambda}\left[\mathbf{\Lambda} + \kappa^2\mathbf{I}\right]^{-1}\mathbf{V}^{\mathbf{t}}\right)_{ij} = \sum_{l,m}^{N_c} \mathbf{V}_{il}\left(\frac{\lambda_l}{\lambda_l + \kappa^2}\right)\delta_{lm}\mathbf{V}^{\mathbf{t}}{}_{mj}, \tag{3.40}$$

where as before N_c is the number of frequency channels (which is equal to the number of eigenforegrounds). This in turn means that our estimator takes the form

$$\begin{aligned}\hat{\mathbf{x}}_i^{Wiener} &= \sum_j^{N_c} \mathbf{V}_{ij}\left(\frac{\lambda_j}{\lambda_j + \kappa^2}\right)(\mathbf{V}^{\mathbf{t}}\mathbf{y})_j \\ &= \sum_j^{N_c} \mathbf{V}_{ij} w_j \hat{\mathbf{a}}_j^{LS} = \sum_j^{N_c} w_j \hat{\mathbf{a}}_j^{LS}\mathbf{v}_j(\nu_i),\end{aligned} \tag{3.41}$$

where $\mathbf{v}_j$ is the j^{th} eigenforeground as defined in Equation 3.23 and $w_j \equiv \lambda_j/(\lambda_j + \kappa^2)$ can be thought of as "Wiener weights" for the j^{th} eigenforeground. Since $w_j \approx 1$ for $\lambda_j \gg \kappa^2$ and $w_j \approx 0$ for $\lambda_j \ll \kappa^2$, this weighting factor preserves eigenmodes that have high signal to noise while suppressing those that have low signal to noise. In words, the Wiener filtering procedure thus amounts to first performing a least-squares fit to obtain estimates for the eigenforeground coefficients, but then in the reconstruction of the foreground spectrum, downweighting modes that were only measured with low signal-to-noise.

In Figure 3-5 we show the Wiener weights for the first ten eigenmodes for an experiment with 50 frequency channels, equally spaced from 100 MHz to 200 MHz, and noise levels given by the fiducial model of Section 3.2.6. The weights are seen to be small after the first few eigenmodes, suggesting that most of the information in our foreground spectrum estimate comes from a mere handful of parameters. The other parameters (*i.e.* eigenforeground coefficients) are too noisy to have much constraining power. Since the Wiener weights tend to unity in the limit of large signal-to-noise, we can define an effective number of measurable parameters n_{eff} by summing the Wiener weights:

$$n_{\text{eff}} \equiv \sum_i^{N_c} w_i. \tag{3.42}$$

For the fiducial scenario in Figure 3-5, the numerical value of n_{eff} is 4.06.

In general, n_{eff} depends on both the nature of the foregrounds and the noise level. If the noise levels were high, the difficulty in measuring the higher (and therefore subdominant) eigenmodes would result in those measurements being noise dominated.

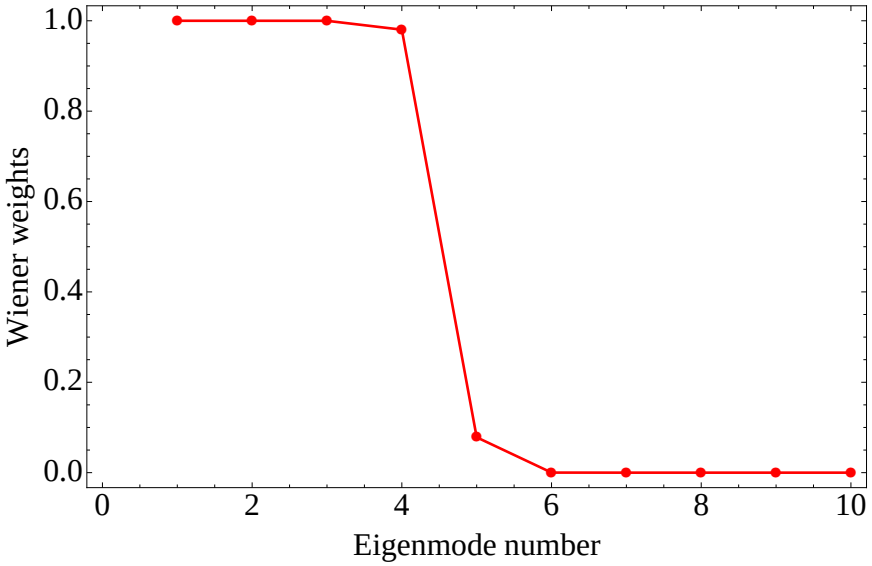

Figure 3-5: First few Wiener weights (defined as $w = \lambda/(\lambda + \kappa^2)$, where λ is the eigenvalue of a foreground mode and κ is the whitened noise) for an experiment with 50 frequency channels, equally spaced from 100 MHz to 200 MHz, and noise levels given by the fiducial model of Section 3.2.6.

The Wiener filter would thus suppress such modes, driving n_{eff} down. In the opposite limit where the noise levels are low, one would expect n_{eff} to be higher, but still to be relatively small. This is because we know from Section 3.3.2 that the foreground spectra are relatively simple functions that are dominated by the first few eigenmodes, and thus even if the noise levels were low enough for the higher modes to be measurable, it is unnecessary to give them very much weight. In general, n_{eff} varies roughly logarithmically with increasing foreground-to-noise ratio since the foreground eigenvalues drop exponentially. This nicely explains one of the results from Liu & Tegmark (2011). There it was found that in performing foreground subtraction by deweighting foreground contaminated line-of-sight Fourier modes, the number of modes that needed to be deweighted increased only logarithmically with the foreground-to-noise ratio. We now see that this is simply a consequence of Equation 3.32.

In Figure 3-6, the red/grey dotted line shows n_{eff} as a function of the total frequency range of an experiment. In changing the frequency range, we keep the integration time and the channel width $\Delta\nu$ constant so that the noise level remains unchanged. Any change in n_{eff} therefore reflects the nature of the foregrounds, and what one sees in Figure 3-6 is that as the total frequency range increases, more and more modes are needed when estimating the spectrum. This is because effects like the spread in the spectral indices are apparent only over large frequency ranges, so

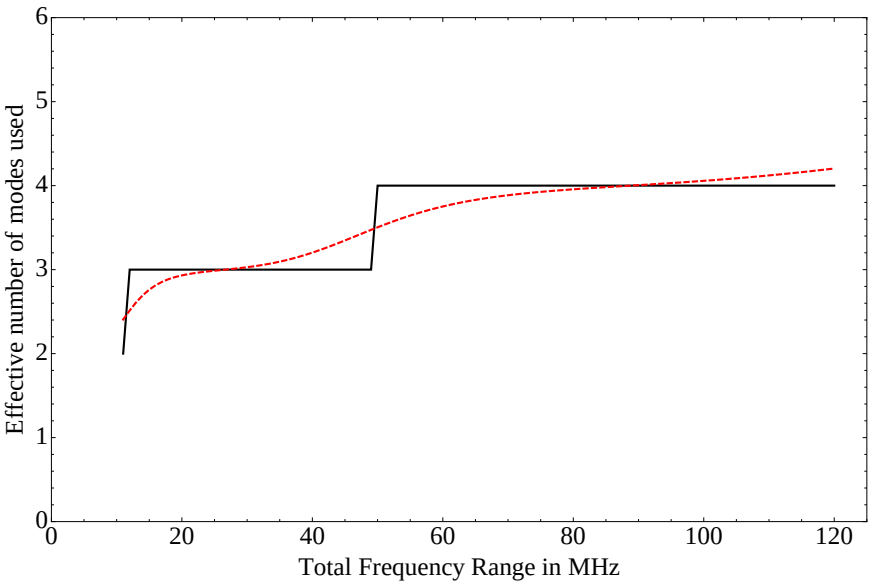

Figure 3-6: Shown with the red/grey dashed curve, the effective number $n_{\rm eff}$ of foreground parameters used (defined by Equation 3.42) in the Wiener filtering method. In solid black is the optimal m (Equation 3.48 adjusted for the fact that m must be an integer) for the truncated least-squares method. In both cases the behavior is shown as a function of the total frequency range of an instrument, with channel width and integration time (and thus noise level) held constant at 1 MHz and 1000 hrs, respectively.

increasing the frequency range makes foregrounds more complicated to model.

To quantify the success of a foreground model generated from Wiener filtered measurements, we consider the quantity $\delta\hat{\mathbf{x}}^w \equiv \hat{\mathbf{x}}^{\rm Wiener} - \mathbf{x}$:

$$\begin{aligned}
\delta\hat{\mathbf{x}}_i^w &= \sum_j^{N_c} \mathbf{V}_{ij}(w_j\hat{\mathbf{a}}_j^{\mathbf{LS}} - \mathbf{a}_j) \\
&= \sum_j^{N_c} \mathbf{V}_{ij}(w_j - 1)\mathbf{a}_j - \sum_{j,k}^{N_c} \mathbf{V}_{ij}w_j\mathbf{V}_{kj}\mathbf{n}_k \\
&= \sum_j^{N_c} (w_j - 1)\mathbf{a}_j\mathbf{v}_j - \sum_j^{N_c} \mathbf{W}_{ij}\mathbf{n}_j
\end{aligned} \tag{3.43}$$

where in the penultimate step we substituted $\hat{\mathbf{a}}^{LS} = \mathbf{a} + \mathbf{V^t n}$ (which follows from Equations 3.25 and 3.34), and in the ultimate step recognized that $\sum_j \mathbf{V}_{ij}w_j\mathbf{V}_{kj}$ is

the Wiener filter $\mathbf{W}$ of Equation 3.40. The covariance of this quantity is

$$\begin{aligned}\langle \delta\hat{\mathbf{x}}^w \delta\hat{\mathbf{x}}_w^t \rangle &= \sum_{i,j}^{N_c} \langle \mathbf{a}_i \mathbf{a}_j \rangle (w_i - 1)(w_j - 1)\mathbf{v}_i \mathbf{v}_j^{\mathbf{t}} + \mathbf{W}\langle \mathbf{n}\mathbf{n}^{\mathbf{t}} \rangle \mathbf{W}^{\mathbf{t}} \\ &= \sum_{i}^{N_c} \lambda_i (w_i - 1)^2 \mathbf{v}_i \mathbf{v}_i^{\mathbf{t}} + \mathbf{W}\mathbf{N}\mathbf{W}^{\mathbf{t}} \\ &= \sum_{i}^{N_c} \lambda_i (w_i - 1)^2 \mathbf{v}_i \mathbf{v}_i^{\mathbf{t}} + \kappa^2 \sum_{i}^{N_c} w_i^2 \mathbf{v}_i \mathbf{v}_i^{\mathbf{t}} \end{aligned} \tag{3.44}$$

where in the penultimate step we used $\langle \mathbf{a}_i \mathbf{a}_j \rangle = \mathbf{S}_{ij} = \mathbf{\Lambda}_{ij} = \lambda_i \delta_{ij}$, and in the last step used Equation 3.40 as well as $\mathbf{N} = \kappa^2 \mathbf{I}$. Note that there are no cross terms in going from Equation 3.43 to 3.44 because such terms would be proportional to $\langle \sum_i a_i \mathbf{v}_i \mathbf{n}^{\mathbf{t}} \rangle = \langle \mathbf{x}\mathbf{n}^{\mathbf{t}} \rangle$, which is assumed to be zero because the foregrounds and the noise are uncorrelated. The diagonal terms in Equation 3.44 give the expected mean-square measurement error at a particular frequency. Averaging over frequency to obtain a figure-of-merit for the entire spectrum, our mean-square measurement error becomes

$$\begin{aligned}\varepsilon^2_{\text{Wiener}} \equiv & \quad \overline{\langle \delta\hat{\mathbf{x}}^w \delta\hat{\mathbf{x}}_w^t \rangle} \\ &= \frac{1}{N_c} \sum_{\alpha}^{N_c} \sum_{i}^{N_c} \lambda_i (w_i - 1)^2 \mathbf{v}_i^2(\nu_\alpha) \\ &\qquad + \frac{\kappa^2}{N_c} \sum_{\alpha}^{N_c} \sum_{i}^{N_c} w_i^2 \mathbf{v}_i^2(\nu_\alpha) \\ &= \quad \frac{1}{N_c} \sum_{i}^{N_c} \lambda_i (w_i - 1)^2 + \frac{\kappa^2}{N_c} \sum_{i}^{N_c} w_i^2, \end{aligned} \tag{3.45}$$

where in the last step we permuted the sums and used the fact that the eigenforeground vectors are normalized. From this expression, we can see that there are essentially two sources of error, each represented by one of the terms in Equation 3.45. The first term is only significant for the higher order eigenmodes, where $w_i \approx 0$. With each element in the sum proportional to λ_i, this term quantifies the error incurred by heavily de-weighting the higher order modes (which are de-weighted so much that they are essentially omitted from the estimator). Of course, we have no choice but to omit these modes, because their measurements are so noisy that there is essentially no foreground information contained in them. The second term is significant only for the first few eigenforegrounds, which have $w_i \approx 1$. Being proportional to κ^2, this

term quantifies the error induced by instrumental noise in the measurement of the lower order modes (which are included in the estimator).

In Figure 3-7, we show (with the red/grey dashed line) the fractional error in an estimate of a foreground spectrum as a function of the total frequency range of an instrument. The channel width $\Delta\nu$ and integration time Δt are held constant, so there is no change in the noise level. From the plot, we see that as the frequency range of the instrument increases, one is able to construct an increasingly accurate estimator of the foreground spectrum. To understand why this is so, note that as the frequency range increases, there are two effects at play — first, a larger frequency range means more complicated foregrounds, possibly increasing the errors in the fits; however, a larger frequency range with fixed channel width results in more data points to fit to, as well as a longer "lever arm" with which to probe spectral features, thus reducing the errors. What we see in Figure 3-7 is that the second outweighs the first. This is unsurprising, since we saw from Figure 3-6 that as the frequency range is increased, the number of modes required to describe the foregrounds increases rather slowly, suggesting that the foreground spectra are only getting marginally more complicated.

Truncated Least-Squares

In a conventional least-squares fitting of a foreground spectrum, one forms an estimator for the spectrum by taking $\hat{\mathbf{a}}^{\mathbf{LS}}$ and multiplying by $\mathbf{V}$. In other words, one fits for all the eigenforeground coefficients (*i.e.* the vector $\mathbf{a}$) and multiplies each coefficient by the spectrum of the corresponding eigenforeground. This, however, is a suboptimal procedure for estimating the true spectrum because the higher order eigenmodes have very low signal-to-noise and measurements of them tend to be noise dominated. Such modes should therefore be excluded (or heavily downweighted), which was what the Wiener filtering of the previous section accomplished. In this section we explore a simpler method where we measure all eigenforeground coefficients, but include only the first m eigenmodes in our estimate of the spectrum, truncating the eigenmode expansion so that the noisier measurements are excluded.

In this truncated least-squares scheme, our estimate of the spectrum for the i^{th} frequency channel takes the form

$$\hat{\mathbf{x}}_i^{LS} = \sum_j^m \hat{\mathbf{a}}_j^{LS} \mathbf{v}_j(\nu_i), \tag{3.46}$$

where m is an integer to be determined later. Written in this way, we see that we can

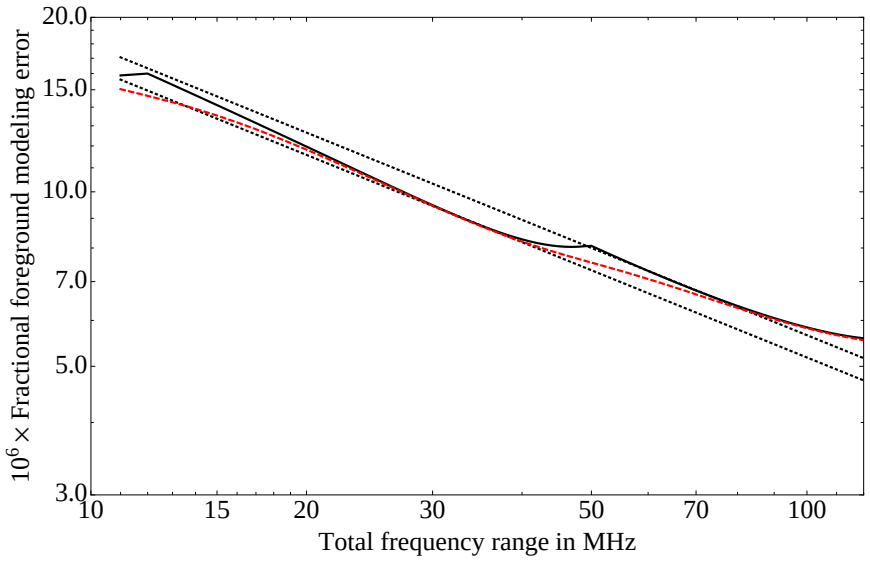

Figure 3-7: Expected error on measured foregrounds divided by the root-mean-square (r.m.s.) foreground intensity to give a fractional foreground modeling error. In dashed red/grey is the error for the Wiener filtering of Section 3.3.4 (the square root of Equation 3.45 divided by the r.m.s.). In solid black is the error for the truncated least-squares method of Section 3.3.4 (Equation 3.49 divided by the r.m.s. and adjusted for the fact that m must be an integer). In both cases we have plotted the errors on a log-log scale as functions of the total frequency range for an instrument with a fixed 1 MHz channel width and 1000 hrs of integration time, ensuring a constant noise level. The black dotted lines are included for reference, and are proportional to $1/\sqrt{N_c \Delta\nu}$.

reuse all the expressions derived in the previous section as long as we set $w_i = 1$ for $i \leq m$ and $w_i = 0$ for $i > m$. Put another way, the truncated-least squares method approximates Wiener filtering by replacing the plot of Wiener weights shown in Figure 3-5 with a step function. Using Equation 3.45, we thus see that the mean-square error for the truncated least-squares method is

$$\begin{aligned} \varepsilon_{\mathrm{LS}}^2 &= \frac{1}{N_c}\sum_{i=1}^{N_c} \lambda_i (w_i - 1)^2 + \frac{\kappa^2}{N_c}\sum_{i=1}^{N_c} w_i^2 \\ &= \frac{1}{N_c}\sum_{i=m+1}^{N_c} \lambda_i + \frac{\kappa^2}{N_c}\sum_{i=1}^{m} 1 \\ &\approx \frac{B}{N_c}\exp\left[-A(m+1)\right] + m\frac{\kappa^2}{N_c}, \end{aligned} \tag{3.47}$$

where in the last step we used the fact that the eigenvalues fall off in a steep exponential (Equation 3.32) that decays quickly enough that we can to a good approximation

omit all but the first term in the sum.

From Equation 3.47, we see that the truncated least-squares method gives large errors for extreme values of m, both large and small — at large m, many modes are included, making the exponential term in the error small, but at the cost of large noise contamination from the second term; at low m, the noise term is small, but too few eigenforegrounds are included in the estimator of the spectrum, and the first term is large. To obtain the best possible error from the truncated least-squares method, we must therefore choose an intermediate value of m that minimizes ε_{LS}^2. Differentiating with respect to m and setting the result to zero allows one to solve for the optimal m:

$$m_{opt} = \frac{1}{A} \ln\left(\frac{AB}{\kappa^2}\right) - 1, \tag{3.48}$$

and inserting this into our expression for the error gives a best error of

$$\varepsilon_{LS}^{best} \sim \frac{\kappa_{fid}}{\sqrt{N_c \Delta\nu \Delta t}} \left[\frac{1}{A}\left[\ln\left(\frac{AB\sqrt{\Delta t \Delta\nu}}{\kappa_{fid}^2}\right) + 1\right] - 1\right]^{\frac{1}{2}}, \tag{3.49}$$

where we have made the substitution $\kappa \to \kappa_{fid}/\sqrt{\Delta t \Delta\nu}$ to emphasize the scalings with integration time Δt and channel width $\Delta\nu$. Note that Equation 3.49 is only approximate, since m must take on an integer value.

In Figure 3-6, the solid black curve shows m_{opt} as a function of $N_c\Delta\nu$, fitting numerically for A and B (Equation 3.32) for each $N_c\Delta\nu$. The channel width $\Delta\nu$ and Δt are held at 1 MHz and 1000 hrs respectively, as described in Section 3.2.6. Since $\Delta\nu$ and Δt are constant, the normalized noise level κ is also constant, and the variation in m_{opt} is due purely to changes in A and B. This variation is seen to be quite weak, and we see that the optimal number of components to solve for is always a rather small number, and that this number increases only slowly with the number of channels N_c (or equivalently with the total frequency range $N_c\Delta\nu$, since the channel width $\Delta\nu$ is being held constant). This is unsurprising, given that the first term in Equation 3.47 decays exponentially while the second term rises only linearly, forcing the optimal m towards small m. With any current-generation experiment, we find m_{opt} to be no larger than 4. Also note that m_{opt} behaves like a discretized version of n_{eff} from the Wiener filtering, as is expected.

In Figure 3-7, we show how the optimal m values of Figure 3-6 translate into the error $\varepsilon_{total}^{best}$, again as a function of $N_c\Delta\nu$. The error is divided by the root-mean-square value of the foregrounds over the entire spectrum to give a rough percentage error, and is plotted as the solid black line, while reference lines proportional to

$1/\sqrt{N_c \Delta\nu}$ are shown in dotted black. The error is seen to decrease mostly with the inverse square root of the total frequency range $N_c \Delta\nu$, as one would expect from the prefactor in Equation 3.49. The deviation from this behavior is due to the fact that the second part of Equation 3.49 depends on A and B, which in turn vary with N_c. This, however, is a rather weak effect, since the relevant parts of Equation 3.49 look a lot like our expression for m_{opt}, which we know from Figure 3-6 rises slowly. The result is that deviations from a $1/\sqrt{N_c \Delta\nu}$ scaling occur only when one is close to the transition in m_{opt} (which remember, must be an integer). Again, we see that that truncated least-squares method closely approximates the Wiener filtering method, with only slightly larger errors.

With both methods, we find that there is very little change in the error if one changes the channel width $\Delta\nu$ while correspondingly adjusting N_c so that $N_c \Delta\nu$ (*i.e.* the total frequency range of the instrument) is kept constant. For the truncated least-squares method, such changes keep the prefactor in Equation 3.49 fixed, and any variations in the error are due purely to the corrections in the second part of the equation. Numerically, the lack of sharp spectral features in the foregrounds means that these corrections are found to be completely subdominant. This implies that to an excellent approximation, the errors in our measured foregrounds are dependent only on the total frequency range of our instrument, and are independent of how we bin our data — binning more coarsely results in a lower noise per frequency bin, but this is canceled out (to a very high precision) by the larger number of bins with which to perform our fits[10]. From this, we see that the decrease in errors seen in Figure 3-7 (where we increased the number of data points by increasing the total frequency range of the instrument) is due not to intrinsically better fits, but rather to a longer "lever arm" over which to probe foreground characteristics.

Which method to use?

In summary, we can see from Figure 3-6 that it is indeed the case that 21 cm foregrounds can be accurately characterized using just a small number ($\sim$ 3 or 4) of independent components. From Figure 3-7, we see that both Wiener filtering and the truncated least-squares method allow foregrounds to be estimated to an accuracy of roughly one part in 10^{-5} to 10^{-6}. This is fortunate since the cosmological signal is expected to be $\sim 10^{-4}$ smaller than the foregrounds, so a failure to reach at least that

[10]This is provided one makes the (usually excellent) assumption that the number of frequency bins is much larger than the number of independent foreground modes, seen in this section and the previous one to be 3 or 4.

level of precision would ruin any prospects of a cosmological measurement.

With the Wiener filtering and the truncated least-squares giving such similar results, for most applications it should not matter which method is used. However, if one requires the very best foreground model achievable, then Wiener filtering should be used, since it was derived by minimizing the error. On the other hand, the least-squares method has the advantage of being simpler, and in addition is more immune to inaccuracies in foreground modeling. This is because Wiener filtering explicitly involves the foreground covariance $\mathbf{S}$, and for the method to minimize the modeling error, it is important that the foreground-to-noise ratio be accurate. On the other hand, the foregrounds only enter the least squares method via the choice of basis (*i.e.* through $\mathbf{V}$), and since this basis spans the vector space, the role of the foreground model in least-squares fitting is simply that of a prior model. Ultimately, though, the Wiener filtering method's need for an estimate of the foreground covariance is unlikely to be an obstacle in practice. As demonstrated in de Oliveira-Costa et al. (2008), it is possible to derive a principal component basis (and the corresponding eigenvalues) empirically.

3.4 The difficulty in measuring physical parameters

In the previous section, we saw that the foregrounds can be accurately described using only a small number of foreground eigenmodes. Put another way, there are very few distinguishing features in radio foreground spectra, so there are really only a few independent parameters in the data. In this section, we will see that this places severe constraints on our ability to measure the physical parameters listed in Table 3.2.5.

We employ a Fisher matrix formalism to estimate the best possible error bars on measurements of these parameters. To do so, we imagine that one has already used Equation 3.34 to compute an estimator $\hat{\mathbf{a}}$ of the expansion coefficient vector from the data[11]. The precise value of this estimator will vary because of the random nature of instrumental noise, whose effect can be quantified by computing the error covariance matrix of $\hat{\mathbf{a}}$. Since both the Wiener filtering method and the truncated

[11]Instead of considering the expansion coefficients, one could instead deal with the foreground spectrum $\mathbf{x}$ and its estimator $\hat{\mathbf{x}}$ directly. However, unlike for the expansion coefficients, the spectra themselves have the unfortunate property of having non-diagonal error covariances (*e.g.* Equation 3.43 for Wiener filtering), which makes the subsequent analysis and interpretation in this section more cumbersome.

least-squares method require as a first step a full least-squares estimation of $\mathbf{a}$, we can use the standard formula for the error covariance Σ for least-squares fitting (Tegmark, 1997b):

$$\Sigma \equiv \langle [\hat{\mathbf{a}} - \langle \mathbf{a} \rangle][\hat{\mathbf{a}} - \langle \mathbf{a} \rangle]^{\mathbf{t}} \rangle = [\mathbf{V^t N^{-1} V}]^{-\mathbf{1}} = \kappa^2 \mathbf{I}, \tag{3.50}$$

where like before we used the fact that in our whitened units, $\mathbf{N} = \kappa^2 \mathbf{I}$. Once again, the angle brackets $\langle \dots \rangle$ denote an ensemble average, or (equivalently from the standpoint of practical observations) an average over many independent lines-of-sight. With this covariance, the probability distribution L of $\hat{\mathbf{a}}$ is given by

$$L(\hat{\mathbf{a}}; \mathbf{\Theta}) = \frac{1}{\sqrt{(2\pi)^{N_c} \det \mathbf{\Sigma}}} \exp\left[-\frac{1}{2} [\hat{\mathbf{a}} - \langle \mathbf{a} \rangle]^t \mathbf{\Sigma}^{-1} [\hat{\mathbf{a}} - \langle \mathbf{a} \rangle] \right], \tag{3.51}$$

where $\mathbf{\Sigma}$ is the error covariance matrix, and $\mathbf{\Theta} \equiv (\theta_1, \theta_2, \theta_3, \dots)$ is a vector of model parameters (such as the parameters in Table 3.2.5), which enters our expression because the true expansion coefficient vector $\mathbf{a}$ depends on these parameters. Interpreted as a function of $\mathbf{\Theta}$ for a given measurement $\hat{\mathbf{a}}$, L is the so-called likelihood function for the parameters that we wish to constrain. The tightness of our constraint is closely related to the Fisher information matrix, which is defined as

$$\mathbf{F}_{AB} \equiv \left\langle \frac{\partial^2 \mathcal{L}}{\partial \theta_A \partial \theta_B} \right\rangle, \tag{3.52}$$

where $\mathcal{L} \equiv -\ln L$. If our estimator for the parameters is unbiased, *i.e.*

$$\langle \mathbf{\Theta} \rangle = \mathbf{\Theta}_0, \tag{3.53}$$

where $\mathbf{\Theta}_0$ is the true parameter vector, then the Cramer-Rao inequality states that the error bars on θ_A (defined by the standard deviation $\Delta\theta_A \equiv \sqrt{\langle \theta_A^2 \rangle - \langle \theta_A \rangle^2}$) satisfy

$$\Delta\theta_A \geq (\mathbf{F}^{-1})_{AA}^{1/2} \tag{3.54}$$

if we estimate all the parameters jointly from the data. Computing the Fisher matrix thus allows us to estimate the ability of an experiment to constrain physical parameters, with the covariance between the parameter estimates equal to $\mathbf{F}^{-1}$ if the data analysis is done in an optimal fashion.

Let us now compute the Fisher matrix for the foreground parameters listed in Table 3.2.5. Inserting Equation 3.51 into Equation 3.52 and performing some matrix

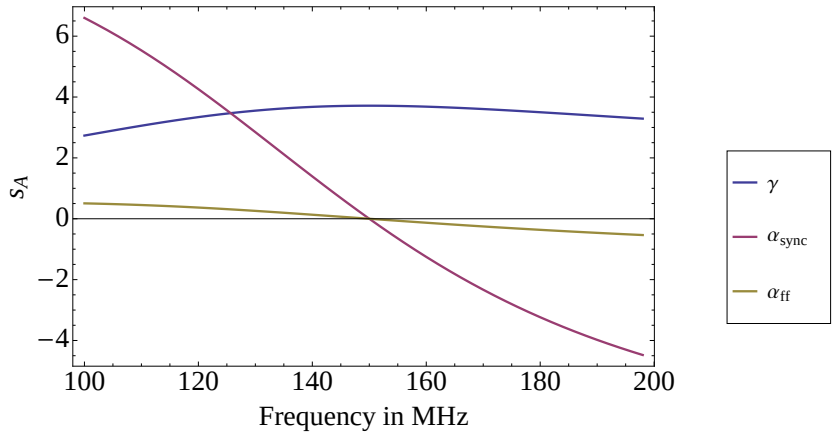

Figure 3-8: Parameter derivative vectors $\mathbf{s}_A$ (Equation 3.59) for γ, α_{sync}, and $\Delta\alpha_{sync}$.

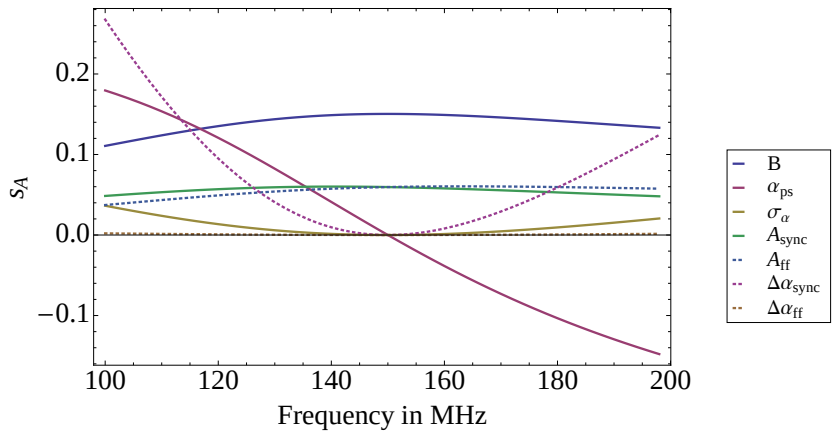

Figure 3-9: Parameter derivative vectors $\mathbf{s}_A$ (Equation 3.59) for B, α_{ps}, σ_α, A_{sync}, A_{ff}, α_{ff}, and $\Delta\alpha_{ff}$.

algebra, one can show (Tegmark et al., 1997) that the Fisher matrix reduces to

$$\mathbf{F}_{AB} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \theta_A}\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \theta_B}\right] + \frac{\partial \langle \mathbf{a^t} \rangle}{\partial \theta_A}\mathbf{\Sigma}^{-1}\frac{\partial \langle \mathbf{a} \rangle}{\partial \theta_B}. \tag{3.55}$$

Referring back to our expression for $\mathbf{\Sigma}$, we see that the first term vanishes because $\mathbf{\Sigma}$ depends only on the noise level and not on the physical parameters. This gives

$$\mathbf{F}_{AB} = \frac{1}{\kappa^2}\left(\frac{\partial \langle \mathbf{a^t} \rangle}{\partial \theta_A}\right)\left(\frac{\partial \langle \mathbf{a} \rangle}{\partial \theta_B}\right). \tag{3.56}$$

Note that this expression depends on the mean vector $\langle \mathbf{a} \rangle$, not the estimator $\hat{\mathbf{a}}$. This is because the Fisher matrix formalism tells us what the error bars are for the *optimal* method, whatever that method happens to be. We thus do not expect $\hat{\mathbf{a}}$ to appear in $\mathbf{F}_{AB}$, for if it did we would have the freedom to plug in a possibly sub-optimal estimator of our choosing, and by construction the Fisher matrix formalism contains information about the optimal errors.

With $\langle \mathbf{a} \rangle$ signifying the true expansion coefficient vector, we know from Equations 3.2 and 3.25 that

$$\langle \mathbf{a} \rangle = \mathbf{V^t}\langle \mathbf{x} \rangle = \mathbf{V^t}\mathbf{m}. \tag{3.57}$$

Inserting this into our expression for the Fisher matrix, we obtain[12]

$$\mathbf{F}_{AB} = \frac{1}{\kappa^2}\left(\frac{\partial \mathbf{m^t}}{\partial \theta_A}\right)\mathbf{V}\mathbf{V^t}\left(\frac{\partial \mathbf{m}}{\partial \theta_B}\right) = \frac{1}{\kappa^2}\mathbf{s}_A \cdot \mathbf{s}_B, \tag{3.58}$$

where we have used the fact that $\mathbf{V^t} = \mathbf{V}^{-1}$ since the eigenforegrounds are orthogonal, and have defined

$$\mathbf{s}_A \equiv \frac{\partial \mathbf{m}}{\partial \theta_A}. \tag{3.59}$$

Each component of Equation 3.58 can be geometrically interpreted as a dot product between two $\mathbf{s}$ vectors in an N_c-dimensional space. As $\mathbf{m}$ encodes the whitened foreground spectrum (Equation 3.22), each $\mathbf{s}$ vector quantifies the change in the expected foreground spectrum with respect to a physical parameter. And since the final covariance matrix on the parameter errors is given by $\mathbf{F}^{-1}$, the greater the dot product between two different $\mathbf{s}$ vectors, the more difficult it is to measure the two corresponding physical parameters. The forms of different $\mathbf{s}_A$'s thus give intuition for

[12]We have implicitly assumed that the whitening procedure performed in Equations 3.22 was with respect to some *fiducial* foreground model, so that $\mathbf{V}$ does not depend on the foreground parameter vector $\mathbf{\Theta}$.

	B	γ	A_{sync}	A_{ff}	α_{ps}	α_{sync}	α_{ff}	σ_α	$\Delta\alpha_{sync}$	$\Delta\alpha_{ff}$
B	1.00	1.00	0.998	0.998	0.0603	0.111	-0.00425	0.683	0.664	0.705
γ	—	1.00	0.999	0.998	0.0603	0.111	-0.00425	0.683	0.664	0.705
A_{sync}	—	—	1.00	0.993	0.113	0.163	0.0489	0.696	0.680	0.713
A_{ff}	—	—	—	1.00	-0.00413	0.0465	-0.0684	0.661	0.639	0.688
α_{ps}	—	—	—	—	1.00	0.997	0.996	0.333	0.389	0.254
α_{sync}	—	—	—	—	—	1.00	0.987	0.399	0.454	0.322
α_{ff}	—	—	—	—	—	—	1.00	0.245	0.303	0.164
σ_α	—	—	—	—	—	—	—	1.00	0.998	0.996
$\Delta\alpha_{sync}$	—	—	—	—	—	—	—	—	1.00	0.987
$\Delta\alpha_{ff}$	—	—	—	—	—	—	—	—	—	1.00

Table 3.2: Dimensionless dot products between $\mathbf{s}_A$ vectors (Equation 3.59) for the foreground parameters listed in Table 3.2.5, or equivalently, a normalized version of the Fisher matrix (given by Equation 3.58) so that the diagonal elements are unity. The derivatives were evaluated at the fiducial foreground parameters for the foreground model described in Section 3.2. The frequency range of the experiment was taken to go from 100 MHz to 200 MHz, with 50 equally spaced channels. The matrix is symmetric by construction, so the bottom left half has been omitted for clarity.

the degeneracies in a set of parameters.

Shown in Figures 3-8 and 3-9 are plots of the $\mathbf{s}_A$ vectors for the foreground parameters shown in Table 3.2.5. For clarity, we have separated the derivatives into two plots that have different vertical scales. Many of the parameters derivatives have similar shapes, suggesting that they will have a large "dot product" in Equation 3.58 and therefore large degeneracies between them. Note that the overall normalization of the curves is irrelevant as far as degeneracies are concerned. This is because two parameters with identically shaped curves but different normalizations are still completely degenerate as one can perfectly compensate for changes in one parameter with smaller (or larger) changes in the other. Thus, to quantify the degree of degeneracy, we can form a matrix of normalized dot products between the $\mathbf{s}_A$ vectors, where the magnitude of each vector is individually normalized to unity. The results are shown in Table 3.2, where we see that the parameters form three degenerate groups — the normalization parameters (B, γ, A_{sync}, A_{ff}), forming a degenerate 4×4 block of $\approx$1's in the top left corner; the spectral index parameters (α_{ps}, α_{sync}, α_{ff}), forming a degenerate block in the middle; and the frequency coherence parameters (σ_α, $\Delta\alpha_{sync}$, $\Delta\alpha_{ff}$), forming a degenerate block in the bottom right corner. Indeed, computing the eigenvalues of the normalized Fisher matrix (*i.e.* of Table 3.2), one finds only three eigenvalues of order unity and higher, with the next largest eigenvalue of order 10^{-3}.

	B	γ	A_{sync}	A_{ff}	α_{ps}	α_{sync}	α_{ff}	σ_α	$\Delta\alpha_{sync}$	$\Delta\alpha_{ff}$
θ'_1	0.368	0.368	0.374	0.36	0.14	0.164	0.108	0.366	0.365	0.366
θ'_2	0.184	0.184	0.155	0.219	-0.532	-0.522	-0.542	-0.053	-0.088	-0.005
θ'_3	0.283	0.283	0.293	0.275	0.167	0.13	0.215	-0.442	-0.429	-0.455
θ'_4	0.006	0.006	0.002	0.01	-0.007	-0.046	-0.049	0.089	0.665	-0.739
θ'_5	-0.001	-0.001	0.001	0.	0.002	0.088	-0.087	-0.81	0.472	0.327
θ'_6	-0.001	-0.001	-0.018	0.019	0.81	-0.293	-0.506	0.004	-0.041	0.008
θ'_7	0.247	0.248	-0.125	-0.369	-0.097	0.646	-0.534	0.05	-0.081	-0.072
θ'_8	0.408	0.41	-0.236	-0.58	0.057	-0.411	0.311	-0.033	0.049	0.043
θ'_9	0.181	0.114	-0.823	0.524	0.004	0.048	0.026	0.	-0.003	0.
θ'_{10}	0.7	-0.712	0.038	-0.026	0.	-0.002	-0.001	0.	0.	0.

Table 3.3: Eigenvectors (Equation 3.60) of the normalized Fisher matrix (Table 3.2). Each row represents an eigenvector, and going from top to bottom the eigenvectors are arranged in descending value of eigenvalue.

This suggests that there are really only three independent foreground parameters that can be measured.

To gain an intuition for what these independent parameters quantify, consider the first few eigenvectors of the normalized Fisher matrix $\widehat{\mathbf{F}}$ shown in Table 3.2, that is, vectors that satisfy

$$\widehat{\mathbf{F}}\theta'_{\mathbf{n}} = \lambda_n \theta'_{\mathbf{n}}. \tag{3.60}$$

These parameter eigenvectors are listed in Table 3.4, where each row gives the weighted average of the original parameters that one must take to form the "eigenparameters". To see what foreground parameters are well characterized by 21 cm tomography, we look at the first few eigenvectors, which can be measured with the highest signal-to-noise ratio:

$$\begin{aligned}
\theta'_1 &\approx 0.4(B+\gamma+\sigma_\alpha+A_{sync}+A_{ff}+\Delta\alpha_{sync}+\Delta\alpha_{ff}) \\
&\quad +0.1(\alpha_{ps}+\alpha_{ff})+0.2\alpha_{sync} \\
\theta'_2 &\approx 0.2(B+\gamma+A_{ff}+A_{sync})-0.1(\sigma_\alpha+\Delta\alpha_{sync}) \\
&\quad -0.5(\alpha_{ps}+\alpha_{sync}+\alpha_{ff})-0.02\Delta\alpha_{ff} \\
\theta'_3 &\approx 0.3(B+\gamma+A_{sync}+A_{ff})+0.2(\alpha_{ps}+\alpha_{ff})+0.1\alpha_{sync} \\
&\quad -0.4(\sigma_\alpha+\Delta\alpha_{sync})-0.5\Delta\alpha_{ff},
\end{aligned} \tag{3.61}$$

where for clarity the coefficients have been rounded to one significant figure (see Table 3.4 for more precise values). The first eigenparameter weights all the parameters equally except for the spectral indices, which are somewhat downweighted. The

first eigenparameter is thus roughly an "everything but spectral indices" parameter. Crudely speaking, this parameter is a generalized normalization parameter because it is mostly comprised of normalization parameters (B, γ, A_{sync}, A_{ff}) and frequency coherence parameters, whose spectral effects are not apparent until the frequency range is large. Examining the third eigenvector, we see that it is most heavily weighted towards the frequency coherence parameters, suggesting that we do have an independent parameter that acts as a "generalized frequency coherence". However, being the third eigenvector, our ability to make measurements of it is lower than with the first eigenvector or the second eigenvector, which is a "generalized spectral index".

To see which degrees of freedom we cannot constrain in our foreground model, we consider the last couple of eigenparameters:

$$\begin{aligned} \theta'_9 &\approx 0.18B + 0.11\gamma - 0.82A_{sync} + 0.52A_{ff} \\ \theta'_{10} &\approx 0.7B - 0.71\gamma, \end{aligned} \tag{3.62}$$

where we have omitted all terms with coefficients less than 0.1. The last eigenvector is dominated by B and γ, which appear with almost equal and opposite magnitudes. This suggests that differences between B and γ are extremely hard to measure. In the penultimate eigenvector, the sum of coefficients for B, γ, and A_{ff} is roughly equal and opposite to the coefficient of A_{sync}. This implies that the collective difference between A_{sync} and the linear combination of B, γ, and A_{ff} is also difficult to tease out from the data, but less so than the difference between B and γ.

That we can only measure a small number of foreground parameters independently is perhaps unsurprising, given that in Figure 3-1 we saw that most of the foreground power comes from a small number of eigenforeground modes that lack distinctive features. Note that since κ enters our expression for the foreground parameter Fisher matrix (Equation 3.58) as an overall multiplicative constant, instrumental noise has no effect on the foreground parameter degeneracies we examined in this section. The parameter degeneracies are due purely to the form of the foregrounds, and have nothing to do with the noise. On the other hand, when quantifying the effective number of parameters in the Wiener filter (Equation 3.42) or when solving for the optimal number of eigenforeground modes to measure in the truncated least-squares method (Equation 3.48), our expressions explicitly depend on the noise level. For instance, in the limit of a noiseless experiment the best characterization of foregrounds is obtained by measuring all the eigenmodes, not the three or four modes found in Sections 3.3.4 and 3.3.4. In general, however, the results of this section show that one

must be able to measure at least three eigenmodes at a reasonable signal-to-noise in order to adequately model the foreground spectrum. If Equation 3.42 or Equation 3.48 suggest that fewer eigenmodes should be used in the model, it is simply because one's experiment is too noisy to allow the foregrounds to be measured accurately.

3.5 Conclusions

In this chapter, we have shown that despite the complicated dependence that low-frequency radio foregrounds may have on physical parameters, the resulting spectra are always rather generic and featureless. The bad news from this came in Section 3.4, where we saw that even the most careful foreground spectrum measurements are unlikely to yield interesting constraints on foreground physics, thanks to high levels of degeneracies between different foreground parameters. This, however, is good news for those who simply consider foregrounds an impediment to 21 cm tomography. In Section 3.3 we saw that it was precisely because the foreground spectra were so featureless that they could be characterized to a greater accuracy than necessary for foreground subtraction using only three or four independent parameters. This bodes well for 21 cm astrophysics and cosmology, for it suggests that extremely detailed and physically motivated foreground models will not be necessary for successful foreground cleaning.

Chapter 4

Will point sources spoil 21 cm tomography?

4.1 Introduction

In Chapter 3, we discussed the problem of foreground modeling. We now turn to actual foreground subtraction. Foreground subtraction in 21-cm tomography is particularly challenging because current-generation instruments are not optimized for imaging. Instead, most experiments are designed to provide cosmological information through *statistical* measurements such as the power spectrum. Images of the 21-cm sky are expected to be of a rather poor quality in terms of both signal-to-noise and synthesized beam, and image-processing techniques devised to optimally mitigate this problem may be too computationally intensive to be useful.

In this chapter, we take a conservative approach and investigate how well foreground cleaning can be done with a computationally simple method using only the the so-called *dirty maps* – distorted maps of the sky as seen by an instrument at many different frequencies. Specifically, we clean out the foregrounds in any given sky direction by fitting a model of the contamination to the corresponding sky pixels in all the dirty maps. This approach of cleaning one sky pixel at a time along the line-of-sight allows one to take advantage of the high spectral resolution offered by most 21 cm tomography experiments. The key idea is that explored in Wang et al. (2006): the point source emission (mainly synchrotron radiation from distant galaxies) is expected to vary smoothly with frequency while the redshifted 21 cm signal of interest varies rapidly, because a small change in frequency (and hence redshift) corresponds to many Mpc in the radial direction and hence a major difference in local

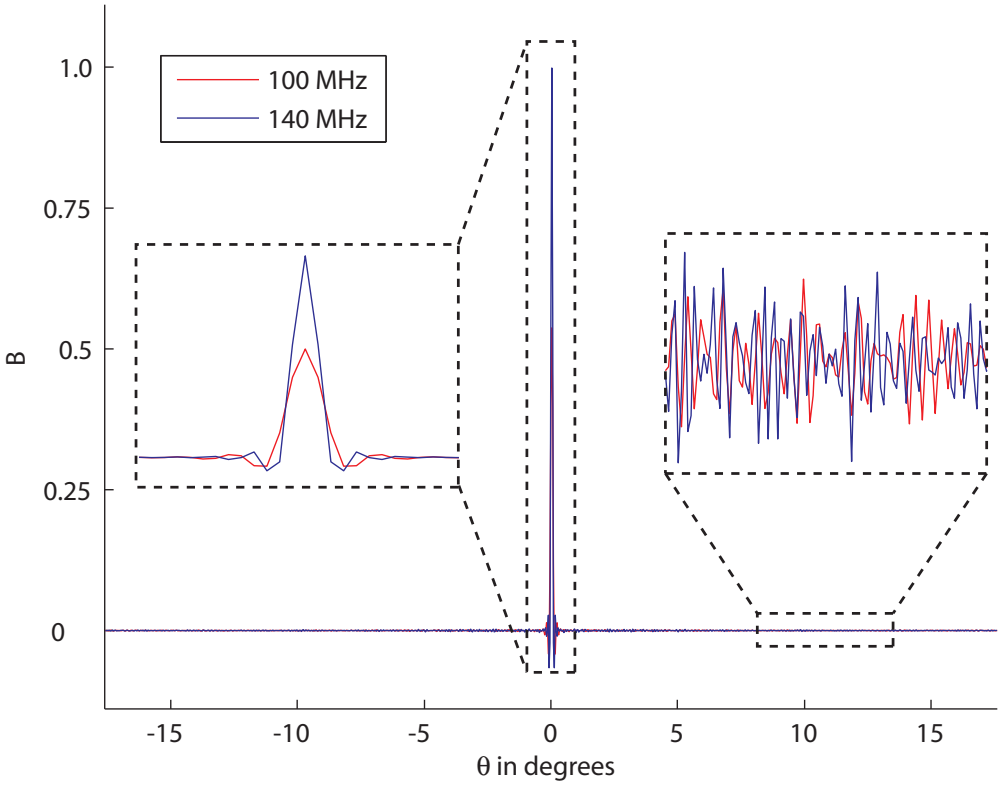

Figure 4-1: Sample synthesized beam profile for a typical 21-cm tomography experiment with 500 antenna tiles. The tiles are distributed within a diameter $D_{\rm array} \sim 1500\,{\rm m}$ according to the density function $\rho(r) \sim r^{-2}$. From the profile, it is clear that a realistic evaluation of foreground subtraction techniques should take into account the fact that beam widths vary as λ/D. This is particularly important when subtracting off unresolved point sources, because a point source can fall on an off-beam "spike" that has an effectively unpredictable dependence on frequency.

density fluctuations etc. High spectral resolution should thus allow one to separate out a rapidly oscillating cosmological signal from the spectrally smooth foreground using a low-order polynomial fit.

While previous studies (Wang et al., 2006; Jelic et al., 2008) have already highlighted the potential of this approach, such studies are incomplete because they do not take frequency dependent beam effects into account. At a given frequency, the dirty map is the true sky multiplied by a broad (say 10°) function known in radio astronomy as the *primary beam* and then convolved with a function referred to as the *synthesized beam*. The primary beam width is of order $\theta \sim \lambda/D_{\rm antenna}$, where $D_{\rm antenna}$ is the physical size of the individual array elements, whereas the synthesized beam

is the Fourier transform of the array layout. Figure 4-1 shows a synthesized beam example illustrating several key features:

- There is a narrow central peak whose width is of order $\theta \sim \lambda/D_{\rm array}$, where $D_{\rm array}$ is the size of the whole antenna array. This peak width is usually referred to as the angular resolution of the experiment.

- There is positive and negative "frizz" extending far beyond the central peak, corresponding to incomplete coverage of the Fourier plane, causing random-looking low amplitude oscillations on the angular resolution scale.

- The synthesized profile is the same at all frequencies except for an overall scaling with the wavelength λ.

As one shifts to higher frequencies, the synthesized beam therefore shrinks like ν^{-1}, causing the contributions from individual point sources to oscillate rapidly as positive and negative parts of the frizz moves across them. This produces an effective foreground signal looking not like a smooth synchrotron spectrum, but more like the rapidly oscillating cosmological signal.

Early work on this problem has provided cautionary but encouraging results (Bowman, 2007). Our goal in this chapter is to quantify in detail how well the simple pixel-by-pixel foreground subtraction strategy can deal with this problem, using simulations including realistic radio arrays and beams. Bowman et al. (2009) have independently studied this problem with a complementary approach, and we compare our results with theirs below. Whereas they perform a detailed case study of how well the MWA experiment can meet this challenge, including all foregrounds, we instead tackle the broader question of how residual point source contamination depends on experimental specifications, and what the experimental design implications are. We ignore non-point-source contributions to the foregrounds, with the expectation that such components (such as Galactic synchrotron radiation) are rather benign relative to the unresolved sources. Being spatially smooth, the off-beam contributions of such foregrounds tend to average out, unlike the point sources, even though their expected amplitude is larger. Bowman et al. (2009) confirm this expectation.

In this chapter, we deal only with the removal of *unresolved* point sources, assuming that resolved and detected sources can be eliminated by standard radio astronomy techniques as well as discarding the most contaminated pixels. We thus envision the cleaning of foregrounds as a two-step process: first masking or deconvolving out bright resolvable point sources that exceed some flux cut, then proceeding with our proposed

algorithm. We focus solely on the viability of the second step, but explore how the results depend on the flux cut from the first step.

The rest of this chapter is organized as follows. In Section 4.2, we outline the simulation methodology. We go through the simulation of foregrounds, the simulation of radio interferometers, and the proposed foreground subtraction strategy itself. The overall results of our analysis are presented in Section 4.3.2, where we consider three possible scenarios for foreground subtraction, ranging from the most pessimistic to the most optimistic. In Section 4.3.3, we quantify which design and algorithm choices are most important, and examine how the interplay between various instrumental and algorithmic parameters gives rise to these results. Table 4.1 lists the parameters that we explore as well as the ranges over which we vary them. In Section 4.4, we summarize our results and discuss future prospects.

Assumptions		**Low-Performance Extreme**	**Fiducial Model**	**High-Performance Extreme**
Experimental	Tile Arrangement	$\rho(r) \sim r^{-2}$	$\rho(r) \sim r^{-2}$	Monolithic with tiles separated by 40 m
	Rotation synthesis	None	6 hours, continuous	6 hours, continuous
	Noise level	$\sigma_T \sim 1\,\mathrm{mK}$	Noiseless [1]	Noiseless
Analysis	Primary beam width adjustments	None	None	Adjusted to be frequency-independent
	Bright point source flux cut S_{cut}	100 mJy	10 mJy	0.1 mJy
	synthesized beam width adjustments	None	None	Resolutions equalised by extra smoothing
	u-v plane weighting	None (natural)	Uniform	Uniform
	Order of polynomial fit	Constant	Quadratic	Quintic
	Range of polynomial fit	80 MHz	2.4 MHz	2.4 MHz

Table 4.1: Range of our parameter space exploration for foreground cleaning. Parameters pertaining both to experimental specifications and to analysis method impact the success of the cleaning.

4.2 Methodology

Our basic approach is to simulate a point source sky, simulate observed maps of it at multiple frequencies, clean these maps and quantify the residual contamination that remains after the cleaning. We repeat this a large number of times to explore the range design and algorithm options listed in Table 4.1.

A key feature of our cleaning method (which, again, is essentially that proposed in Wang et al. (2006), but with dirty map effects taken into account) is that the algorithm is blind, in the sense that our method does not rely on any physical modeling of foregrounds. This is important because the foregrounds relevant to 21-cm tomography are still relatively poorly understood (as compared to, say, the foregrounds relevant to cosmic microwave background experiments), and models are often based on interpolations and extrapolations from other frequency bands (de Oliveira-Costa et al., 2008). Because of this, we choose to use only the generic property that radio frequency foregrounds are smooth functions of frequency. This means that in our analysis, the foreground and instrumental *simulation* steps are not only separate from each other, but also completely decoupled from the foreground *subtraction* step. Thus, in what follows we divide our description of the methodology into each of those steps.

4.2.1 Step I: Simulation of Foregrounds

While point sources tend to cluster and are therefore not randomly distributed, the clustering is rather weak, and for simplicity we make the the assumption that they are uncorrelated. We thus simulate the contribution to each sky pixel independently, with its flux being the sum of the fluxes of a large number of of randomly generated point sources. The brightness of each point source is randomly drawn from the source count distribution

$$\frac{dn}{dS} = (4.0\,\mathrm{mJy}^{-1}\,\mathrm{sr}^{-1})\left(\frac{S}{880\,\mathrm{mJy}}\right)^{-1.75}, \tag{4.1}$$

[1]Our baseline simulations are run without noise because we show in Section 4.3.2 that the noise contribution can be included analytically and is unimportant for evaluating the effectiveness of our foreground subtraction strategies.

which is applicable at $\nu_* \equiv 150\,\mathrm{MHz}$ (Di Matteo et al., 2002; Lidz et al., 2008). The spectral dependence of each point source is given by

$$S(\nu) = S(\nu_*)\left(\frac{\nu}{\nu_*}\right)^{-\alpha}, \tag{4.2}$$

where the spectral index α is randomly chosen from a Gaussian distribution

$$p(\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\alpha-\alpha_0)^2}{2\sigma^2}\right], \tag{4.3}$$

with $\alpha_0 = 2.5$ and $\sigma = 0.5$ (Tegmark et al., 2000). Putting this all together, the brightness of each pixel is found by simply summing the simulated point sources within that pixel:

$$I = \left(\frac{dB}{dT}\right)^{-1}\Omega_{sky}^{-1}\sum_{i=1}^{N} S_i^*\left(\frac{\nu}{\nu_*}\right)^{-\alpha_i}, \tag{4.4}$$

where S_i^* is the flux of the i^{th} point source at 150 MHz, and $dB/dT = 6.9\times10^5(\nu/\nu_*)^2\,\mathrm{mJy\,K^{-1}\,sr^{-1}}$ is the standard conversion between intensity and brightness temperature. We use a pixel size of $\Omega_{sky} = 4.3\,\mathrm{arcmin}^2$ in our simulations. In generating our point sources, equation 4.1 is truncated at some maximum flux $S_{max} = S_{cut}$ because the brightest point sources are assumed to be detected and separately removed as mentioned above. We leave the threshold S_{cut} as a free parameter, and we investigate how the effectiveness of our foreground subtraction algorithm changes as S_{cut} is varied from 0.1 mJy to 100 mJy. At the dim end of the distribution, we truncate at a minimum flux $S_{\min}$ because the source count distribution diverges as $S \to 0$. We find that it is unnecessary to have $S_{\min}$ be any lower than 10^{-3} mJy, as the total flux I has converged by then. These values for $S_{\min}$ and S_{max} mean that we typically simulate $N \sim 2000$ point sources point sources for the sum in equation 4.4.

All the simulations were performed on 1024×1024 grids, and we verified that increasing the resolution to 4096×4096 had essentially no effect on the results. In the line-of-sight direction, the lowest frequency that we simulate is $\nu = 158.73\,\mathrm{MHz}$ (corresponding to $z = 8$), and (except for in section 4.3.3) the frequency increases by $\Delta\nu = 0.03\,\mathrm{MHz}$ from one frequency slice to another over 80 slices. In section 4.3.3 we continue to simulate 80 slices starting at $\nu = 158.73\,\mathrm{MHz}$, but we increase $\Delta\nu$ to 0.3 MHz and 1.0 MHz to quantify the effect of increasing the range of the polynomial fit. A sample sky at $\nu = 159\,\mathrm{MHz}$ is shown in figure 4-2.

Performing the sum in equation 4.4 at all frequencies for all 1024^2 pixels would take on order a month with our software. We therefore accelerate our algorithm by

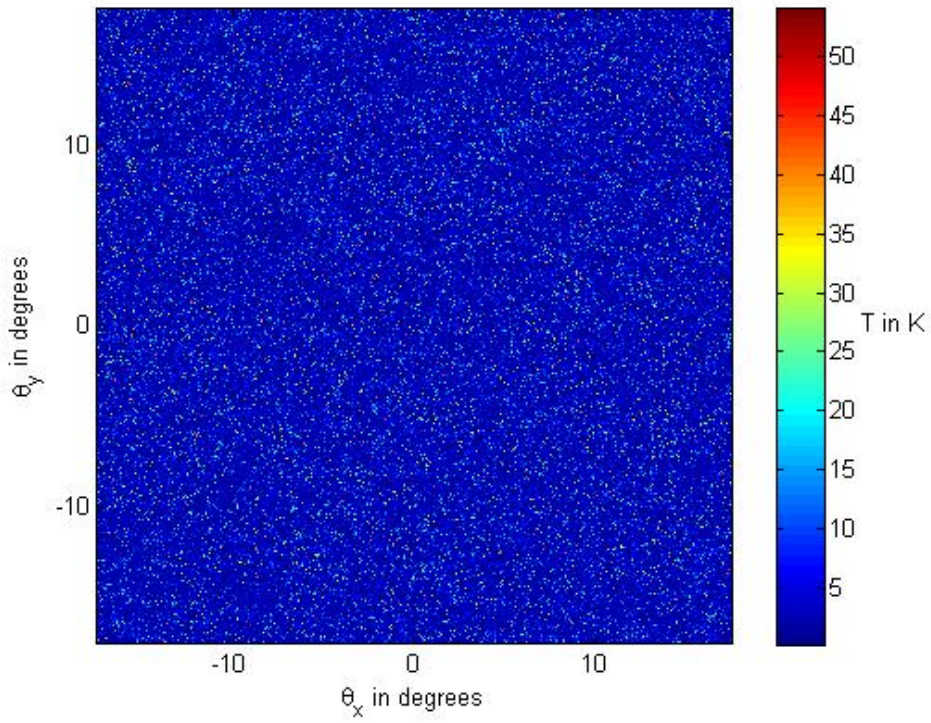

Figure 4-2: A sample sky with point source foregrounds at $\nu = 159\,\mathrm{MHz}$

exploiting the fact that the pixels are independent random variables. We first generate a database of 1000 pixels for which we compute the full frequency dependence. We then give each sky pixel the frequency dependence of a random pixel from the archive. It is easy to show that this procedure does not bias the residual power spectra low or high, but merely increases the variance in our estimation of these power spectra. By computing the scatter between multiple analyses using different random seeds and database sizes, we find that this excess scatter becomes unimportant for our purposes once the database exceeds a few hundred pixels.

We re-emphasize that the process used here pertains only to the *simulation* of foregrounds. The foreground subtraction itself is blind and therefore does not depend on the assumed properties of the foregrounds.

4.2.2 Step II: Simulation of Radio Interferometer

As mentioned above, the response of a radio interferometer array can be described by two beam functions: the primary beam and the synthesized beam. The primary beam function encodes the power response of an individual interferometer element to different parts of the sky, while the synthesized beam is a convolution kernel that

describes the interferometric effects of the array as a whole. To a good approximation, what an interferometer sees is not the true sky, but the true sky multiplied by the primary beam and then convolved with the synthesized beam.

In our simulations, we approximate the primary beam by a Gaussian. Its width scales as $\lambda/D_{\rm antenna}$, and is chosen so that the width matches simulated antenna patterns of the Murchison Widefield Array (MWA) dipole antennas at 150 MHz (approximately 30° full-width-half-max). The synthesized beam is found by taking the Fourier transform of the distribution of baselines (in the so-called u-v plane), which depends on the arrangement of tiles in the interferometer array. While our simulations of course require specific realizations of the array layout, we expect our results to hold for any radio array whose tile distribution is qualitatively similar to the various scenarios we consider in the results section. Roughly speaking, this means that our results should be applicable to any array with a relatively large number of short baselines and a number of long baselines that span up to $\sim 1\,$km, giving an angular resolution that is on the order of arcminutes. For an array with very different dimensions, our results can be straightforwardly scaled.

Integration time affects the response of the interferometer array in two ways that are important here. First, detector noise averages down with time as $t^{-1/2}$ as long as it is uncorrelated over different time periods. A typical point source sensitivity for a current-generation experiment like the MWA is $S = 0.27\,$mJy, with a 4.6 square arcminute pixel, and a 32 MHz bandwidth. This corresponds to an r.m.s. detector noise per a pixel of (Wang et al., 2006):

$$\sigma_T^{\rm MWA} = 0.22\,{\rm K}\left(\frac{32{\rm MHz}}{\Delta\nu}\right)^{1/2}\left(\frac{1\,{\rm hour}}{t}\right)^{1/2}, \tag{4.5}$$

where $\Delta\nu$ is the bandwidth, and t is the integration time. The MWA has a channel bandwidth of 32 kHz, and with ∼1000 hours of integration the detector noise level is $\sim 0.2\,$K. In our simulations we consider not only current-generation experiments with this level of noise, but also the noise levels of hypothetical future experiments, which we take to be of order $\sim 1\,$mK.

The second effect of taking measurements over an extended period of time pertains to the concept of rotation synthesis. The rotation of the Earth during observations means that interferometer baselines are not static points on the u-v plane, but instead sweep out of u-v tracks. This results in a change in the properties of the synthesized beam, as one can see in figure 4-3. In our analysis, we explore a range of rotation synthesis scenarios from none at all (i.e. taking a snapshot of the sky) to having 6

hours of rotation synthesis.

4.2.3 Step III: Foreground Subtraction

As mentioned above, we use a pixel-by-pixel line-of-sight cleaning strategy for removing the unresolved foregrounds below the flux threshold. For each pixel, its frequency dependence is fit by a low-order polynomial. Such a polynomial is by definition a smooth function, and thus the hope is that by subtracting off the fit from the signal, one subtracts off mainly the smooth foregrounds and not much of the cosmological signal, which is expected to oscillate wildly with frequency. We leave the order of the polynomial as a free parameter, and explore the effects of varying it from 0 (constant) to 5 (quintic). This involves a tradeoff between two separate effects. On one hand, the higher the order of the fit, the lower the residual foregrounds will be. On the other hand, high order fits come at the expense of fitting more power out of the cosmological signal. Crudely speaking, one therefore wishes to select the *lowest* order capable of adequately subtracting off foregrounds.

4.3 Results

In this section, we quantify the extent to which point source contamination can be removed with line-of-sight cleaning, and how this depends on the many parameters in Table 1. To get a sense of the extent to which the parameters matter, we then analyze three scenarios designed to bracket the range of possibilities in Section 4.3.1. As seen in Figure 4-4, this range is broad, extending all the way from utter failure to recover the cosmological signal to success in pushing foregrounds two orders of magnitude below the signal. After discussing some general findings that are independent of all these parameters in Section 4.3.2, we devote Section 4.3.3 to examining the parameters one by one to quantify which ones make the greatest difference.

4.3.1 Three scenarios

We now analyze three scenarios designed to bracket the range of possibilities:

1. **Pessimistic scenario (PESS)**

    - *Experimental assumptions:* In this scenario, we simulate an array whose tiles are arranged in a radial density profile that goes as r^{-2}. The array

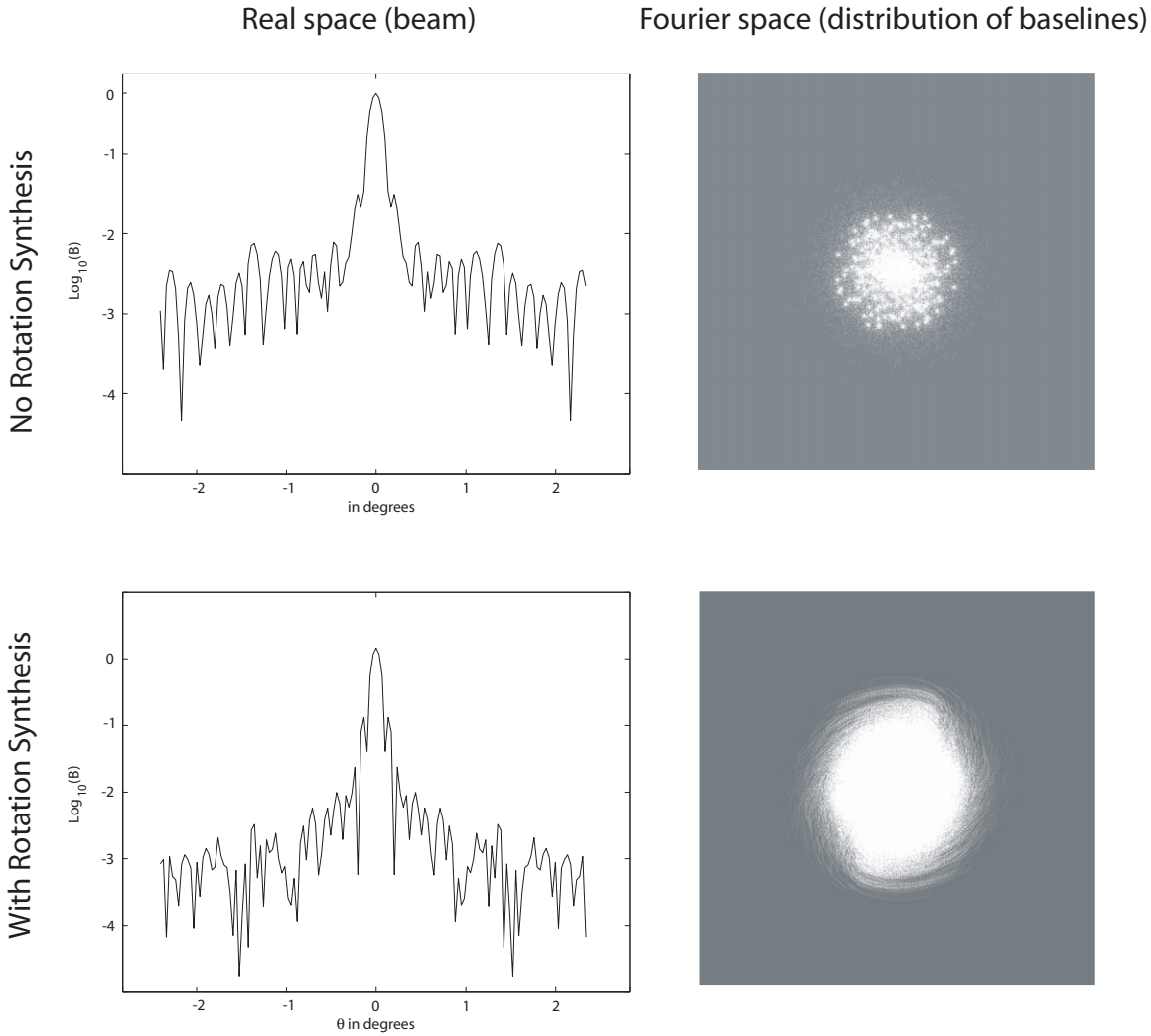

Figure 4-3: The left hand column shows sample beam profiles (a real-space description of the beam) while the right hand column shows the corresponding u-v distribution of baselines (a Fourier-space description of the beam). The top row illustrates an array with no rotation synthesis, while the bottom row shows an array with 6 hours of rotation synthesis. The real-space beams are normalized so that their peaks are at 1.

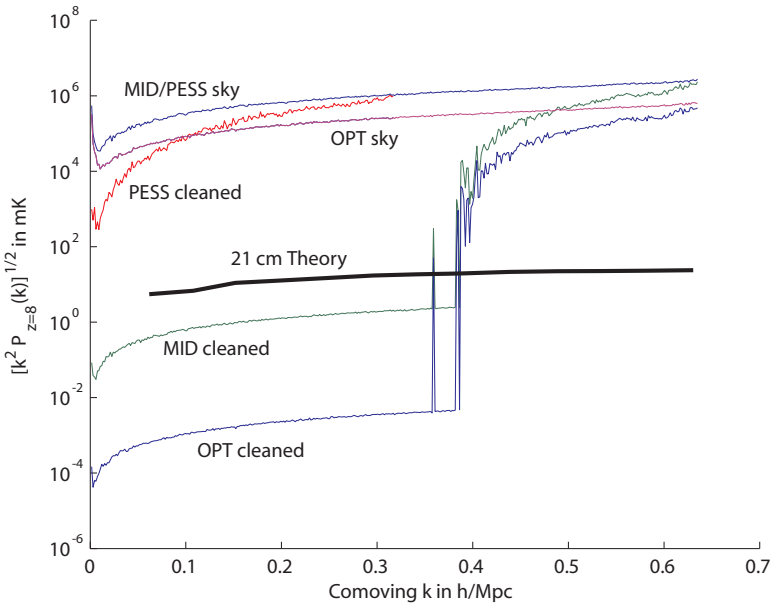

Figure 4-4: 2D power spectra of foregrounds and foreground residuals for the various scenarios outlined in the text. It is clear that except for the most pessimistic scenario, a dirty-map pixel-by-pixel cleaning strategy can be used to get within at least striking distance of the cosmological signal. As we will explain in Chapter 5, the rise in foreground residuals towards higher k is due to incomplete u-v coverage as one moves away from the origin in u-v space. The spike just prior to the general rise is due to the specific baseline distribution used in this chapter's simulations. The u-v coverage simply happens to have a hole at some location closer to the origin than the beginning of the general "thinning" of baselines towards the edge of the coverage, giving rise to a premature spike

takes a single snapshot of the sky.

- *Analysis:* No extra weighting is imposed on the u-v plane (see Section 4.3.3 for details on weighting schemes), and no attempt is made to adjust for the fact that both the primary beam and the synthesized beam have frequency-dependent widths. It is assumed that only sources that are 10 mJy or brighter can be removed prior to the subtraction of unresolved point sources. We also assume that the highest order polynomial that one can fit without taking out significant power from the cosmic signal is a quadratic.

2. **Fiducial scenario (MID)** — this scenario is designed to be reasonably representative of current-generation experiments.

   - *Experimental assumptions:* Same as PESS except that the array samples the u-v plane continuously for 6 hours.
   - *Analysis:* Same as PESS except that the u-v plane measurements are weighted so that all parts of the u-v plane that are covered are given a uniform weight.

3. **Optimistic scenario (OPT)**

   - *Experimental assumptions:* Same as MID.
   - *Analysis:* Same as MID except that we assume that bright point sources can be removed down to 1.0 mJy, and that one can safely fit up to a cubic polynomial.

Our fiducial model is a middle-of-the-road (MID) case intended to be representative of current generation experiments, while the pessimistic (PESS) and optimistic (OPT) scenarios refer to worst and best case experimental expectations, respectively. The PESS and OPT scenarios in general differ less than the low-performance and high-performance extremes outlined in Table 4.1 because the parameters of these extreme scenarios are in some cases unrealistic, and are considered in Section 4.3.3 solely to better understand how the various parameters affect foreground subtraction.

As our measure of how well or poorly the cleaning works, we use the two-dimensional spatial power spectrum of the cleaned map. We do this because the power spectrum of the cosmological signal is the key quantity that the 21 cm tomography community aims to measure in the near term, and the residual point source power spectrum will simply add to this. In all our power spectrum plots (starting with Figure 4-4),

we have removed the distorting effect of convolution with the synthesized beam, so that $P_{2D}(k)$ for the point sources is a constant (k-independent) white noise spectrum. The detector noise also has a constant spectrum, and the cosmological 21 cm signal plotted is what would be seen by an ideal instrument making a distortion-free image of the sky. The deconvolution step is trivial to perform in Fourier space, corresponding simply to dividing the measured power spectrum by the radial distribution of baselines.

Rather than plot $P_{2D}(k)$ itself, we plot the dimensionless quantity $[k^2 P_{2D}(k)]^{1/2}$, which can be roughly interpreted as the dimensionless fluctuation lever, analogous to the standard quantity $\delta_T/T \propto [\ell^2 C_\ell]^2$ for the cosmic microwave background and the quantity $\Delta \propto [k^3 P_{3D}(k)]^{1/2}$ for three-dimensional galaxy surveys. In the figures that follow, $[k^2 P_{2D}(k)]^{1/2}$ is always plotted for the frequency slice at $\nu = 158.73\,\mathrm{MHz}$. This translates to a redshift $z = 8$, around the "sweet spot" of most current generation 21 cm experiments.

Our cleaning method is linear, i.e., the "data cube" containing the cleaned maps at different frequencies is simply some linear combination of the input data cube, and the weights in this linear combination (implementing the polynomial fits) are fixed, independent of what the data cube contains. This means that if we have three data cubes, containing cosmological 21 cm signal, detector noise and point source signal, respectively, we get the same result if we sum them and then clean as if we first clean them individually and then sum them. Since these three components are all statistically independent, this also implies that their post-cleaning power spectra simply add. We take advantage of this fact (which we also verify numerically in Section 4.3.3) to simplify our analysis, performing most of our simulations with both the cosmological signal and the noise set to zero. A perfect foreground subtraction algorithm would thus produce a residual power spectrum that is zero everywhere.

Figure 4-4 shows the results. For comparison, the figure also shows the simulated EOR signal for $z = 8$ (taken from McQuinn et al. (2006a) and McQuinn et al. (2007)). Figure 4-4 shows that except for in the PESS scenario, our simple foreground subtraction algorithm is successful on large scales but not on small scales. Let us now focus on understanding this better, and clarifying the key properties of the cleaning method.

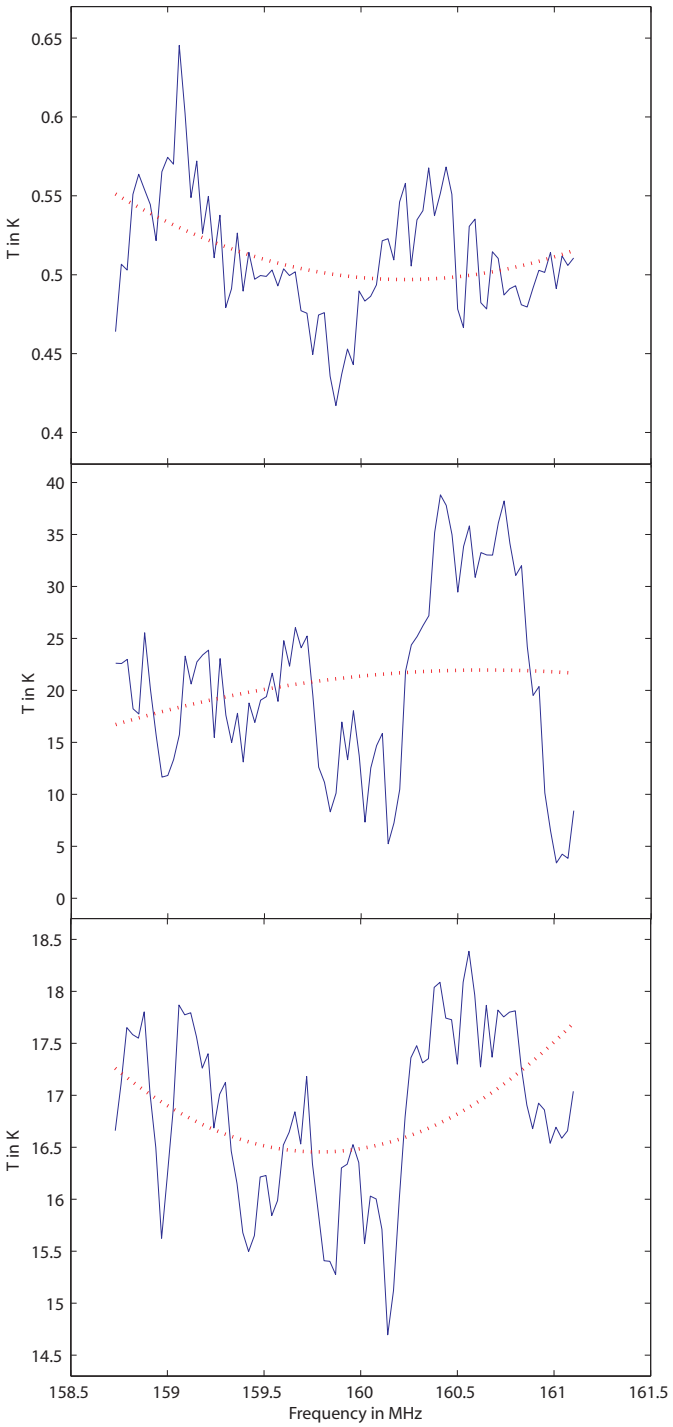

Figure 4-5: The spectra of a three typical dirty-map pixels, with fits given by the dotted curves. The top panel assumes no instrumental noise and no point sources brighter than 0.1 mJy. The middle panel assumes no instrumental noise and no point sources brighter than 100.0 mJy. The bottom panel assumes an instrumental noise level of $\sigma_T = 200$ mK and no point sources brighter than 0.1 mJy.

4.3.2 Why the method can work well on large scales

In the top panel of figure 4-5, we show the spectrum for a pixel randomly chosen from the dirty map produced by a noiseless but otherwise MWA-like instrument that has been continuously observing a fixed patch in the sky for 6 hours. A bright point source flux cut of $S_{cut} = 0.1\,\mathrm{mJy}$ has been assumed. In figure 4-5 we can see that while the overall trends of the spectrum can be captured by a low-order polynomial fit, the finer features generally remain as residuals.

The effectiveness of the polynomial fit depends on a variety of factors. We find that the greatest variation comes from varying S_{cut} and the noise level. In the middle panel of figure 4-5, we show a typical pixel with $S_{cut} = 100.0\,\mathrm{mJy}$[2]. Similarly, the addition of noise can decrease the quality of the fit. In bottom panel of figure 4-5, we have added $\sigma_T = 200\,\mathrm{mK}$ (note the larger amplitudes on the y-axis compared to what is shown in the top panel).

From figure 4-5, the reader may be surprised that our subtraction strategy works at all. Indeed, it seems from the figure that low order polynomial fits do extremely poorly, and that most of the foregrounds will remain after the subtraction. So why, then, does Figure 4-4 show that the foregrounds can be suppressed by many orders of magnitude?

The above-mentioned linearity holds the key to the success. To understand why, consider again the 3-dimensional data cube, and imagine it arranged so that the 2-dimensional dirty maps are all horizontal, stacked vertically on top of one another so that the vertical direction corresponds to frequency. To generate a plot such as Figure 4-4, we perform two operations on this cube of simulated data:

1. Clean one pixel (vertical column) at a time

2. Fourier transform one frequency map (horizontal slice) at a time

Since both of these steps are linear, they can each be thought of as multiplying all the data (rearranged into a single vector) by some matrix. Since the two steps mix data purely horizontally and purely vertically, respectively, it is easy to see that the corresponding two matrices must commute. In other words, we get the same result if we perform the steps in the opposite order (our simulations confirm this). An analogous and more familiar example is the 3-dimensional Fourier transform, which decomposes into successive Fourier transforms in the vertical and two horizontal

[2]Note that in general, dirty-map pixels may be negative. Such negative brightness temperatures arise from the fact that in radio arrays, the signal from a single tile is never correlated with itself. We thus never sample the origin of the u-v plane, and therefore lose the mean of the signal.

directions, and again gives the same result regardless of the order in which these operations are performed.

This means that we get the exact same result in Figure 4-4 if we first Fourier transform the maps, then perform the cleaning one Fourier mode at a time instead of one pixel at a time. Figure 4-6 shows the spectral fit to a sample pixel in the Fourier (u-v) plane, revealing a very smooth curve that can be fit with exquisite accuracy. The success for this particular Fourier mode hinges on the fact that it lies in the central part of the Fourier plane which has been well sampled by the array at all frequencies. This explains the low plateau in the residual power spectra in the left side of Figure 4-4. The sharp rise in residual power on smaller scales corresponds to incomplete Fourier coverage, with certain interferometric baselines missing. Converting this intuitive understanding back to real space, the small-scale synthesized beam frizz seen in Figure 4-1 is largely averaged away when expanding the sky into long-wavelength Fourier modes, whereas the small-scale modes are severely affected. The characteristic scale separating "short" and "long" Fourier modes is determined by the *longest baseline radius for which u-v coverage is complete*. As one moves outwards from the origin in Fourier space, one eventually reaches a pixel where u-v coverage is incomplete at some frequency along the line of sight. Once there is any such missing information along the frequency direction, our simple fitting algorithms fail to find a good fit and the result is a large increase in residuals. Since this increase is caused by incomplete u-v coverage, the scale at which this occurs depends on the rotation synthesis scenario being considered. We find, however, that for most reasonable cases the upswing takes place between $k = 0.3h/\mathrm{Mpc}$ and $0.4h/\mathrm{Mpc}$.

The reader may also be concerned (based on figure 4-5) that there may be much residual power in the frequency direction, leading to problems for experiments aiming to measure the 3D power spectrum. A full discussion of the 3D power spectrum is beyond the scope of this chapter, but for now we note that plots like figure 4-5 are simply *real space* plots in the line-of-sight direction. Fourier transforming in the perpendicular direction, the fit becomes quite good in the central parts of the $u - v$ plane as mentioned above. Fourier transforming in all three directions (which is what one would ultimately do in order to find estimate the 3D power spectrum), one does not expect the residual power to depend on both the line-of-sight and transverse wave number. Indeed, Bowman et al. (2009) find that while there is some contamination along the line-of-sight, there do exist clean regions in the full 3D Fourier space.

Aside from residual foregrounds, another problem that may arise is the fact that blind algorithms such as ours have no way to distinguish between signal and fore-

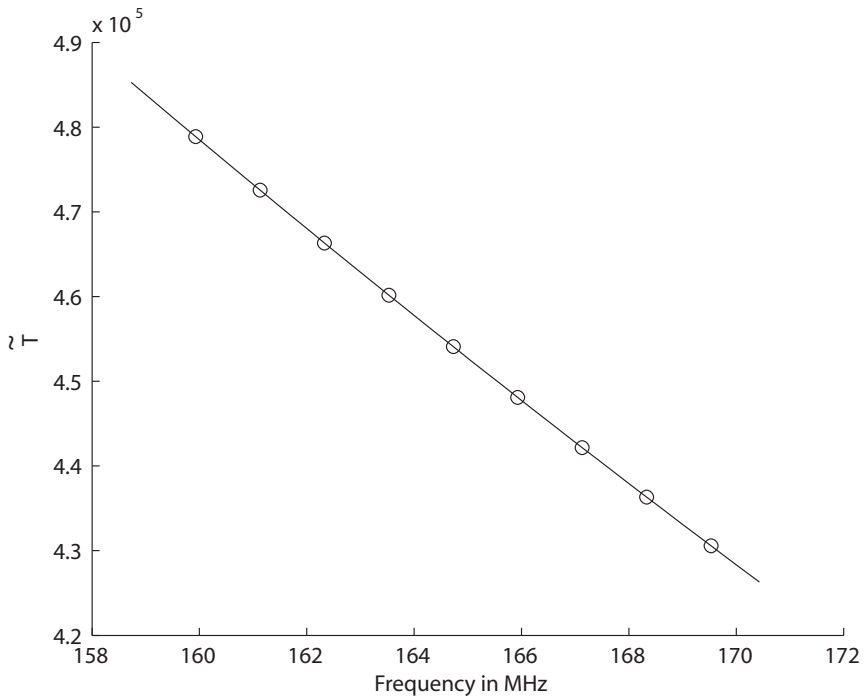

Figure 4-6: The spectrum of a Fourier space pixel, taken from the part of the plane where u-v coverage is complete. A fit to the spectrum is shown.

ground except for through differences in smoothness as a function of frequency. Thus, if the signal possesses any smooth large-scale component, the foreground subtraction algorithm may inadvertently remove this component of the signal with the foregrounds. Deciding whether this is a problem or not (and quantifying the seriousness if it *is* indeed a problem) requires the detailed examination of the properties of the signal, which is beyond the scope of this chapter. Intuitively, however, one can say (at the very least) that there will be no loss of cosmological signal in the transverse Fourier components, since our algorithm subtracts along the line-of-sight.

4.3.3 Exploration of parameter space

Above we saw that there was a huge difference in foreground removal ability between the three scenarios. Let us now investigate which of the instrumental and algorithmic parameters in Table 4.1 make the greatest difference. We do this by varying one of them at a time, over the range given in Table 4.1, while keeping all others fixed.

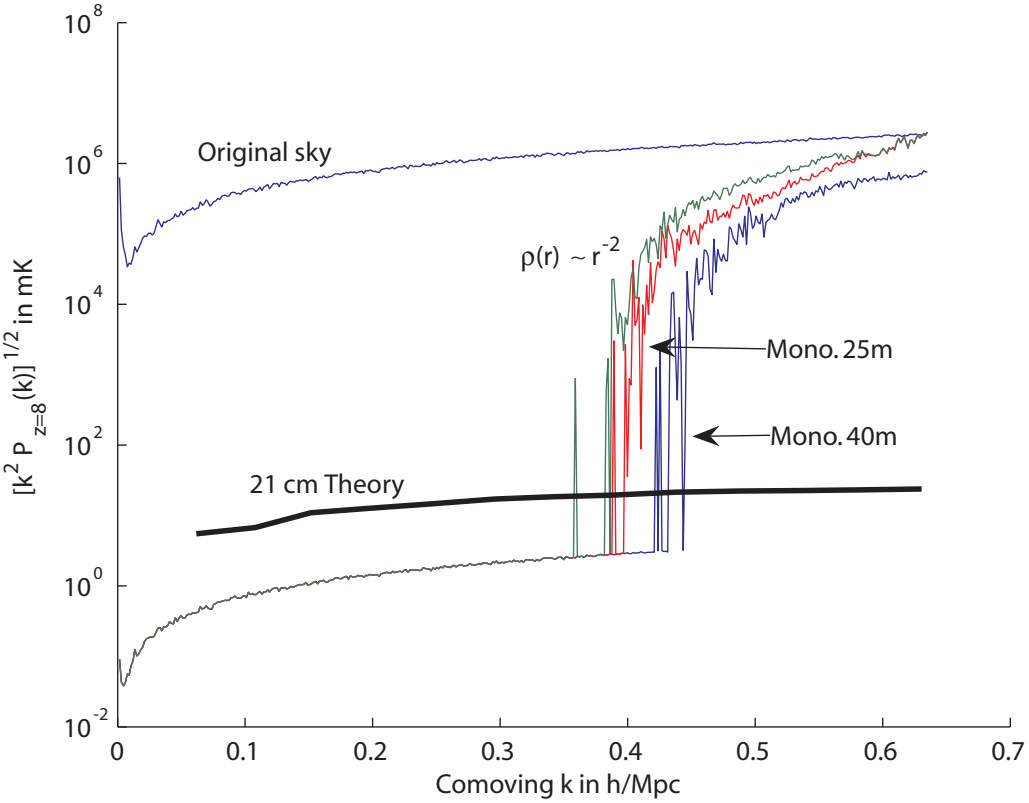

Figure 4-7: Dependence on array layout: 2D power spectra for the fiducial model with various arrangements of tiles. Monolithic and r^{-2} arrangements both seem to work well.

Array Layout

In a radio array, it is the arrangement of antenna elements that determines the distributions of baselines, and it is the distribution of baselines that determines the shape of the synthesized beam. Thus, our ability to subtract foregrounds depends strongly on the layout of the array.

In an experiment such as the MWA, each antenna element consists of a tile of 16 crossed dipole antennas, arranged in a 4 by 4 grid with the dipoles spaced roughly 1 m from each other. In this configuration, the dipoles give rise to a complicated primary beam pattern which may include structures such as sidelobes. As mentioned in previous sections, we neglect these complications and assume that each tile of dipoles has a primary beam that takes the form of a Gaussian with a 30° full-width-half-max.

We consider two different types of tile arrangement. One has the density of tiles

varying as r^{-2}, where r is the radius from the centre of the array. The other, which we categorize as a "monolithic" arrangement, is a regular square grid densely packed near the centre of the array. A few tiles are placed farther from the centre to provide some long baselines for calibration purposes and for good angular resolution. In both cases, we have a total of 500 tiles in our simulations; for the monolithic cases 400 tiles form the central core, while 100 tiles scattered outside the core provide some short baselines (which can be formed between two close-by tiles that are both outside the core) and many long baselines (which are formed by pairing up a tile within the core to one outside the core). We consider two realizations of the monolithic cases, one with tiles in the central core separated by 25 m and with them separated by 40 m. In both cases, the combination of short-to-medium-length baselines formed between tiles in the core and the extra few short baselines provided by closely separated tiles outside the core allow complete coverage of the u-v plane near the origin with the help of rotation synthesis.

The results are shown in figure 4-7. From the plot, it is clear that good foreground subtraction can be done using either an r^{-2} arrangement or a monolithic arrangement, and that on large scales there is no difference in performance. Monolithic arrangements, however, seem capable of pushing foreground subtraction to slightly finer scales. The explanation for this is that with an r^{-2} arrangement, the high density of tiles at small r means that the inner parts of the u-v plane are vastly oversampled given the rotation synthesis, and one can spread out the tiles slightly to extend the perfectly covered regions of the u-v plane without introducing any gaps thanks to the tight monolithic core.

Noise Level

As mentioned above, we consider three different scenarios in our analysis of noise effects: a noise level $\sigma_T \sim 200\,\mathrm{mK}$ (representative of current-generation instruments), a noise level $\sigma_T \sim 1\,\mathrm{mK}$ (hopefully representative of next-generation instruments), and a hypothetical noiseless case. The resulting power spectra are shown in figure 4-8. For all but the noiseless case, there are in fact two sets of curves plotted, corresponding to two different computations—in one case, noise was included prior to the foreground subtraction, whereas in the other a noiseless foreground subtraction was performed, and the resulting power spectrum was added to a power spectrum of noise (which is a constant). One obtains the same result in both cases. Indeed, the two sets of curves lie on top of each other and are visually indistinguishable, confirming that our linear cleaning method leaves the noise power essentially unaffected. It is

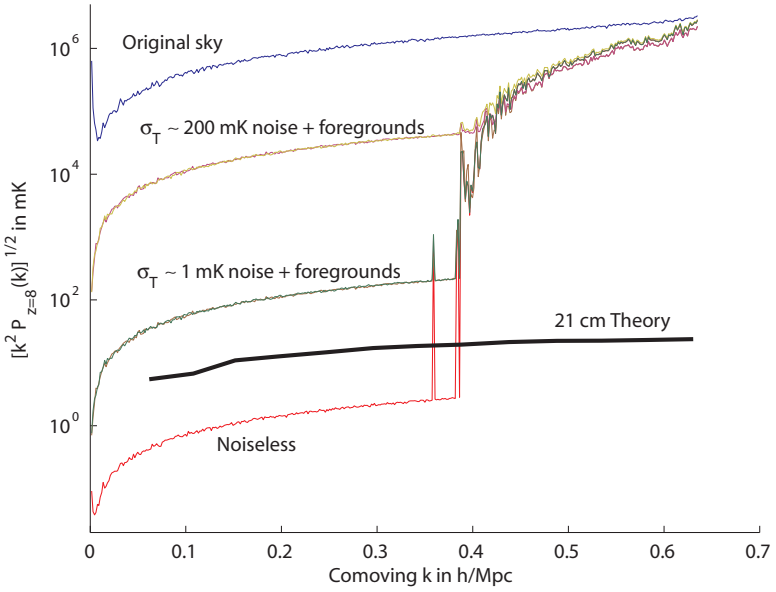

Figure 4-8: Effect of noise: 2D power spectra for the fiducial model with various levels of instrumental noise show that the power spectrum of detector noise simply gets added to the power spectrum of the residual foregrounds.

therefore unnecessary to run simulations with noise, since the results can be predicted analytically given a noiseless simulation.

A naive reading of Figure 4-8 suggests the pessimistic conclusion that current-generation 21 cm tomography experiments have no hope of seeing a cosmological signal. However, the additive contribution to the power spectrum from noise can be completely eliminated by measuring the power spectrum by cross correlating maps made at different times, since their noise will be uncorrelated. The WMAP team have successfully used an analogous procedure for noise bias removal, where they cross correlated maps made not at different times but with different receivers (Hinshaw & et. al. [WMAP collaboration], 2007). In contrast, the contribution from residual point sources can not be eliminated in this way.

Just like for WMAP, the noise will still contribute to the *error bars* $\Delta P(k)$ on the measured power spectrum, but these error bars can be shrunk by averaging many Fourier modes with comparable wave number k. For the case where noise and signal have a Gaussian distribution, $\Delta P(k) = \sqrt{2/N}P(k)$ where $P(k)$ is the total power spectrum (including noise and residual point sources) and N is the number of Fourier modes averaged. For example, the noise power exceeds the signal power around the third acoustic peak of the CMB power spectrum measured by WMAP, but N is large

enough that the error bars nonetheless become significantly smaller than the cosmic signal. In our case, binning radially in annuli of thickness $k = 0.1h/\mathrm{Mpc}$, we typically have $N \sim 10^4$ to 10^5 (depending the value of k), which would suggest percent level relative errors on the plotted power spectrum curves.

In summary, our results regarding noise are quite encouraging: the promising forecasts that have been made in the past for what can be learned from 21 cm cosmology all included detector noise, but not the point source contribution in the full complexity that we are modeling. Our results show that the fact the point sources can be removed without having much effect on the noise levels.

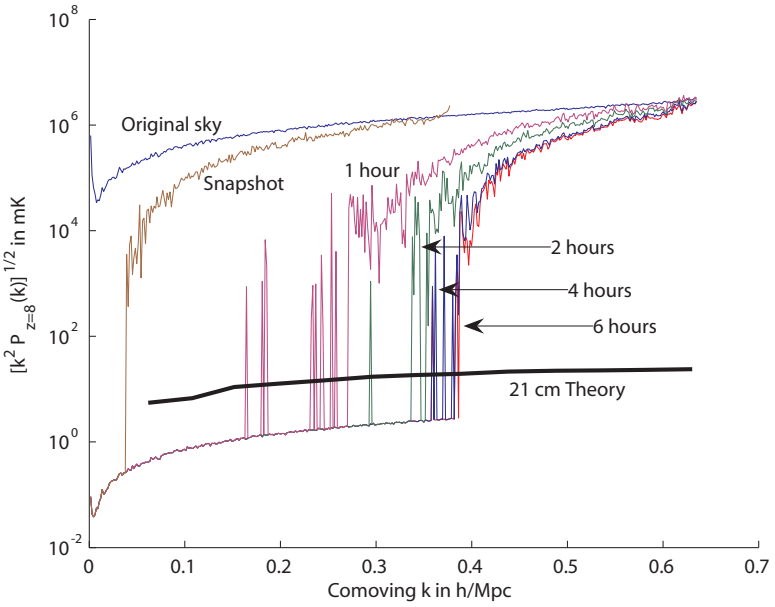

Figure 4-9: Effect of rotation synthesis: 2D power spectra are show for the fiducial model but with various total rotation synthesis times. The effect of lengthening integration time to increase u-v coverage is seen to saturate after about 4 hours.

Rotation Synthesis

In a maximally optimistic situation, rotation synthesis can dramatically boost one's ability to image the sky, since for every baseline, one effectively obtains an entire arc of baselines on the u-v plane. However, rotation synthesis is not always as readily available as one might hope. For instance, a declination zero object as viewed from the equator does not rotate but merely translates across the sky, allowing no rotation synthesis.

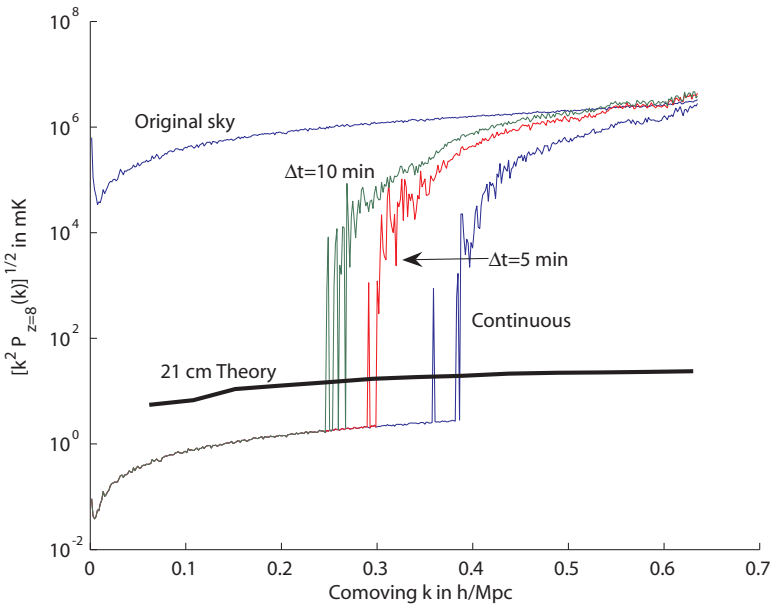

Figure 4-10: Effect of temporal binning: 2D power spectra for the fiducial model but with varying binning time Δt, showing that one should strive for continuous u-v coverage.

As another example, consider a patch of the sky that lies close to the horizon. The short time interval between the rising and setting of the patch means that even if continuous rotation synthesis is possible, one may not obtain sufficient total rotation synthesis time. (Observing the patch again on the next night does not alleviate the problem, for a sidereal day later one is simply sampling the same points on the u-v plane again). One can also imagine a situation where u-v coverage is poor even for a patch that remains in the sky for most of the night because of limitations in hardware calibration and data flow. In such a scenario, each baseline would sweep out a long track on the u-v plane, but this track would only be sparsely sampled by a series of snapshots of the sky. We thus use two separate parameters to parameterize the quality of rotation synthesis: the total rotation synthesis time and the time Δt between snapshots of the sky.

As expected, the performance of our foreground subtraction algorithm improves as one increases the total rotation synthesis time. In figure 4-9, we show the effect of allowing integration time to increase from a single snapshot to 6 hours (beyond which there is no significant improvement in foreground subtraction). Each curve was simulated by assuming continuous rotation synthesis (i.e. $\Delta t \to 0$) using an instrument located at latitude and longitude $(\lambda, \phi) = (-27°, 0°)$ for a field with right

ascension and declination $(\alpha, \delta) = (60^\circ, -30^\circ)$[3]. We see that the first few hours of integration give rise to dramatic improvements, and that the gains begin to saturate after about 4 hours. Intuitively, the saturation occurs because after several hours of integration, one is revisiting parts of the plane that have already been well sampled by other baselines, and so there is no further improvement[4]. As mentioned earlier in this section, the effectiveness of rotation synthesis depends on the location of the array as well as the sky coordinates of the patch being observed. We find, however, that our conclusion is fairly robust to changes in array location and sky coordinates. In other words, after 4 hours of observation one has essentially exhausted the potential of rotation synthesis, however small or large this potential may be.

In figure 4-10, we instead fix the integration time at 6 hours and vary Δt. It is clear that one should strive for continuous u-v coverage if possible. It should also be noted that with both parameters, it is generally the smallest scales that benefit from rotation synthesis – the r^{-2} array layout provides enough short baselines to ensure reasonable coverage on large angular scales even without exploiting Earth rotation.

Primary beam width adjustments

As mentioned in Section 4.2.2, the width of a radio array's primary beam is proportional to λ. Thus, even if one has perfect coverage of the u-v plane, the sky will look different at different frequencies. In doing foreground subtraction, we consider two possible ways to analyze the data:

1. Do nothing, and simply accept the fact that the primary beam width changes with frequency.

2. Smear the u-v plane data to smooth out the primary beams so that all the primary beams have the same width as the widest beam in the frequency range.

The results are shown in figure 4-11. A close look at the plot reveals that adjusting for the frequency dependence of the primary beam does in fact bring about an improvement, albeit a small one, in our ability to subtract foregrounds.

[3]The latitude, right ascension, and declination were chosen to match planned observations by the MWA. The longitude was set to zero for computational simplicity since its precise value does not result in qualitative changes in rotation synthesis.

[4]That does of course not imply that one should stop integrating after 4 hours of observation, for repeated measurements of the same u-v points increases signal-to-noise. Indeed, current observation plans call for thousands of hours of integration.

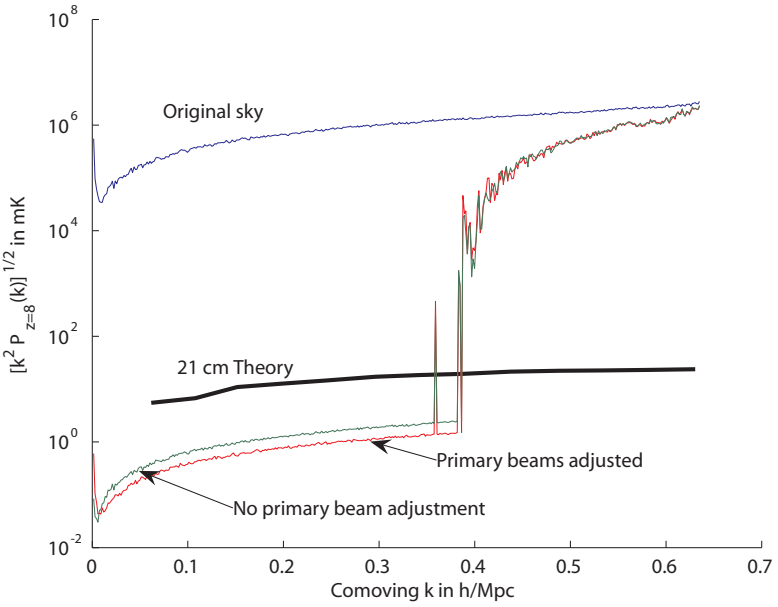

Figure 4-11: Effect of primary beam equalization: 2D power spectra for the fiducial model but with different algorithms for dealing with the fact that the primary beam width changes with frequency. Adjusting for the frequency dependence is seen to improve foreground subtraction slightly.

synthesized beam width adjustments

The synthesized beam width is also expected to scale as λ, as illustrated in figure 4-1. In addition to the "frizz" issue that we have focused on so far, a second potential cause for concern is the width change of the central peak, as it means that sources slightly off from the centre of the beam will appear dimmer at higher frequency. This can potentially degrade the spectral fits. One way of adjusting for this is to convolve the dirty maps with frequency-dependent kernels whose frequency dependence exactly compensates for the changing width of the central part of the synthesized beam. The effect of this extra step is to ensure that, aside from the frizz effect, the sky has been convolved with the exact same beam at all frequencies. In other words, the angular resolution is made frequency independent by degrading the resolution of all maps to that of the lowest frequency map. In figure 4-12, we show the profiles of the beams from figure 4-1 after convolution with Gaussians of appropriate widths.

As mentioned above, the procedure tested in this section deals with the central peak, and has no mitigating effect on the "frizz". Since the "frizz" is responsible for non-cosmological line-of-sight structure, one expects the beam adjustment to have very little effect in the frequency direction. Figure 4-13 shows how this beam adjust-

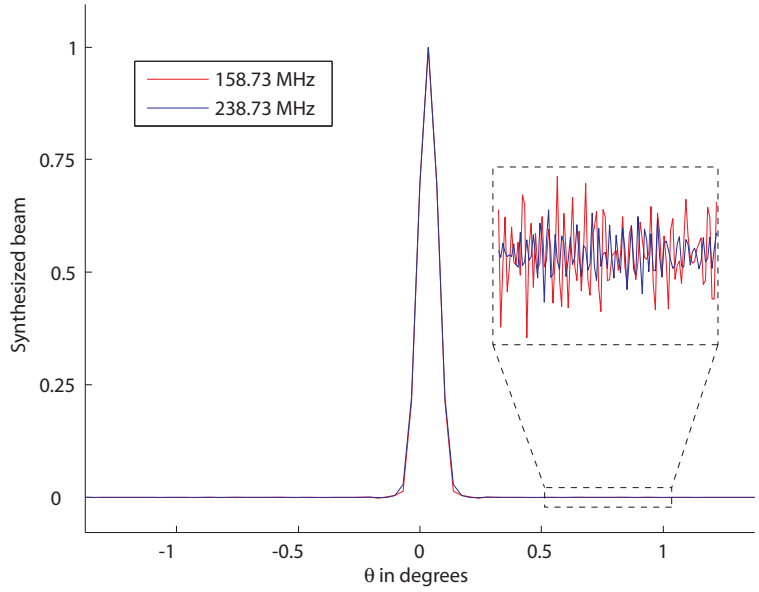

Figure 4-12: Beam profiles after an extra Gaussian convolution, designed to make the heights and widths of the central peaks frequency-independent. Parts of the beam beyond the central peak, however, will in general remain dependent on frequency.

ment affects the the foreground subtraction in the spatial directions. On the largest and smallest scales, the extra convolution is seen to have no effect, and on intermediate scales smaller scales it is seen to make things slightly worse. This procedure of adjusting for changes in angular resolution with frequency therefore does more harm than good. The harm presumably comes from the exacerbating the frizz-related problems (by causing departures from uniform weighting in the Fourier plane discussed in Section 4.3.3 below). The good comes from correcting for the above-mentioned dimming effect. However, since λ and hence the angular resolution varies by only a couple of percent across our frequency band, the point dimming effect will be very small for most point sources. Moreover, it will be a smooth function of frequency which can be accurately matched by our fitting polynomial, thus making angular resolution adjustments rather redundant.

Flux Cuts

Since it remains unclear down to what flux cut $S_{\rm cut}$ one will be able to resolve and remove point sources with upcoming 21 cm tomography experiments, we examine how varying $S_{\rm cut}$ affects our algorithm for cleaning out unresolved point sources. In figure 4-14, we vary S_{cut} from 0.1 mJy to 100.0 mJy. The results are shown in Figure 4-14,

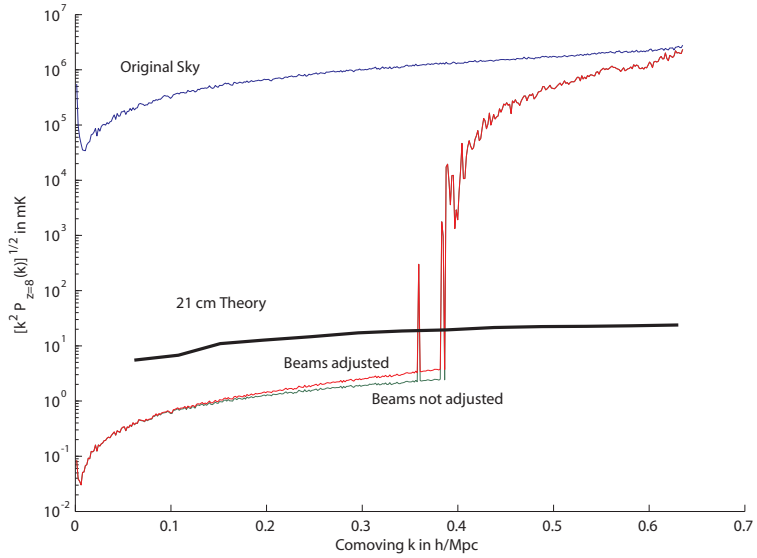

Figure 4-13: Effect of synthesized beam adjustment: the attempt to smooth all maps to a common resolution before cleaning is seen to do more harm than good.

and reveal a simple and useful scaling: raising the input power spectrum by some factor by increasing the flux cut raises the output by the same factor. This means that there is no need to rerun simulations with many flux cut levels, as the results from a single simulation can be analytically scaled to apply to all other cases. All we need to extract from simulations is the factor by which the cleaning algorithm suppresses the point source fluctuations — in this case, the suppression is seen to be about six orders of magnitude on large scales.

Quantitatively, the residuals are seen to lie below the theoretical 21-cm signal as long as the flux cut is below the 100.0 mJy level. However, there is some variability in results with frequency, both from variations in the residual power spectrum and in the magnitude of the theoretical curve. To be conservative, it is therefore prudent to aim to be able to remove bright point sources down to the 10 mJy level. This is consistent with the results in Bowman et al. (2009), where it was found that foreground contaminants could be subtracted to an acceptable level starting from a sky model with a flux cut of 10 mJy.

These results were derived using quadratic fits. In section 4.3.3, we will see that the order of the polynomial has a strong effect on residuals, so if this 10 mJy goal cannot be met, then the residual power spectrum can alternatively be brought down further by increasing the order of the fitting polynomial.

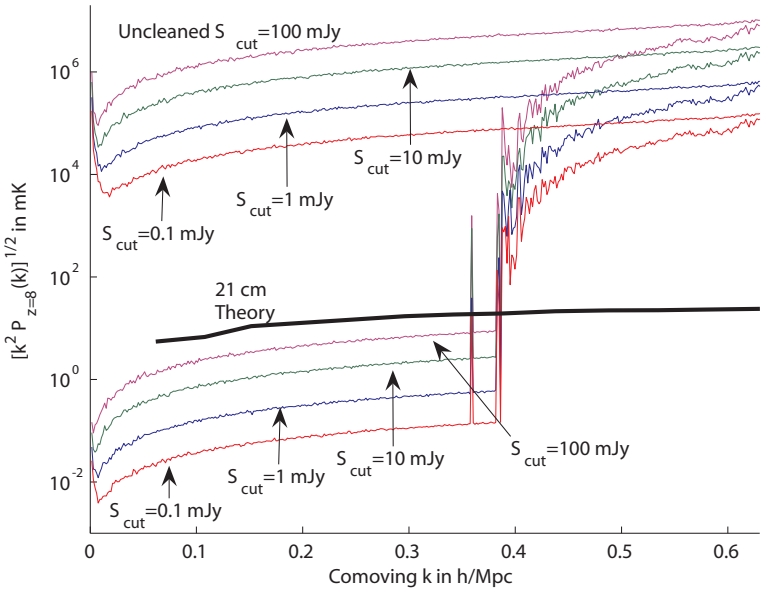

Figure 4-14: Effect of flux cut for bright source removal: 2D power spectra for the fiducial model but with various bright point source flux cuts.

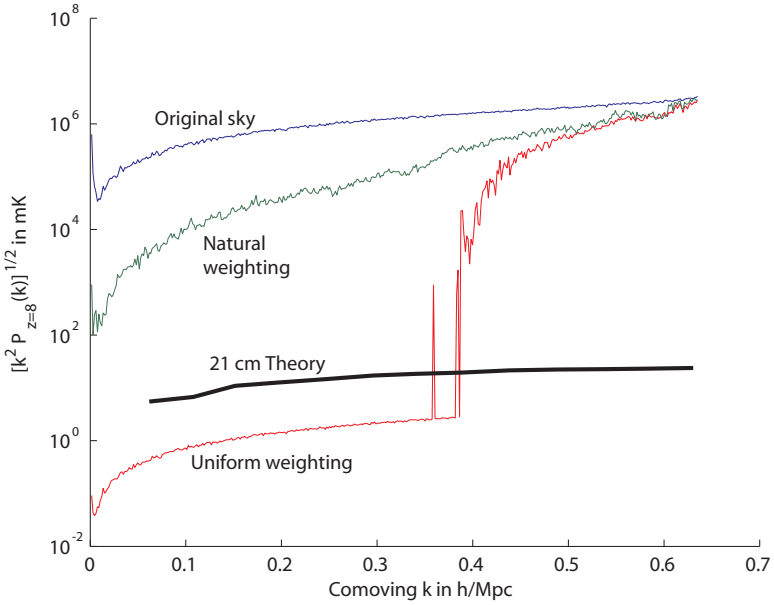

Figure 4-15: Effect of u-v plane weighting: 2D power spectra for the fiducial model but with two different weighting schemes for the u-v plane. A uniform weighting of the u-v plane is seen to far outperforms natural weighting.

Weighting scheme on u-v plane

In general, a radio array will not provide uniform coverage of the u-v plane – typically, some parts of the plane will be sampled more than once, while other parts will not be sampled at all. To compensate for this, one may decide to weight different parts of the u-v plane differently. In figure 4-15, we examine a "uniform" weighting scheme (where all sampled parts of the u-v plane are given equal weight) as well as a "natural" weighting scheme (where the u-v measurements are not given any weighting beyond their "natural" density as determined by the instrument). It is clear that a uniform weighting far outperforms a natural one. An intuitive explanation for why the uniform weighting so outperforms the natural weighting can be found in Bowman et al. (2009). Because the filling of the u-v plane is accomplished slowly by a discrete set of baseline loci, the final distribution of baselines will not be particularly smooth. The frequency dependence of the synthesized beam then maps this u-v lumpiness into an incoherence in the frequency direction that is difficult to fit out using smooth polynomials. With a uniform weighting, one artificially normalizes the baseline distribution so that each pixel in the sampled part of the u-v plane has the same weight, thus ensuring that the distribution of baselines is smooth.

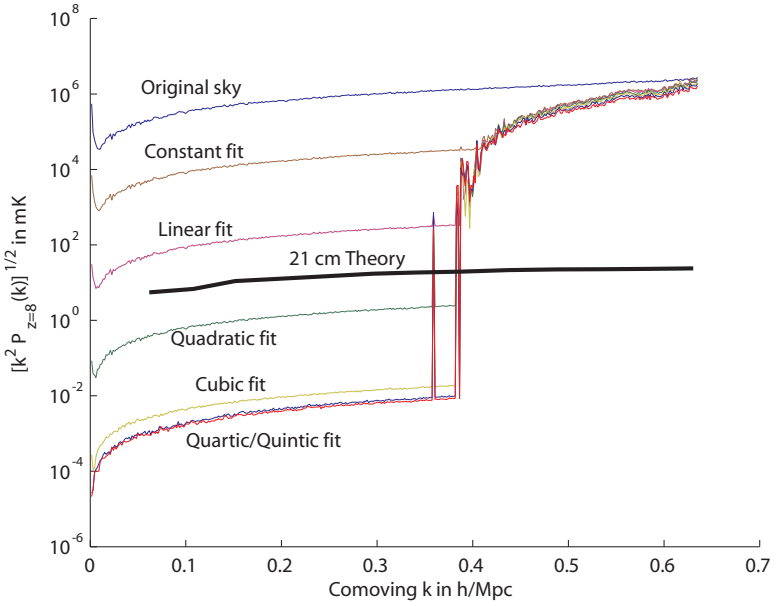

Figure 4-16: Effect of fitting range: 2D power spectra for the fiducial model but with the foreground subtraction performed by fitting polynomials of various degrees to the spectra.

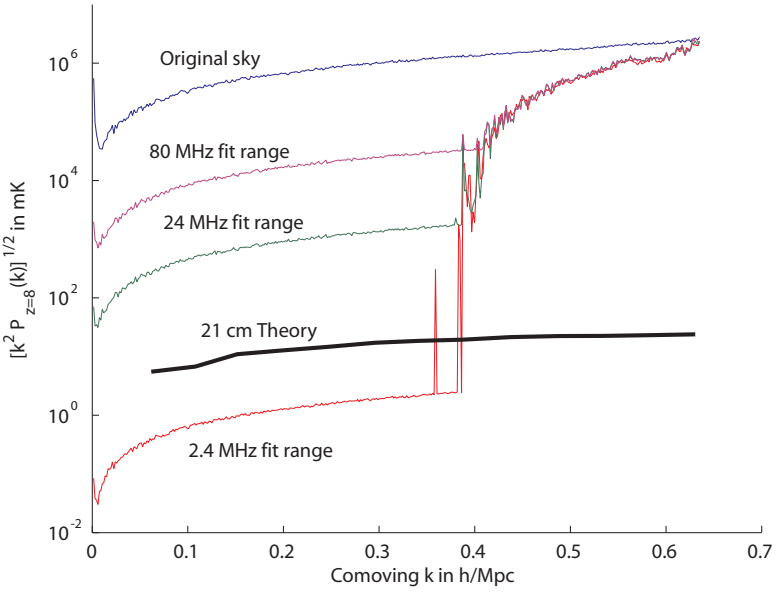

Figure 4-17: Effect of fitting range: 2D power spectra for the fiducial model but with the foreground subtraction performed by fitting quadratic polynomials to the spectra over a variety of frequency ranges.

Order of the polynomial fit

As we mentioned above, one should aim to fit the spectra with polynomials that are as low-order as possible, with the constraint that the residuals need to be below the expected cosmic signal level. Figure 4-16 shows that a quadratic fit satisfies this requirement for our fiducial case. If S_{cut} cannot be pushed down to 10 mJy, however, then a higher order fit may be necessary. For example, $S_{cut} = 100$ mJy would require a cubic fit.

Range of the polynomial fit

A choice must be made regarding the frequency range over which polynomial fits are applied to spectra of dirty map pixels. In general, one expects post-cleaning foreground residuals to increase with increasing frequency range, and this is what is seen in figure 4-17, where the foreground cleaning was performed using quadratic fits over various frequency ranges. If quadratic polynomials are used, one should therefore not fit over the broad frequency range probed by a typical 21-cm tomography experiment. Instead, one should divide the spectra into smaller frequency bands and subtract foregrounds individually from each band.

The effect of changing the frequency range is of course rather degenerate with the effect of changing the degree of the fitting polynomial, since what matters is ultimately the number of degrees of freedom that are fit out per unit frequency. As we increase either the polynomial order or the number of frequency bands, we are removing cosmological signal on ever smaller scales along the line of sight. One should therefore fit over as broad a frequency range as possible subject to the constraint that the predicted residuals lie comfortably below the expected cosmic signal. From figure 4-17, we see that a frequency range of a few MHz appears appropriate for quadratic fits.

4.4 Summary and Discussion

We have studied the problem of cleaning out point source foreground contamination from 21 cm tomography data, and investigated how the level of residual contamination after the cleaning process depends on various experimental and algorithmic parameters. Our results show that while successful foreground removal is far from guaranteed, the signs are encouraging. For instance, the fact that the post-cleaning residual foregrounds lie below the simulated cosmological signal in the fiducial model of Section 4.3.2 is promising, since the scenario in question is considered to be representative of current-generation experiments, and its cleaning algorithms are both simple and computationally cheap. These conclusions agree with those of the independent and concurrent analysis of Bowman et al. (2009), which is complementary by focusing on the specifics of the MWA experiment. As a further consistency check on our calculations and software, we and the authors of that paper agreed on a test example that we both simulated, obtaining consistent results.

We also identified a number of aspects of the problem that can be understood analytically. Noise and cosmic signal decouple from the point source problem as long as the cleaning method is linear. The power spectrum of unresolved point sources in the observed sky simply gets suppressed by a fixed k-dependent factor, so any shifts in this input power spectrum due to revised source count estimates or altered flux cuts for resolved source removal simply scale the output power spectrum by the same amount. Finally, simple Fourier space considerations explain why line-of-sight cleaning works as well as it does.

Based on our analysis, we can make several specific recommendations regarding the foreground subtraction of unresolved point sources, pertaining to both instrumentation and data reduction.

- Instrumental recommendations:
  - **Tile arrangement** should ensure good u-v after rotation synthesis; we found both a monolithic arrangement and an r^{-2} arrangement to work well with respect to foreground subtraction.
  - **Rotation synthesis** is important, but once an array has achieved 3 to 4 hours of integration, the advantage gained from further u-v coverage is minimal.
  - **Time between snapshots** taken of the sky should be made as short as possible. Ideally, one should have continuous u-v coverage, as is planned for experiments such as the MWA.
- Data reduction/algorithmic recommendations:
  - **Uniform weighting** of the u-v plane outperforms "natural weighting".
  - **Adjusting for changes in the primary beam with frequency** through an extra convolution kernel in u-v space results in a modest improvement in foreground subtraction. However, this conclusion may change if the primary beam has sidelobes.
  - **Adjusting for changes in the synthesized beam with frequency** appears not to be worthwhile.
  - **Quadratic fits** across a few MHz are adequate for subtracting off spectrally smooth foreground components from the data as long as bright point sources can be resolved and removed down to the 10 mJy level — otherwise higher orders should be considered.
  - **The frequency range** over which the polynomial fitting is performed should be narrow enough to allow one to see cosmic signal, but broad enough to prevent over-fitting which excessively removes cosmic signal. For quadratic fits, a range of a few MHz appears appropriate.
  - **The ability to remove bright point sources** prior to the subtraction of unresolved point sources is crucial, with the corresponding flux cut S_{cut} being the most important parameter of all in our analysis. If one can safely fit a cubic to the spectra without loosing too much cosmic signal, then one needs to be able to remove resolved points sources of flux down to 100 mJy.

Our results were deliberately conservative, obtained with an extremely simple line-of-sight cleaning procedure, and it is likely that one can do better. An interesting

question for future work is determining the optimal procedure. For example, other classes of fitting functions may be worth considering, and it may be better to replace multiple polynomial fits in disjoint bands by a single fit to a function with more parameters, say a spline. Such work should quantify the extent to which one is removing power from the cosmological signal, and optimize the tradeoff between more foreground removal and less signal removal. Such an optimization should ideally be done using a more complete sky model, including diffuse Galactic synchrotron radiation.

For now, however, what is reassuring for the field of 21-cm tomography is the fact that all of the recommendations listed above can be followed without any substantial revisions to current experimental designs. Our results on point source cleaning suggest that the foreground problem is surmountable, allowing 21 cm cosmology to live up to its great potential.

Chapter 5

An Improved Method for 21 cm Foreground Removal

5.1 Introduction

The previous chapter and other studies have examined the feasibility of foreground subtraction in neutral hydrogen tomography, and have generally found that variations of the line-of-sight approach pioneered by Wang et al. (2006), McQuinn et al. (2006b), and Zaldarriaga et al. (2004) should be able to clean out foreground contamination to an acceptable degree. Wang et al. (2006); Bowman et al. (2009); Jelic et al. (2008); Liu et al. (2009b); Gleser et al. (2008) performed simulations that took into account various instrumental effects and found that the foreground cleaning could be accomplished despite the many limitations of first generation instruments.

In this chapter we propose a new variation on the traditional line-of-sight methods. In particular, we describe a cleaning algorithm that (unlike most proposals) is implemented in Fourier space. As we discuss in section 5.3, this allows one to completely sidestep any problems that may arise from the frequency dependence of an instrument's beam, which was previously the limiting factor in the quality of foreground cleaning at high-wavenumber spatial Fourier modes (Bowman et al., 2009; Liu et al., 2009b). The increase in performance at such wavenumbers can be easily seen in figure 5-1, where we have plotted a 2D power spectrum of the data at $z = 8$.

The rest of the chapter is organized as follows. In section 5.2 we review the old method used in Bowman et al. (2009); Liu et al. (2009b), and in section 5.2.1 we recast it as an algorithm in Fourier-space. The Fourier-space description is then used to introduce our new method in section 5.3. We conclude in section 5.4.

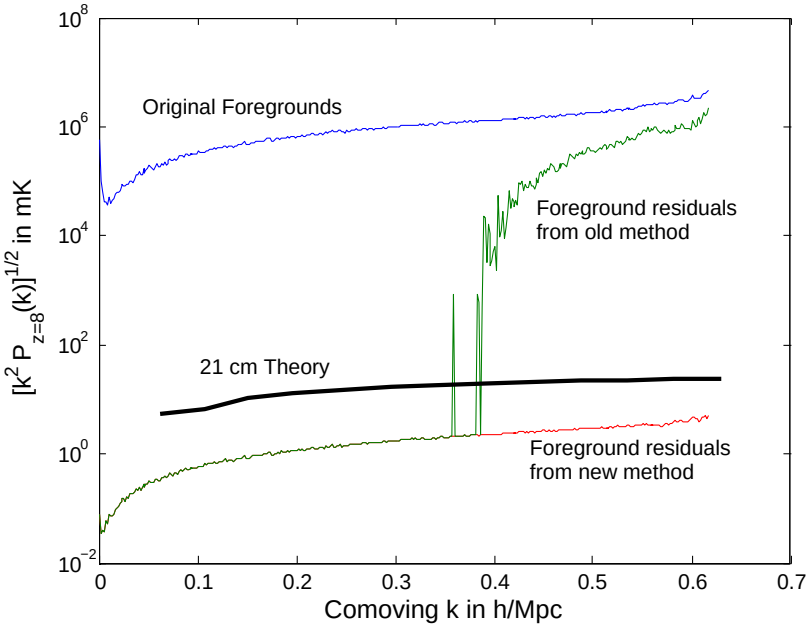

Figure 5-1: 2D power spectra of foregrounds and foreground residuals using the "old method" (Bowman et al., 2009; Liu et al., 2009b) and the "new method" (this chapter). At low-k the two methods give identical results, while at high-k the new method does much better.

5.2 Review of Old Method

In general, the data collected from a typical 21-cm tomography experiment can be thought of as populating a "datacube": stacks of 2D images separated by redshift or frequency along the line-of-sight direction due to the resonant nature of the 21-cm hyperfine transition. Along the transverse directions, the axes are usually labeled in one of three ways:

1. Real-space coordinates θ_x and θ_y. In this case the datacube is a literal map of 21-cm emission and foreground contaminants.

2. Interferometer coordinates u and v. Under the correct convention, these are simply the Fourier-conjugate coordinates to θ_x and θ_y. The datacube is a stack of 2D maps in Fourier space.

3. Fourier-space coordinates k_x and k_y. These are the Fourier-conjugate coordinates to the physical lengths x and y. Up to factors of 2π (depending on one's Fourier convention), $(k_x, k_y) \sim (u, v)/D_M$, where D_M is the transverse comoving distance.

In a typical experiment the data (in the form of visibilities) are collected in uv-coordinates, while the results are presented in either real-space coordinates (in the case of sky maps) or in Fourier-space coordinates (in the case of power spectra). Foreground removal is often done in either real-space coordinates (as demonstrated in Wang et al. (2006); Bowman et al. (2009); Jelic et al. (2008); Liu et al. (2009b); Gleser et al. (2008)) or in uv-space (as we propose in this chapter).

We first review the real-space removal algorithms. The fundamental idea behind all such algorithms is the fact that the 21-cm signal is expected to oscillate rapidly with frequency while the relevant foreground contaminants are spectrally smooth. The contaminants along a given line-of-sight can therefore be separated from the signal by plotting the flux as a frequency and subtracting off a smooth component (such as a low-order polynomial) from the total signal. What remains is the cosmological signal and a (hopefully small) residual contamination.

Previous studies have simulated the aforementioned real-space algorithms and have estimated the level of residual contamination that can be expected for current experiments (**?**Bowman et al., 2009) as well as how the residuals depend on the properties of a generic interferometer (Liu et al., 2009b). Although these papers have highlighted the fact that the quality of foreground subtraction is highly dependent on a large number of parameters (both instrumental and those pertaining to data

analysis), they also suggest that the qualitative behavior is rather generic. In what follows we examine the qualitative behavior that emerges, emphasizing the various features and their mathematical origin.

Consider the spectrum shown in the top panel of figure 5-2, where we show the frequency dependence of a single pixel in real-space coordinates (i.e. the frequency dependence of a particular line-of-sight), as seen by a typical 21 cm tomography interferometer[1]. The spectrum contains *foregrounds only*, with no noise or cosmological signal. Since the foregrounds are known (and are simulated[2]) to be spectrally smooth, this suggests that the rapid oscillations seen in the figure are *caused by the instrument*. This is bad news for the subtraction algorithm, as it means that simply fitting out the smooth component of a spectrum will leave residuals that can be confused with the cosmological signal. Indeed, it can be seen from the figure that the fit seems rather poor.

One way of understanding the rapid oscillations is to consider the interferometer's beam in real-space. The left-hand panel of figure 5-3, shows that the beam of a typical interferometer contains "frizz" outside the central peak that oscillates rapidly with angle. Since beam widths scale as λ/D, this angular oscillation translates into an oscillation in frequency, which is what is seen in figure 5-2. Alternatively, the behavior of figure 5-2 can be understood by considering the effect of an interferometer's beam in uv-space. An interferometer samples pixels in the uv-plane, and with enough of these uv-pixels one can produce a real space image by Fourier transforming. Thus, the spectrum of a single pixel in real space like that shown in figure 5-2 can be thought of as a linear combination of the spectra of different uv pixels sampled by the interferometer. Exactly which pixels are sampled depends on the layout of the interferometer in question, but in a typical 21-cm tomography experiment the uv coverage is complete near the origin and drops off as one moves farther out.

In general, the spectra such as that shown in figure 5-2 can be decomposed into the two components: a component that is formed from the linear combination of uv-pixels where the interferometer's coverage is complete (i.e. the inner parts of the plane), and a component that is formed from uv-pixels residing in parts of the uv-plane where coverage is sparse (i.e. the outer regions). These components are shown by the solid black lines in the bottom panel of figure 5-2. The line with "x" markers

[1]We use the Murchison Widefield Array as our fiducial model for the simulations in this chapter, but it should be noted that the algorithm we propose in section 5.3 can be applied to data collected by any interferometric configurations.

[2]The simulation methodology used in this chapter was the same as that used in Liu et al. (2009b). Please see Liu et al. (2009b) for details.

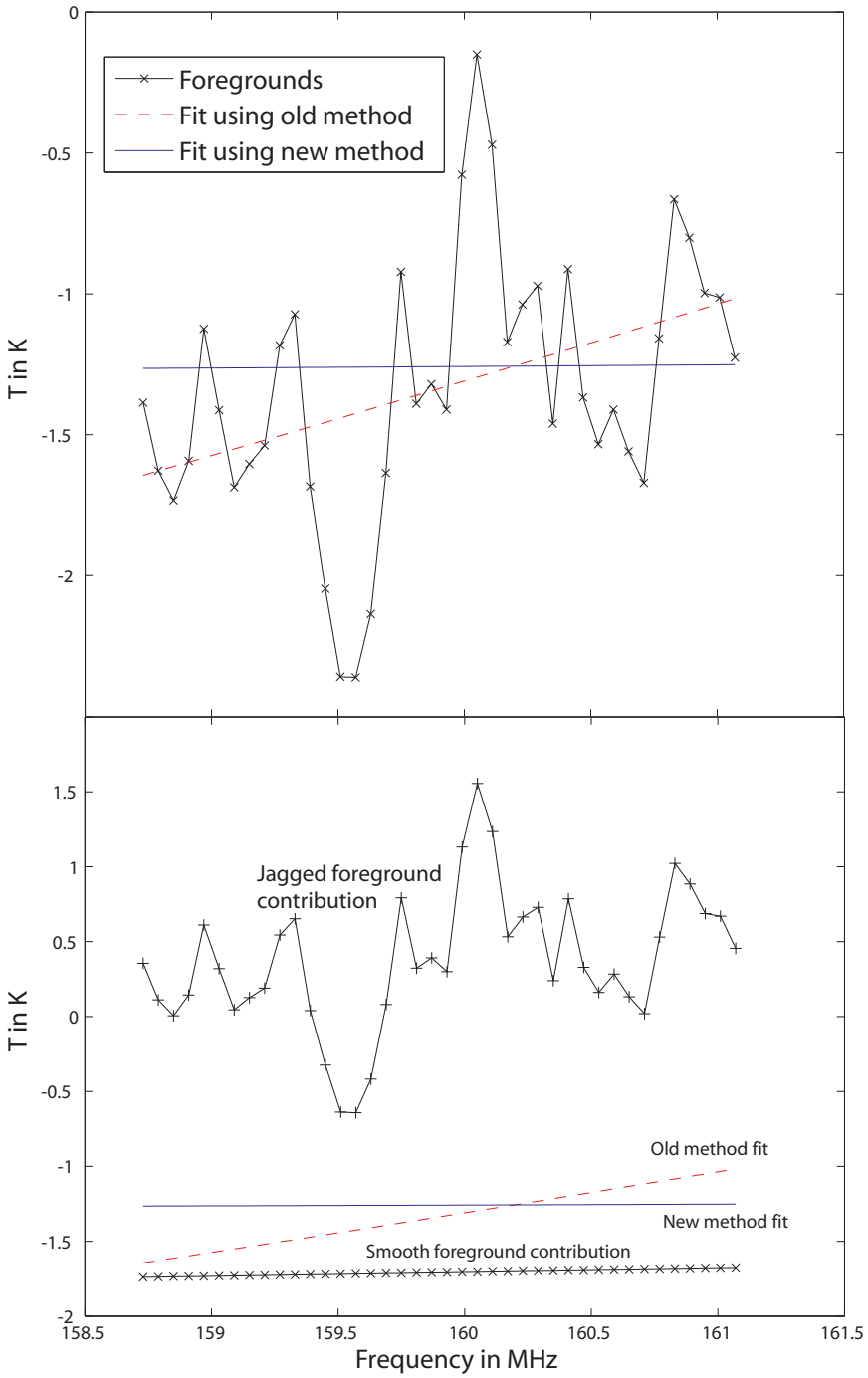

Figure 5-2: The spectrum of a typical real-space pixel. Top panel: Instrumental effects result in a jerky dependence on frequency, even though the foregrounds are intrinsically smooth. The red/dotted line gives the foreground fit using the old method, while the blue/solid line gives the analogous real space "fit" using the new method (see section 5.3 for details). Bottom panel: the signal seen by the instrument is decomposed into a smooth component coming from the central parts of the uv-plane and a jagged component from the outer parts of the plane.

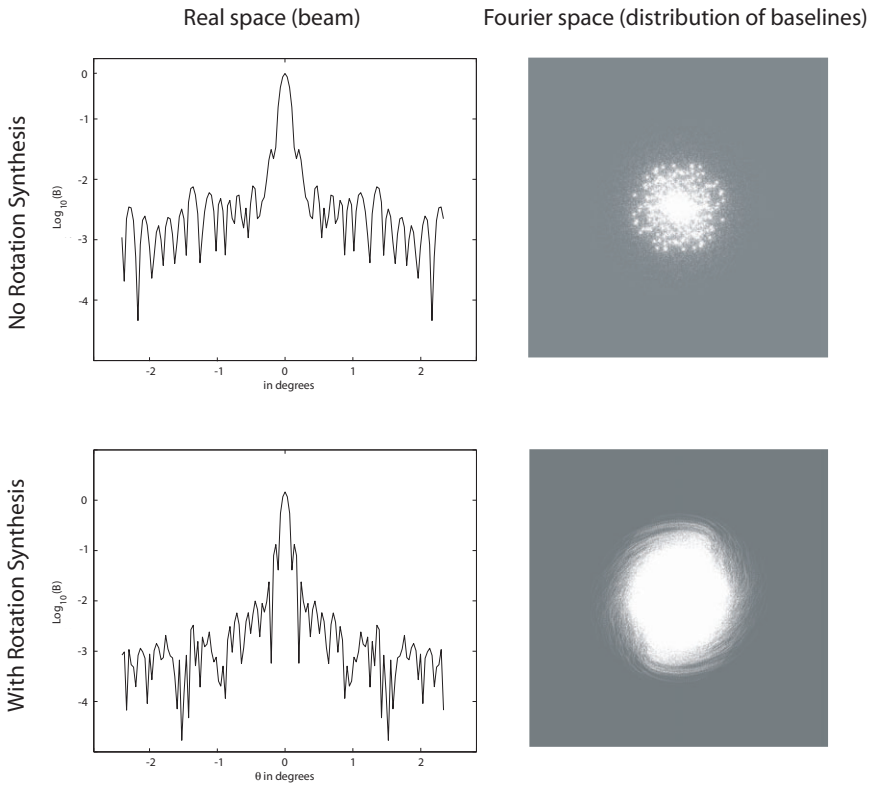

Figure 5-3: The left hand column shows sample beam profiles (a real-space description of the beam) while the right hand column shows the corresponding uv-distribution of baselines (a Fourier-space description of the beam). The top row illustrates an array with no rotation synthesis, while the bottom row shows an array with 6 hours of rotation synthesis. The real-space beams are normalized so that their peaks are at 1.

(showing the part of the signal originating from the inner parts of the plane) is seen to be smooth, whereas the line with "+" markers (showing the contribution from the outer parts) is what contributes the rapid oscillations. This decomposition explains why real-space pixel-by-pixel foreground subtraction algorithms have been shown to be adequate even though the fits themselves seem terrible at first sight. Even though the smooth fits cannot subtract off the jerky component of the spectrum, they are capable of fitting out the smooth component that comes from the central parts of the Fourier plane. Indeed, this is exactly what is seen in figure 5-1, where the low-k parts of the power spectrum are cleaned effectively whereas the high-k parts remain contaminated. It is simply the case that by examining pixels in real space, one is viewing a "bad" linear combination of pixels that mixes together the well-fit, central located uv-pixels with the outer uv-pixels where sparse baseline coverage results in jerky spectra that are badly fit.

5.2.1 Fourier space description of decontamination

In the previous section we examined how foreground fits of real-space pixels could be understood by considering the flux in each pixel as being a linear combination of different uv-pixels. We now show that one can go further and perform the fits themselves in uv-space and get exactly the same results. With slight modifications, this will lead to section 5.3's discussion of an improved method for subtracting foregrounds at high k.

Consider the steps that must be taken to perform the foreground subtraction outlined above. The data is collected by the interferometer in Fourier space i.e. in a u-v-ν datacube. This data must then be Fourier transformed in the two transverse directions, giving an x-y-ν datacube. Fitting is subsequently performed in the frequency direction. Mathematically, we can express this as follows. Let $\tilde{y}_{ij\alpha}$ represent the initial datacube, with the first two (Latin) indices being the two spatial indices and the last (Greek) index being the frequency index. With no loss of generality, we can fold the first two indices into one and write $\tilde{y}_{j\alpha}$ instead. In this notation the Fourier transform can be written as

$$y_{k\alpha} = \sum_i F_{ki}\tilde{y}_{i\alpha} \tag{5.1}$$

where F is the Fourier matrix and y is the real-space analog of $\tilde{y}$. The fit in the frequency direction can be represented by yet another linear operator[3] G, and so we have

$$\overline{y}_{k\beta} = \sum_\alpha G_{\beta\alpha}y_{k\alpha} = \sum_{i,\alpha} G_{\beta\alpha}F_{ki}\tilde{y}_{i\alpha} \tag{5.2}$$

where $\overline{y}$ represents the fit. In the last expression, note that G possesses only Greek indices whereas F only has Latin indices. This means that the two operations performed in our algorithm – the 2D spatial Fourier transform (F) and the fitting in the frequency direction (G) – in fact commute (i.e. $FG = GF$).

The fact that the Fourier transform commutes with the fitting means that we can perform the two operations in either order. In other words, we can think of the foreground fitting and subtraction as taking place in Fourier space without changing any of the results (which is something that we have also verified numerically). Viewing the process as a pixel-by-pixel fitting in uv-space reveals exactly why there exists

[3]Explicitly, if X is an $n \times m$ matrix such that X_{ij} equals the frequency of the ith frequency channel taken to the jth power, then G is given by $X[X^tN^{-1}X]^{-1}X^tN^{-1}$, where N is the noise covariance matrix (Wang et al., 2006).

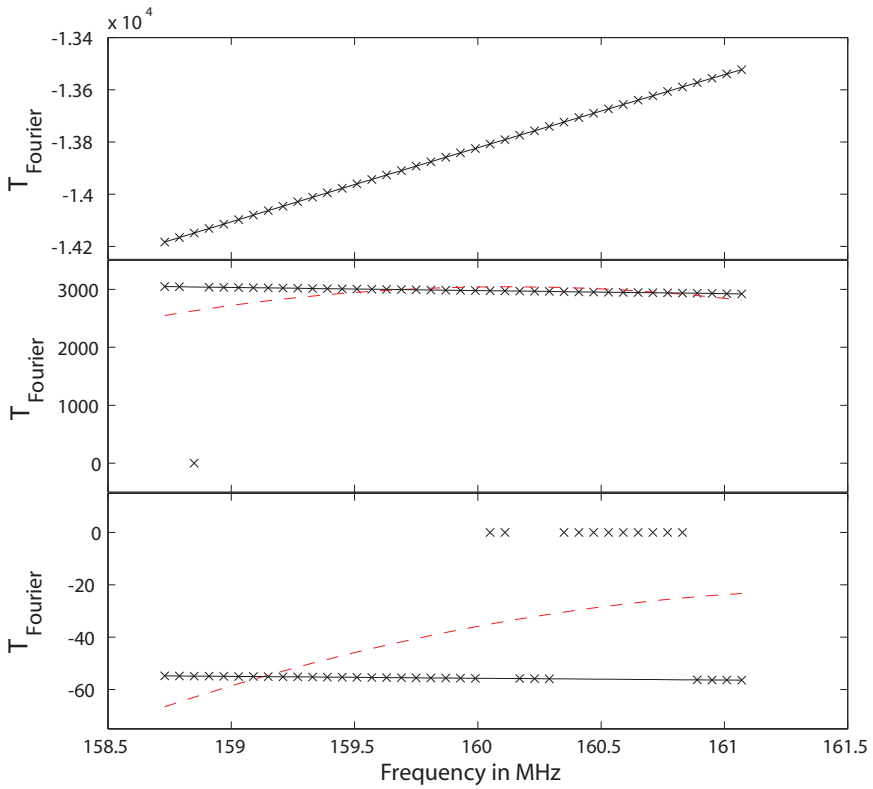

Figure 5-4: Spectra of various uv pixels from different parts of the plane. From the top panel to the bottom panel, one is moving away from the origin. It is clear that the data can be easily fit by low-order polynomials in the top panel, but that the old method of fitting (dashed red curves) becomes inadequate when baseline coverage begins to drop out. The solid black curves show the fits done using the new method describe in section 5.3.

such a vast difference between the quality of the cleaning at low-k and at high-k, and why the transition between the two regimes appears as such an abrupt jump in the power spectra. In figure 5-4 we show typical spectra from different parts of the uv-plane. The top panel shows a typical pixel from the inner part of the plane. The spectrum is plotted using so-called uniform weighting, so that in every Fourier pixel the interferometer acts as an on/off switch: the interferometer "reads" a value of 0 if no baselines fall in that pixel, and 1 otherwise (regardless of how many baselines are binned into the pixel). It is evident that a simple polynomial fit does extremely well.

On the other hand, when one moves out to regions of the uv-plane where baseline coverage becomes sparse, the fit becomes poor. A glance at the bottom two panels of figure 5-4 makes the problem clear – when coverage is sparse, at certain frequencies there is no baseline coverage, and a simple polynomial fit is unable to deal with this. We emphasize that the trouble is not with incomplete Fourier coverage per se. It

is the fact that the incomplete coverage is changing with frequency. In other words, foreground subtraction becomes poor in this regime because the frequency dependence of the beam (or "mode-mixing", as emphasized in Bowman et al. (2009); Liu et al. (2009b)) becomes important on these small (high-k) scales. Note that even though this problem exists when the spectra are being fit in real space, it is not *apparent* unless one fits in uv-space, where the pixels are "good" linear combinations of the data.

5.3 New method

We now propose a slight modification to the foreground subtraction algorithm that evades the aforementioned problem. From figure 5-4, one can see that an alternate way of phrasing the problem is to say that the old fitting algorithm, being mathematically equivalent to a fitting in real space, is unable to distinguish between pixels with no data and pixels with values that happen to be zero. In uv-space, however, one can easily identify pixels with no baselines, and so one can simply skip frequencies where data is unavailable. In fact, one can find the optimal fit (in the sense of having minimal r.m.s. errors) by employing an inverse-variance weighted fit. In this scheme, the weight of each point in the least-squares sum is proportional to N, the number of baselines that are binned into a particular uv-pixel at a particular frequency. This way, points with lower signal-to-noise are given less weight[4].

In figure 5-4 it can be seen that since missing frequencies are now given zero weight in the fit, one obtains excellent fits even for uv-pixels where baseline coverage is sparse. This improves the subtraction of foregrounds at frequencies where there *is* data, whereas at the skipped frequencies nothing has been compromised since no foregrounds were detected by the instrument in the first place.

The effect that the frequency-skipping has on the 2D power spectrum is shown in figure 5-1. To be conservative, we have also tested our new algorithm using a completely independent pipeline with a different foreground model (for details, please

[4]It is important to emphasize that while we are performing a weighted fit to the data, the data itself is still uniformly weighted (as described in section 5.2.1). The key distinction here is the difference between fitting to weighted data and making a weighted fit. In the former, one multiplies the uv-space data by a weighting function and fits to the modified data, while in the latter one does not apply any weighting function, but allows for the fit to treat different points with varying importance. As an example of the difference, consider a linear fit to a series of points that happen to lie exactly on a straight line. If one fits to a weighted version of this data, the result will not be a perfect fit unless the weighting function is a constant. On the other hand, applying a weighted fit will give a perfect fit regardless of the weighting used.

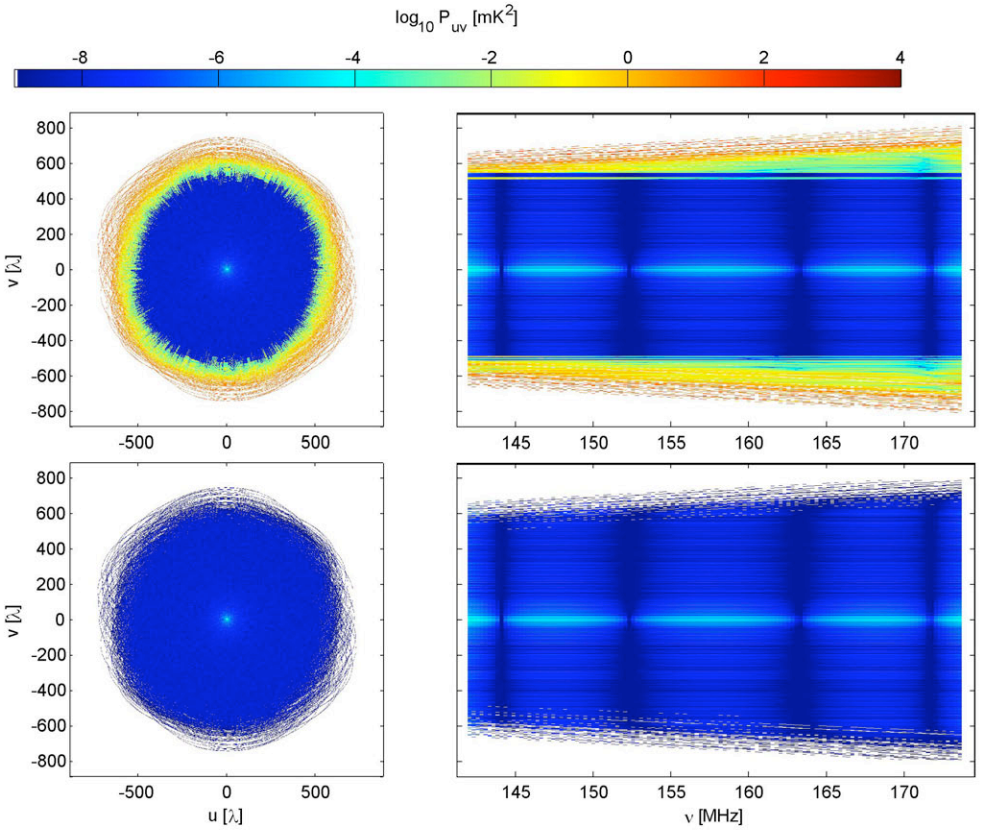

Figure 5-5: Post-subtraction residuals shown in the uv-plane (left column) and as a function of frequency (right column) for both the old method (top row) as well as the new method (bottom row). The new method does offer any increase in performance at low-k, but avoids the large increase in residuals at high-k.

see Bowman et al. (2009)). The results from the second pipeline are shown in figure 5-5, and the fact that the results agree demonstrate the fact that uv-plane cleaning is generally applicable and not dependent on the foreground model. Qualitatively, one can see that at low-k there is no improvement from the old method because in that regime one is limited by the fact that simple low-order polynomials will not in general be perfect fits to the foregrounds, even though the foregrounds are smooth functions. At high-k, however, one avoids the dramatic increase in post-subtraction foreground residuals, because previously the limitation at high-k was the mode-mixing problem. With our new method, the limiting factor is the ability of the fitting function to match the form of the foregrounds. For example, the fact that the foreground residuals in figure 5-1 are a constant factor ($\sim 10^6$) off from the original foregrounds regardless of scale (or equivalently, regardless of location on the uv-plane) means that the residuals

are due entirely to the quality of the fit. In other words, the residuals of one part in $\sim 10^6$ come from the fact that the second-order polynomials used in the fits to produce figure 5-1 are good fits to the foregrounds only to one part in $\sim 10^6$. With the chief limitation now being the fitting itself, one can in principle subtract foregrounds up to Fourier modes that correspond to the longest baselines, although as one is forced to skip many frequencies at high k, the signal-to-noise of the data is reduced.

As a weighting scheme that weights data points according to their information content, inverse-variance weighting not only gives higher signal-to-noise data points greater weight, but also automatically incorporates frequency-skipping, since the skipped frequencies are simply those with $N = 0$ and therefore no information. While both effects contribute to an improvement in the quality of foreground subtraction, we find that the frequency-skipping is the dominant contribution to the improvements in foreground subtraction. At noise levels that are typical for current-generation 21 cm tomography experiments, we find that if only the frequency-skipping is done (i.e. if all frequencies that are *not* skipped are given the same weight), the resulting residual power spectra are only about 1% greater than if both techniques are used.

It is important to note that whereas without the skipping of empty frequencies the transverse Fourier transform commuted with the fitting of the foregrounds, under the new scheme proposed here the two operations no longer commute. This is because the frequencies of the pixels that need to be skipped require knowing the baseline distribution (which lives in uv-space) and therefore depend on the location of the uv-pixel being cleaned. Mathematically, this means that in equation 5.2 the fitting operator G acquires an extra i (spatial) index and the two sums no longer commute. The significance of this is that the fit can no longer be done in real-space. To apply this new algorithm for foreground subtraction, one *must* work in Fourier space.

However, while the skipping of frequencies in our fit dictates that we must *work* in Fourier space, the improvements brought about by the new algorithm can still be *seen* in real space. Consider the solid (blue) fit in figure 5-2. This fit was obtained by taking the uv-space fits generated by the new algorithm and Fourier transforming real space to give a real space "fit". It is clear from the figure that the new method does a much better job of tracking the behavior of the smooth foreground component. (The "fit" is displaced vertically from the smooth component because the jagged component itself has a non-zero mean that must also be fit out). On the other hand, the slope of the fit from the old method is biased by the jagged foreground contribution (which, remember, is an instrumental artifact that arises from incomplete baseline coverage), and does a worse job tracking the smooth foregrounds.

The fact that our new method traces the smooth foreground component better means that it can be used to get better estimates of the foregrounds themselves. One simply Fourier transforms the fits produced by the new algorithm to get real-space, multi-frequency maps of the foregrounds. Such maps will be of a higher quality than those that are simply imaged by the instruments. This is because our new fitting algorithm can be interpreted as one where the missing frequencies are not so much skipped as interpolated over. By fitting low-order polynomials over the frequencies where data is available, one is essentially deriving a foreground model that can be extrapolated to other frequencies. Without missing frequencies in the spectra, the real space foreground maps will not have artificially jagged foreground components, and will therefore be a more accurate representation of the true foregrounds.

5.4 Conclusions

In this chapter, we have shown that there is an easy explanation for the increased foreground residuals at high-k: a frequency-dependent incompleteness of baseline coverage in the outer parts of the uv-plane makes the foregrounds in certain uv-pixels difficult to fit out using a simple unweighted polynomial fit. The solution to this problem is to weight the fit so that frequencies with no information are given zero weight, while other frequencies are given an inverse-variance weighting. As seen in figure 5-1, this allows foreground cleaning to be performed at much higher k, paving the way for higher quality power spectrum measurements in neutral hydrogen tomography.

Chapter 6

A Method for 21 cm Power Spectrum Estimation in the Presence of Foregrounds

6.1 Introduction

Having explored foreground subtraction on its own in previous chapters, we now add power spectrum estimation to the picture and consider the two data analysis steps in a single framework. Power spectrum estimation in 21 cm tomography has similarities to and differences from power spectrum estimation for both the CMB and galaxy surveys. As it is for the CMB, foreground contamination is a serious concern, especially since the foregrounds are so strong at the lower frequencies ($\sim 100 - 200\,\mathrm{MHz}$) that EoR experiments target. Unlike the CMB, however, 21 cm tomography probes a three-dimensional volume (just like with galaxy surveys), so the final quantity of interest is not an angular power spectrum but a three-dimensional power spectrum. Ultimately, one wishes to obtain the spherically averaged matter power spectrum as a function of redshift $P(k, z)$, because one expects the cosmological power to be isotropic. However, it is often preferable to first form a cylindrically averaged power spectrum $P(k_\perp, k_\parallel; z)$, since the isotropy may be destroyed by redshift-space distortions (Barkana & Loeb, 2005a), the Alcock-Pacyznski effect (Ali et al., 2005; Nusser, 2005; Barkana, 2006), as well as residual foregrounds that arise from imperfect foreground removal.

The problem of foreground removal in 21 cm tomography has been extensively studied in the literature. As noted in previous chapters, early foreground cleaning proposals focused on using the angular structure of the foregrounds (Di Matteo et al.,

2002, 2004; Oh & Mack, 2003; Santos et al., 2005; Zaldarriaga et al., 2004), while recent studies have suggested that a line-of-sight (LOS) spectral subtraction method may be the most promising approach (Wang et al., 2006; Gleser et al., 2008; Bowman et al., 2009; Liu et al., 2009b; Jelic et al., 2008; Harker et al., 2009; Liu et al., 2009a; Harker et al., 2010). Though the detailed algorithmic implementations of the LOS method vary amongst these studies, the core idea is always to take advantage of the smooth spectral behavior of foregrounds to separate them from any cosmological signals, which are expected to fluctuate rapidly with frequency. Simulations of this approach have confirmed it as a successful foreground cleaning technique in the sense that the post-subtraction foreground residuals can be suppressed to a level smaller than the expected amplitude of the cosmological signal. Thus, foreground contamination is unlikely to be an insurmountable obstacle to the initial *detection* of the 21 cm cosmological signal.

What remains unclear, however, is whether LOS subtraction techniques adversely affect the quality of one's final estimate of the cosmological power spectrum. For instance, although spectrally smooth components of the cosmological signal are expected to be small, if any are present they will be inadvertently subtracted off with the foregrounds when LOS cleaning is performed. LOS methods are thus not *lossless*, in the sense that their destruction of cosmological information leads to a degradation of error bars on the final power spectra (for a more rigorous definition of information loss, see Tegmark (1997c) or Section 6.2 of this chapter for a quick review). Foreground subtraction may also lead to residual systematic noise and foreground *biases* in estimates of the power spectrum, a fact that was explored numerically using simulations in Bowman et al. (2009); Harker et al. (2010); Petrovic & Oh (2011). Finally, foreground subtraction along the LOS may lead to correlated errors in power spectrum measurements, which can limit their usefulness for estimating cosmological parameters.

The goal of this chapter is to adapt the existing mathematical formalism for power spectrum estimation from the CMB and galaxy survey literature in a way that not only respects the unique geometrical properties of 21 cm tomography experiments, but also allows us to robustly deal with the issue of foreground contamination. This will permit a more complete analysis of the errors involved in estimating 21 cm power spectra, in a way that automatically incorporates the effect that foreground removal has on the final results. In particular, we will be able to quantify not just the errors on the power spectrum itself (the "vertical" error bars), but also the correlations between the parts of the power spectrum (the "horizontal" error bars, or equivalently

the power spectrum window functions[1]). In addition, we will quantify the noise and foreground biases that are introduced by foreground subtraction and power spectrum estimation so that such biases can be systematically removed. We accomplish all this by considering the foregrounds not as an additional signal to be removed, but rather as a form of correlated noise. Doing so in our formalism allows us to build on the numerical simulation work of Bowman et al. (2009), Harker et al. (2010), and Petrovic & Oh (2011) to *analytically* show how LOS foreground subtraction methods are lossy and leave residual noise and foreground biases. The formalism also naturally suggests an alternative method for foreground subtraction that leads to smaller error bars and eliminates the residual biases.

The rest of this chapter is organized as follows. In Section 6.2, we introduce the mathematical notation and formalism that we use for power spectrum estimation. Existing LOS methods are cast in this language, and in addition we introduce our alternative foreground subtraction scheme. In Section 6.4, we test our methods on a specific foreground model, which we describe in Section 6.3. Provided that the foregrounds satisfy certain generic criteria (such as having a smooth frequency dependence), the details of the foreground model should have no effect on our qualitative conclusions. In Section 6.5, we show how these qualitative conclusions can be understood through a simple toy model. We summarize our conclusions and discuss broader implications in Section 6.6.

6.2 The Mathematical Framework

In this section we review the power spectrum estimation formalism introduced by Tegmark (1997c), and adapt it for 21 cm tomography. Readers interested in the mathematical details of power spectrum estimation are encouraged to peruse Bond et al. (1998); Tegmark (1997c); Tegmark et al. (1998, 2004). Our development builds on the quadratic methods presented in Tegmark (1997c); Tegmark et al. (1998). The goal is to derive new expressions appropriate to 21 cm power spectrum estimation (such as Equation 6.3 and the expressions in Section 6.2.3) as well as to systematically place different power spectrum estimation methods in a common matrix-based

[1]These are not to be confused with the instrumental window functions that are conventionally defined in radio astronomy to quantify the effect of an instrument's beam and noise properties (and as used, for instance, in related 21 cm power spectrum papers such as Morales & Hewitt (2004) and Bowman et al. (2006)). In this chapter, the "window function" can be thought of as a convolution kernel that produces the measured power spectrum when applied to the true power spectrum. Please see Section 6.2 for a rigorous mathematical definition.

framework.

6.2.1 Quadratic Estimators

Our objective is to use the measured 21 cm brightness temperature distribution $T_b(\mathbf{r})$ to estimate a cylindrically symmetric power spectrum[2] $P_T(k_\perp, k_\parallel)$ which is conventionally defined by the equation

$$\langle \widehat{T}_b(\mathbf{k})^* \widehat{T}_b(\mathbf{k}') \rangle = (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') P_T(k_\perp, k_\parallel), \tag{6.1}$$

where $\widehat{T}_b$ is the three-dimensional spatial Fourier transform of T_b, and δ is the Dirac delta function. As written, this definition contains the *continuous* functions T_b and P_T. However, any practical numerical scheme for estimating the power spectrum must necessarily be discrete. The simplest way to do this is to divide the measured spatial distribution of brightness temperature into discrete voxels, so that T_b takes the form of a data vector $\mathbf{x}$ that is essentially a list of the brightness temperatures measured at various points on a three-dimensional grid. Note that in forming this vector, one should *not* use the full radial extent of the data that are available in a typical 21 cm tomography experiment. Instead, one should split the full data into *multiple* $\mathbf{x}$ vectors, each of which spans only a very narrow range in redshift so that the evolution of the cosmological signal is negligible. The analysis that we describe in this chapter should then be performed *separately* on each vector to produce a number of power spectra at different redshifts. Failure to do so would violate the assumption of translational invariance that was implicit when we defined the power spectrum using Equation 6.1.

To discretize the power spectrum, we parametrize it as a piecewise *constant* function[3], such that $P_T(k_\perp, k_\parallel) = p_{ab}$ for $k_a^\perp \leq k^\perp < k_{a+1}^\perp$ and $k_b^\parallel \leq k^\parallel < k_{b+1}^\parallel$. If the index a runs over M different values and the index b runs over N values, the power spectrum can then be stored in an MN-dimensional vector p_α, where index pairs (a, b) have been folded into a single index α. Parameterizing the power spectrum

[2]Note that in this chapter, we are not examining the steps required to estimate the *matter* power spectrum. Instead, we are focussing on estimating the *temperature* power spectrum $P_T(k_\perp, k_\parallel)$, which is an important first step for finding the matter power spectrum. For details on how one might extract a matter power spectrum from a temperature power spectrum, see for example Barkana & Loeb (2005a); Mao et al. (2008).

[3]In a practical application of the methods described in this chapter, it is often preferable to instead parametrize as a piecewise constant function the *ratio* of the power spectrum to a prior. As shown in Hamilton & Tegmark (2000), doing so tends to give better behaved window functions (Equation 6.8). Here for simplicity we employ a white (*i.e.* constant) prior simply because there are as yet no observational constraints on the form of the 21 cm power spectrum.

in this way (where each component p_α is referred to as the *band power* of the band α) represents no significant loss of information as long as the widths of the k bins are small compared to the physical scales over which P_T varies appreciably (Tegmark et al., 1998).

A *quadratic method* for estimating our discretized power spectrum is one where the estimator of p_α takes the quadratic form

$$\widehat{p}_\alpha = (\mathbf{x} - \mathbf{m})^t \mathbf{E}^\alpha (\mathbf{x} - \mathbf{m}) - b_\alpha \tag{6.2}$$

for some family of symmetric matrices $\mathbf{E}^\alpha$ and constants b_α (one for each k-region in our piecewise constant discretized power spectrum). The vector $\mathbf{m}$ is defined as the *ensemble average* over different random realizations of the random variable data vector $\mathbf{x}$, *i.e.* $\mathbf{m} \equiv \langle \mathbf{x} \rangle$. The hat ($\widehat{\ }$) denotes the fact that what we have here is an *estimate* of the true power spectrum p_α from the data. The matrix $\mathbf{E}^\alpha$ encodes the Fourier transforms, binning, and — crucially — the weighting and foreground subtraction steps required in going from the calibrated data vector $\mathbf{x}$ to the corresponding estimate of the power spectrum p_α. For example, *if one chooses to forgo any form of foreground subtraction* and to form a power spectrum using a completely uniform weighting of all voxels, the matrix takes the form

$$\begin{aligned}
\mathbf{E}^\alpha_{ij}\Big|_{\text{no fg}} &= (\mathbf{C}_{,\alpha})_{ij} \equiv \int_{V_k^\alpha} e^{i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j)} \frac{d^3\mathbf{k}}{(2\pi)^3} \\
&= \frac{1}{(2\pi)^3} \int_0^{2\pi} \int_{k_{a-1}^\perp}^{k_a^\perp} \int_{k_{b-1}^\parallel}^{k_b^\parallel} e^{i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j)} k^\perp dk^\parallel dk^\perp d\varphi_k \\
&= \frac{(2\pi^2)^{-1}}{(r_{ij}^\perp)^2 r_{ij}^\parallel} \left(\sin k_b^\parallel r_{ij}^\parallel - \sin k_{b-1}^\parallel r_{ij}^\parallel \right) \times \\
&\quad \left[k_a^\perp r_{ij}^\perp J_1(k_a^\perp r_{ij}^\perp) - k_{a-1}^\perp r_{ij}^\perp J_1(k_{a-1}^\perp r_{ij}^\perp) \right],
\end{aligned} \tag{6.3}$$

where $r_{ij}^\parallel \equiv r_i^\parallel - r_j^\parallel$ is the radial line-of-sight distance between spatial grid points $\mathbf{r}_i$ and $\mathbf{r}_j$, $r_{ij}^\perp \equiv |\mathbf{r}^\perp| = |\mathbf{r}_i^\perp - \mathbf{r}_j^\perp|$ is the projected perpendicular distance, $\varphi_k \equiv \arccos(\widehat{\mathbf{k}}_\perp \cdot \widehat{\mathbf{r}}_{ij}^\perp)$ is the angle between $\mathbf{k}_\perp$ and $\mathbf{r}_{ij}^\perp$, V_k^α is the volume in Fourier space of the α-th bin, and J_1 is the 1^{st} Bessel function of the first kind. (Recall that indices a and b specify the $k_\perp$ and $k_\parallel$ bins, respectively, and are folded into the single index α). The notation $\mathbf{C}_{,\alpha}$ is intended to be suggestive of the connection between the correlation function

(or covariance matrix) $\mathbf{C}$ of the data and its discretized power spectrum p_α:

$$\mathbf{C} \equiv \langle (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t \rangle = \mathbf{C}_{fg} + \mathbf{N} + \sum_\alpha p_\alpha \mathbf{C}_{,\alpha}, \tag{6.4}$$

where in the last equality we were able to take advantage of the fact that the foregrounds, instrumental noise, and signal are uncorrelated to write the total covariance as the sum of individual covariance contributions ($\mathbf{C}_{fg}$ for the foregrounds and $\mathbf{N}$ for the instrumental noise) and the contribution from the cosmological signal (the final term) with no cross-terms. In this form, we see that $\mathbf{C}_{,\alpha} \equiv \partial \mathbf{C}/\partial p_\alpha$ is simply the derivative of the covariance with respect to the band power. Intuitively, the last term in Equation 6.4 is simply an expansion of the correlation function of the cosmological function in binned Fourier modes. As a more familiar example, consider a situation where one is trying to estimate the three-dimensional power spectrum $P_T(\mathbf{k})$ instead of $P_T(k_\perp, k_\parallel)$. In such a case, there would be no spherical or cylindrical binning in forming the power spectrum, and we would have $(\mathbf{C}_{,\alpha})_{ij} \sim e^{i\mathbf{k}_\alpha \cdot (\mathbf{r}_j - \mathbf{r}_i)}$. Equation 6.4 then simply reduces to the well-known fact that the power spectrum is the Fourier representation of the correlation function.

Different choices for the matrix $\mathbf{E}^\alpha$ will give power spectrum estimates with different statistical properties. A particularly desirable property to have is for our power spectrum estimate to be free from noise/foreground bias, so that the final estimator $\widehat{p}_\alpha$ depends only on the cosmological power and not on noise or foregrounds. Taking the expectation value of Equation 6.2 and substituting Equation 6.4 yields

$$\widehat{p}_\alpha = \sum_\beta W_{\alpha\beta}\, p_\beta + \mathrm{tr}\left[(\mathbf{C}_{fg} + \mathbf{N})\mathbf{E}^\alpha\right] - b_\alpha, \tag{6.5}$$

where $W_{\alpha\beta}$ is a matrix that we will discuss below. From this expression, we see that to eliminate the noise and foreground biases, one should pick

$$b_\alpha = \mathrm{tr}\left[(\mathbf{C}_{fg} + \mathbf{N})\mathbf{E}^\alpha\right]. \tag{6.6}$$

The presence of $\mathbf{C}_{fg}$ in this expression means that in our formalism, we can consider foreground subtraction to be a two-step process. The first step acts on the data $\mathbf{x}$ directly through $\mathbf{E}^\alpha$ in forming the quantity $(\mathbf{x}-\mathbf{m})^t\mathbf{E}^\alpha(\mathbf{x}-\mathbf{m})$. As we will see explicitly in Sections 6.2.2 and 6.2.3, this step involves not just the Fourier transforming and binning as outlined above, but also a *linear* foreground subtraction acting on the data. The result is a biased first guess at the power spectrum. The second step (where

one subtracts off the b_α term) is a *statistical* removal of foregrounds from this first guess, which acts on a *quadratic* function of the data (since it is applied to the power spectrum estimate and not the data) and is similar in spirit to the methods suggested in Morales et al. (2006). Our formalism builds on the work in Morales et al. (2006) and allows one to compute the appropriate statistical foreground removal term b_α for any quadratic power spectrum estimate — one simply plugs the corresponding $\mathbf{E}^\alpha$ into Equation 6.6.

With b_α chosen appropriately, Equation 6.5 reduces to a relation between our power spectrum estimate and the true power spectrum:

$$\widehat{\mathbf{p}} = \mathbf{W}\mathbf{p}, \tag{6.7}$$

where we have grouped the components of the true power spectrum p_α and our power spectrum estimate $\widehat{p}_\alpha$ into the vectors $\mathbf{p}$ and $\widehat{\mathbf{p}}$ respectively[4], and $\mathbf{W}$ is the *window function matrix*, given by

$$\mathbf{W}_{\alpha\beta} = \mathrm{tr}[\mathbf{C}_{,\beta}\mathbf{E}^\alpha]. \tag{6.8}$$

In general, $\mathbf{W}$ will not be a diagonal matrix, and thus each component of our power spectrum estimate vector $\widehat{\mathbf{p}}$ (*i.e.* each band power) will be a weighted sum of different components of the true power spectrum. Put another way, the power spectrum estimate at one particular point in the $k_\perp$-$k_\parallel$ plane is not merely a reflection of the true power spectrum at that point, but also contains contributions from the true power spectrum at nearby locations on the $k_\perp$-$k_\parallel$ plane. Neighboring points of a power spectrum estimate are thus related to one another, and give rise to "horizontal error bars" in one's final power spectrum estimate. Equation 6.8 allows one to quantify the extent to which foreground subtraction causes unwanted mode-mixing between different parts of k-space, again because the $\mathbf{E}^\alpha$ matrix can be written to include any linear foreground subtraction process. As a general rule of thumb, the broader one's window functions the greater the information loss in going from the true power spectrum to our power spectrum estimate. The window functions are thus a useful diagnostic for evaluating one's estimation method, and we devote Section 6.4.1 to examining the window functions from various methods for estimating the power spectrum.

In addition to computing the window functions, a complete evaluation of one's

[4]This grouping is not to be confused with our earlier grouping of data into the vector $\mathbf{x}$. The index on $\mathbf{x}$ runs over the spatial grid points of one's measured brightness temperature distribution, whereas the index of vectors like $\mathbf{p}$ runs over different bands of $k_\perp$ and $k_\parallel$.

power spectrum estimation technique should also involve a computation of covariance matrix $\mathbf{V}$ of the band powers:

$$\mathbf{V}_{\alpha\beta} = \langle \widehat{p}_\alpha \widehat{p}_\beta \rangle - \langle \widehat{p}_\alpha \rangle \langle \widehat{p}_\beta \rangle. \tag{6.9}$$

Roughly speaking, the diagonal elements of $\mathbf{V}$ give the "vertical error bars" on our power spectrum estimate. If the signal is Gaussian, Equation 6.9 can be written as

$$\mathbf{V}_{\alpha\beta} = \sum_{ijkl} [\mathbf{C}_{ik}\mathbf{C}_{jl} + \mathbf{C}_{il}\mathbf{C}_{jk}]\mathbf{E}^\alpha_{ij}\mathbf{E}^\beta_{kl}. \tag{6.10}$$

Ultimately, one of course wishes to *minimize* the variances (*i.e.* error bars) on our power spectrum estimate. For a deconvolved power spectrum estimate[5], *i.e.*, one where $\mathbf{W} = \mathbf{I}$ so that

$$\langle \widehat{\mathbf{p}} \rangle = \mathbf{p}, \tag{6.11}$$

the smallest possible error bars that can be obtained for a given experimental set-up can be computed using the Fisher matrix formalism. The *Fisher information matrix* is defined as (Fisher, 1935)

$$\mathbf{F}_{\alpha\beta} \equiv -\left\langle \frac{\partial^2}{\partial p_\alpha \partial p_\beta} \ln f \right\rangle, \tag{6.12}$$

where f is the probability distribution for the data vector $\mathbf{x}$, and is dependent on both $\mathbf{x}$ and the band powers p_α, which we defined above. Doing so and assuming that the fluctuations are Gaussian allows one to write the Fisher matrix as

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left[\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta}\mathbf{C}^{-1}\right]. \tag{6.13}$$

By the Cramer-Rao inequality, the quantity $(\mathbf{F}^{-1})^{1/2}_{\alpha\alpha}$ represents the smallest possible error bar on the band power p_α that any method satisfying Equation 6.11 can achieve if one is estimating all the band powers jointly, while $(\mathbf{F}_{\alpha\alpha})^{-1/2}$ gives the best possible error if the other band powers are already known. In our particular case, this means that if a given foreground subtraction and power spectrum estimation technique (specified by the set of $\mathbf{E}^\alpha$s) yields a covariance matrix $\mathbf{V}$ that is equal to

[5] A power spectrum estimator that respects Equation 6.11 is sometimes also said to be an unbiased estimator. This is conceptually separate from our earlier discussion about choosing b_α appropriately to eliminate *noise and foreground bias*. Unfortunately, both usages of the word "bias" are standard. In this chapter we reserve the term for the latter meaning, and instead use "unwindowed" or "deconvolved" when referring to the former.

$\mathbf{F}^{-1}$, the technique is optimal in the sense that no other *unwindowed* method will be able to produce a power spectrum estimate with smaller error bars. This method will turn out to be closely related to the inverse variance scheme that we introduce in Section 6.2.2, where we apply the formalism that we have developed so far to 21 cm tomography. Similarly, in Section 6.2.3 we will take the traditional line-of-sight subtraction algorithms suggested by Wang et al. (2006); Bowman et al. (2009); Liu et al. (2009b,a) and recast them in our mathematical framework. We will see in Section 6.4 that the traditional methods result in a substantial loss of information, resulting in error bars that are larger than one obtains with the inverse variance method.

6.2.2 Inverse Variance Foreground Subtraction and Power Spectrum Estimation

Suppose we form the quantity

$$q_\alpha \equiv \frac{1}{2}(\mathbf{x}-\mathbf{m})^t \mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) - b_\alpha, \tag{6.14}$$

and take $\widehat{\mathbf{p}} = \mathbf{F}^{-1}\mathbf{q}$ to be our power spectrum estimate. In Tegmark (1997c) it was shown that this estimate is precisely the unwindowed (*i.e.*, $\mathbf{W} = \mathbf{I}$), optimal estimator hinted at in the last section. This estimator gives error bars that are exactly those specified by the Cramer-Rao bound.

In practice, however, using this estimator tends to give power spectra that are quite noisy. This is because window functions naturally have a width of order the inverse size of the survey volume, so insisting that $\mathbf{W} = \mathbf{I}$ enforces a rather ill-posed deconvolution that greatly amplifies the noise. This results in band power estimates that have error bars that are both large (despite being the "best" allowed by Cramer-Rao) and anticorrelated between neighboring bands (see Tegmark et al. (2004) for a more extensive discussion). To avoid these issues, it is often preferable to smooth the power spectrum on the $k_\perp$-$k_\parallel$ plane, which mathematically means having a nondiagonal $\mathbf{W}$. This in turn implies that we have $\widehat{\mathbf{p}} = \mathbf{W}\mathbf{p}$, which allows us to evade the Cramer-Rao bound since the $\widehat{\mathbf{p}} = \mathbf{p}$ requirement is no longer enforced. In other words, by allowing some smoothing of our power spectrum estimate, we can construct an estimator with smaller error bars than those given by the Cramer-Rao inequality.

In this chapter, we will construct such an estimator by choosing

$$\mathbf{E}^{\alpha} = \frac{1}{2\mathbf{F}_{\alpha\alpha}}\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}, \tag{6.15}$$

where $\mathbf{F}_{\alpha\alpha}$ are diagonal elements of the Fisher matrix given by Equation 6.13. Just like with Equation 6.14, the choice of Equation 6.15 for $\mathbf{E}^{\alpha}$ represents an inverse variance weighting of the data. This can be seen by inserting Equation 6.15 into Equation 6.2:

$$\widehat{p}_{\alpha} = \frac{1}{2\mathbf{F}_{\alpha\alpha}}(\mathbf{x}-\mathbf{m})^{t}\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) - b_{\alpha}. \tag{6.16}$$

Since the covariance matrix $\mathbf{C}$ is symmetric, using Equation 6.15 corresponds to using an inverse variance weighted data vector ($\mathbf{C}^{-1}\mathbf{x}$) to perform our power spectrum estimate. This weighting procedure acts as our foreground subtraction step, since in Equation 6.4 we included the foregrounds in our covariance matrix. The residual noise and foreground bias is subtracted by the b_{α} term, which is obtained by substituting Equation 6.15 into Equation 6.6. The (nondiagonal) window functions are obtained by inserting Equation 6.15 into Equation 6.8, and it is by imposing the normalization condition $\mathbf{W}_{\alpha\alpha} = 1$ (as is standard) that the $(\mathbf{F}_{\alpha\alpha})^{-1}$ normalization factor appears. Throughout this chapter, Equations 6.15, 6.16, and the corresponding window functions are what we refer to as "the inverse variance method".

In Tegmark (1997c), it was shown that in the limit that the data vector $\mathbf{x}$ is drawn from a Gaussian distribution[6], the inverse variance method beats all other power spectrum estimators, in the sense that *no other method — windowed or not — can deliver smaller error bars on the final power spectrum estimates*. To compute these error bars, we insert Equation 6.15 into Equation 6.10 and (assuming Gaussianity in a manner similar to Tegmark (1997c)), obtain

$$\mathbf{V}_{\alpha\beta} = \frac{\mathbf{F}_{\alpha\beta}}{\mathbf{F}_{\alpha\alpha}\mathbf{F}_{\beta\beta}}. \tag{6.17}$$

[6]Gaussianity is certainly *not* a good assumption for 21 cm tomography. However, even if not strictly optimal, an inverse variance weighting of data is often desirable. Moreover, in the presence of strongly non-Gaussian signals, the power spectrum itself is a non-optimal statistic in that it fails to capture all the cosmological information present in the raw data. Thus, the mere act of replacing the data by a power spectrum is in this sense a Gaussian approximation, and in that limit the inverse variance method is the optimal one. Note also that while the formulae for the error bar estimates (such as Equation 6.18) may suffer from inaccuracies if there are non-Gaussianities, the expressions for the window functions (Equation 6.7) and the noise and foreground bias (Equation 6.6) remain strictly correct. This is because the window functions and the bias terms are derived from manipulating Equation 6.2, which only involves *second moments* of the data vector $\mathbf{x}$ (and are therefore completely describable in terms of covariances), whereas the error bars come from Equation 6.9, which implicitly depends on *fourth moments* of the data.

In particular, the "vertical error bars" on a particular band power are given by the diagonal elements of $\mathbf{V}$, giving

$$\Delta p_\alpha \equiv \mathbf{V}_{\alpha\alpha}^{1/2} = \frac{1}{\sqrt{\mathbf{F}_{\alpha\alpha}}}. \tag{6.18}$$

This means that the inverse variance method delivers error bars that are smaller than those given by the Cramer-Rao bound, because here our error bars are equal to $(\mathbf{F}_{\alpha\alpha})^{-1/2}$, something which can only be achieved by an unwindowed estimator if all but one of the band powers are known beforehand. The windowing has thus achieved our goal of producing a less noisy power spectrum with smaller error bars. Note that this outcome was by no means guaranteed — for instance, in Section 6.4 we will find that the traditional line-of-sight methods described in Section 6.2.3 essentially smooth the $k_\perp$-$k_\parallel$ plane too much, resulting in window functions that are so broad that there is a substantial loss of information, which in turn causes larger error bars than one obtains with the inverse variance method.

Comparing the inverse variance method to the unwindowed estimator $\widehat{\mathbf{p}} = \mathbf{F}^{-1}\mathbf{q}$ discussed earlier, we see that the difference lies in whether one normalizes the power spectrum using $\mathbf{F^{-1}}$ or $(\mathbf{F}_{\alpha\alpha})^{-1}$. While we have just seen that the latter gives smaller error bars on the power spectrum, the choice of normalization becomes irrelevant as one goes beyond the power spectrum to constrain cosmological parameters. This is because both choices consist of multiplying by an invertible matrix, and thus no information is lost. This is not true for the traditional line-of-sight algorithms, where the non-optimal error bars on the power spectrum are due to an irreversible loss of information, which will in turn cause larger error bars on cosmological parameters.

6.2.3 Line-of-Sight Foreground Subtraction

In a typical[7] line-of-sight (LOS) foreground subtraction method, the measured signal from each LOS is plotted as a function of frequency (or, if one prefers, of radial distance) and a low-order polynomial is fit to the data. The low-order fit is then subtracted from the data, with the hope that the remaining signal varies sufficiently rapidly with frequency to be dominated by the cosmological contribution to the signal and not by foregrounds (which are spectrally smooth and therefore expected to be well-approximated by low-order polynomials). Mathematically, if one arranges the elements of the data vector $\mathbf{x}$ so that one cycles through the radial/frequency direction

[7] As noted above, variants of the LOS method exist, but we expect the conclusions of this chapter to be independent of our specific implementation.

most rapidly and the perpendicular/angular directions less rapidly, the action of a LOS foreground subtraction can be described by the equation $\mathbf{z} = \mathbf{D}\mathbf{x}$, where $\mathbf{z}$ is the foreground-cleaned data and $\mathbf{D}$ is a block diagonal matrix. That $\mathbf{D}$ is block diagonal is simply an expression of the fact that different lines-of-sight are independently cleaned in this algorithm, *i.e.* there is no attempt to use the angular structure of foregrounds for cleaning. If one's data cube has $n_\parallel$ pixels along the line of sight direction and a total of $n_\perp$ pixels in the perpendicular directions, then $\mathbf{D}$ consists of $n_\perp$ blocks, each of size $n_\parallel \times n_\parallel$ and of the form

$$\mathbf{D}_{single\,block} = \mathbf{I} - \mathbf{X}[\mathbf{X}^t\mathbf{X}]^{-1}\mathbf{X}^t, \tag{6.19}$$

where $\mathbf{I}$ is identity matrix, $\mathbf{X}$ is an $n_\parallel \times (m+1)$ matrix such that $\mathbf{X}_{ij}$ equals the ith frequency (or radial pixel number) taken to the $(j-1)$th power, and m is the order of the polynomial fit. In Bowman et al. (2009); Liu et al. (2009b) it was found that a *quadratic* polynomial ought to be sufficient for cleaning foregrounds below the level of the expected cosmological signal. In general, one should select a polynomial that is as low an order as possible to mitigate the possibility of accidentally removing part of the cosmological signal during foreground subtraction.

Once foreground subtraction has been performed, we can form the power spectrum by Fourier transforming, binning, and squaring. The Fourier transform and binning are accomplished by the matrix $\mathbf{C}_{,\alpha}$, which is given by Equation 6.3. Subsequently multiplying by $\mathbf{z}^t$ squares the results, so the final estimate of the power spectrum takes the form

$$\widehat{p}_\alpha^{LOS} = \mathbf{z}^t\mathbf{C}_{,\alpha}\mathbf{z} = (\mathbf{x}-\mathbf{m})^t\mathbf{D}\mathbf{C}_{,\alpha}\mathbf{D}(\mathbf{x}-\mathbf{m}). \tag{6.20}$$

Comparing this to Equation 6.2, it is clear that the LOS foreground subtraction methods (and subsequent power spectrum estimations) proposed in the literature are quadratic methods with $\mathbf{E}^\alpha = \mathbf{D}\mathbf{C}_{,\alpha}\mathbf{D}$ and $\mathbf{b}_\alpha = 0$. The window functions can be readily computed by substituting $\mathbf{E}^\alpha = \mathbf{D}\mathbf{C}_{,\alpha}\mathbf{D}$ into Equation 6.8, giving

$$\mathbf{W}_{\alpha\beta}^{LOS} = \mathrm{tr}[\mathbf{C}_{,\alpha}\mathbf{D}\mathbf{C}_{,\beta}\mathbf{D}]. \tag{6.21}$$

As we shall see in Section 6.4, the off-diagonal elements of this matrix are large, which implies that LOS foreground subtraction has the effect of introducing large correlations between the power spectrum estimates at different parts of the $k_\perp$-$k_\parallel$ plane.

Comparing the LOS algorithm to the optimal inverse variance algorithm described

in Section 6.2.2, we can identify three areas in which the LOS algorithm is non-optimal:

1. **The method produces power spectra that still contain a residual noise and foreground bias.** This conclusion has been numerically confirmed by Bowman et al. (2009), Harker et al. (2010), and Petrovic & Oh (2011), and trivially falls out of our analytic framework — Equation 6.6 shows that to eliminate the residual bias, one must subtract

$$\begin{aligned} b_\alpha^{LOS} &= \mathrm{tr}\left[(\mathbf{C}_{fg}+\mathbf{N})\mathbf{E}^\alpha\right] \\ &= \mathrm{tr}\left[(\mathbf{C}_{fg}+\mathbf{N})\mathbf{D}\mathbf{C}_{,\alpha}\mathbf{D}\right] \end{aligned} \tag{6.22}$$

   from the initial estimate, so the setting of $\mathbf{b}_\alpha^{LOS}$ to zero means the method is biased. Fortunately, since $\mathbf{D}$ and $\mathbf{C}_{,\alpha}$ are known *a priori* and $\mathbf{C}_{fg}$ and $\mathbf{N}$ can be modeled (see Section 6.3 for details), this bias can be easily removed.

2. **The method does not make full use of the available data.** The foreground cleaning in the LOS method can be thought of as a projection of the data vector $\mathbf{x}$ into the subspace orthogonal to low order polynomials in the frequency direction[8]. To see that, note that the matrix $\mathbf{D}$ defined above takes the form of a symmetric ($\mathbf{D} = \mathbf{D}^t$) projection matrix ($\mathbf{D}^2 = \mathbf{D}$), so using the vector $\mathbf{z} = \mathbf{D}\mathbf{x}$ to estimate our power spectrum essentially amounts to limiting our analysis to the subset of data orthogonal to the polynomial modes. Such a procedure, where one projects out the modes that are *a priori* deemed contaminated, is exactly analogous to similar techniques in CMB data analysis (where one customarily removes the monopole and dipole modes, as well as pixels close to the Galactic plane) and galaxy survey analysis (where one might project out some purely angular modes in order to protect against incorrectly modeled dust extinction).

   The projecting out of contaminated modes will necessarily result in larger error bars in one's final power spectrum estimate, because cosmological information is irreversibly destroyed in the projection procedure. Indeed, this has been a concern with LOS subtraction, for any component of the cosmological signal that is non-orthogonal to low-order polynomials in frequency will be inadvertently subtracted from the data. At best, if the cosmological signal happens

[8] A complementary treatment of LOS foreground subtraction that also casts the subtraction as a projection of data onto a subspace of polynomials can be found in McQuinn et al. (2006b).

to be completely orthogonal to the polynomial modes, estimating the power spectrum from the projected data will give error bars that are identical to an optimal method that uses the full data, since the optimal method can simply assign zero weight to the contaminated modes.

3. **The method is a non-optimal estimator even for the projected data.** The prescription given by Equation 6.20 calls for a uniform weighting of the projected data $\mathbf{z}$ in the estimation of the power spectrum. In Tegmark et al. (1998), it was shown that ideally one should instead apply inverse variance weighting to the remaining data. Since the covariance matrix of the reduced data $\tilde{\mathbf{C}} \equiv \mathbf{DCD}$ is singular[9], the inverse variance weighting is accomplished using the so-called *pseudoinverse*, given by

$$\mathbf{M} \equiv \mathbf{D}[\tilde{\mathbf{C}} + \eta \mathbf{X}\mathbf{X}^t]^{-1}\mathbf{D}, \tag{6.23}$$

where η is a non-zero constant and the matrix $\mathbf{X}$ is the same as that defined for Equation 6.19. This pseudoinverse can be shown to be independent of η Tegmark (1997c), and has the property that $\tilde{\mathbf{C}}\mathbf{M} = \mathbf{I}$ in the subspace of remaining modes, as one would desire for an inverse. To estimate a power spectrum with an inverse variance weighting of the projected data, one simply acts with the pseudoinverse after the projection:

$$\mathbf{z} = \mathbf{MDx} = \mathbf{Mx}, \tag{6.24}$$

where in the final step we made use of the fact that $\mathbf{D}^2 = \mathbf{D}$, so $\mathbf{MD} = \mathbf{M}$ from Equation 6.23.

With this weighting, the resulting power spectrum estimate becomes optimal for the projected data, in that the covariance matrix $\mathbf{V}_{\alpha\beta}$ becomes equal to the expression given in Equation 6.17, except that one uses not the full Fisher matrix $\mathbf{F}$ but the Fisher matrix of the projected data $\tilde{\mathbf{F}}$. This Fisher matrix can be proven (Tegmark, 1997c) to take the same form as Equation 6.13, except with the pseudoinverse taking the place of the inverse covariance:

$$\tilde{\mathbf{F}}_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left[\mathbf{C}_{,\alpha}\mathbf{M}\mathbf{C}_{,\beta}\mathbf{M}\right]. \tag{6.25}$$

In Section 6.4 we will use Equation 6.25 (inserted into Equation 6.18) to estimate the

[9] An unsurprising result, given that we have thrown away select modes in $\mathbf{x}$.

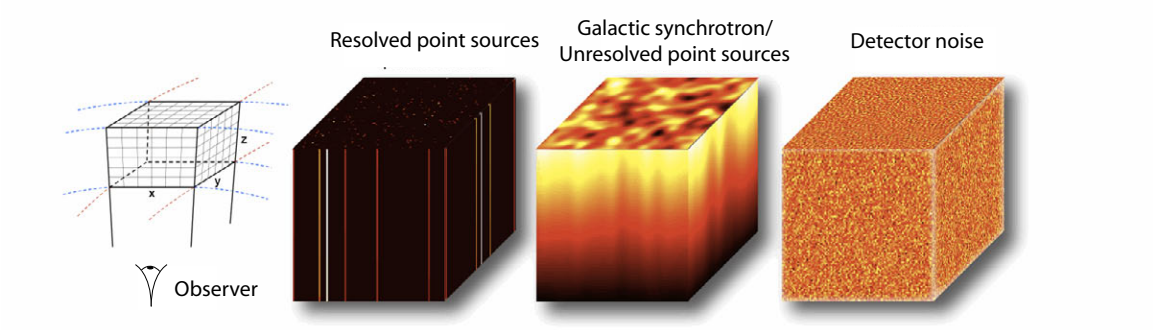

Figure 6-1: Schematic visualization of various emission components that are measured in a 21 cm tomography experiment. From left to right: The geometry of a data "cube", with the line-of-sight direction/frequency direction as the z-axis; resolved point sources (assumed to be removed in a prior foreground step and therefore not included in the model presented in Section 6.3), which are limited to a few spatial pixels but have smooth spectra; Galactic synchrotron radiation and unresolved point sources, which again have smooth spectra, but contribute to every pixel of the sky; and detector noise, which is uncorrelated in all three directions.

errors for the LOS method, even though the LOS method as proposed in the literature does not optimally weight the data. This is because once $\mathbf{M}$ has been computed, the optimal weighting can be accomplished relatively easily, and moreover most forecasts of the ability of 21 cm tomography to constrain cosmological parameters assume that the weighting is optimal (see for instance Mao et al. (2008)). The LOS method that we use to generate the results in Section 6.4 should therefore be considered an improved version of the traditional methods found in the literature. Even so, we will find that the inverse variance method does better.

6.3 Foreground and Noise Modeling

Unlike the original LOS subtraction discussed in the first part of Section 6.2.3, the optimally weighted version of the LOS subtraction scheme and the pure inverse variance scheme of Section 6.2.2 perform foreground subtraction in ways that are *not* blind. In other words, to inverse variance weight the data it is necessary to have a model for the covariance matrix $\mathbf{C}$, which in turn means that one must have a foreground model, since the foregrounds appear in Equation 6.4. We note, however, that since Equation 6.7 depends only on geometry (via $\mathbf{C}_{,\beta}$) and one's chosen power spectrum estimation *method* (via $\mathbf{E}^{\alpha}$), the window function matrix $\mathbf{W}_{\alpha\beta}$ remains strictly correct even if the foreground model is not. In other words, even though the $\mathbf{E}^{\alpha}$ matrix for our inverse variance scheme involves factors of $\mathbf{C}^{-1}$, this does not detract

from the accuracy of our window functions, since an incorrect $\mathbf{C}$ will simply result in a different $\mathbf{E}^\alpha$, which will in turn give different window functions that nonetheless accurately reflect what was done to the data. It should also be noted that even with schemes where the foreground subtraction is completely blind, a proper estimation of the error bars will necessarily involve a foreground model. We also re-emphasize that the methods presented in Section 6.2 are generally applicable to *any* foreground model, and that the foreground model described in this section is used only for the numerical case study presented in Section 6.4.

The foreground model that we use in this chapter contains the following features:

1. Following Wang et al. (2006); Bowman et al. (2009); Liu et al. (2009b,a), we assume that bright, resolved point sources above some flux S_{max} have been identified and removed from the data. This may be done using traditional radio astronomy algorithms such as CLEAN (Högbom, 1974; Clark, 1980) or more recently developed techniques suitable for low-frequency radio interferometers (e.g. Bernardi et al. (2011); Pindor et al. (2010)).[10]

2. Based on extrapolations from CMB data (Tegmark et al., 2000), we assume that free-free emission is negligible compared to Galactic synchrotron radiation at the relevant frequencies. Since both free-free emission and Galactic synchrotron radiation have spectra that are well-described by the same functional form (Wang et al., 2006), we can safely ignore free-free emission for the purposes of this chapter, because its contribution would be subdominant to the uncertainties inherent in the Galactic synchrotron parameters.

3. With bright point sources removed and free-free emission safely ignorable, the foreground sources that remain are unresolved point sources and Galactic synchrotron radiation.

Since the two foreground sources in our model are independent in origin, we may assume that their contributions to the signal are uncorrelated, thus allowing us to write the foreground covariance $\mathbf{C}_{fg}$ as the sum of an unresolved point source covariance $\mathbf{C}_{pt}$ and a Galactic synchrotron covariance $\mathbf{C}_{sync}$. Figure 6-1 provides a visualization of these various components, and in the following subsections we consider each term in turn.

[10]For an analysis of how such bright point sources can affect the errors on one's final power spectrum, see Datta et al. (2010). Their analysis is complementary to what is done in this chapter, in that they consider the effects of bright point source removal, whereas we consider the effects of spatially extended foregrounds.

6.3.1 Unresolved Point Sources

Consider first the unresolved point source covariance. Since the unresolved point sources can still be considered "local" in radial distance compared to the cosmological signal, the frequency dependence of these foregrounds is due entirely to the chromaticity of the sources. This means that correlations in the LOS/frequency "direction" are independent of the spatial distribution of the sources, and the covariance matrix can be modeled as a separable function in frequency and angular position:

$$\begin{aligned} \mathbf{C}_{pt}(\mathbf{r},\mathbf{r}') &\equiv \langle(\mathbf{x}(\mathbf{r})-\mathbf{m}(\mathbf{r}))(\mathbf{x}(\mathbf{r}')-\mathbf{m}(\mathbf{r}'))^t\rangle \\ &\equiv \Gamma(\nu,\nu')\Theta(\mathbf{r}_\perp,\mathbf{r}'_\perp). \end{aligned} \tag{6.26}$$

The function $\Theta(\mathbf{r}_\perp,\mathbf{r}'_\perp)$ encodes the spatial distribution of point sources, and is therefore where the clustering of point sources manifests itself. While point sources are known to be clustered, the clustering is weak and for many applications it is permissible to ignore the clustering (Tegmark & Efstathiou, 1996). If clustering is ignored, the number of sources in a given pixel on the sky can be modeled using a Poisson distribution, and the resulting angular power spectrum of point sources is flat (*i.e.* independent of angular scale) (Tegmark & Villumsen, 1997). This gives rise to a perpendicular correlation function/covariance Θ proportional to $\delta(\mathbf{r}_\perp-\mathbf{r}'_\perp)$.

It should be emphasized, however, that it is certainly not necessary to assume that point sources are unclustered. Whether or not clustering is included, the foreground cleaning and power spectrum estimation procedure described in Section 6.2.2 remains the same. One simply inserts a different covariance matrix $\mathbf{C}$ into the quadratic estimator. In fact, the inclusion of clustering will aid foreground subtraction if the clustering pattern differs significantly from that of the cosmological signal.

Unfortunately, only future experiments will be able to probe the angular clustering of point sources on the fine arcminute scales probed by 21 cm tomography. For simplicity, we therefore model the Θ function as a Gaussian in the perpendicular separation $r_\perp \equiv |\mathbf{r}_\perp-\mathbf{r}'_\perp|$. Without loss of generality, we can normalize so that $\Theta=1$ when $\mathbf{r}_\perp=\mathbf{r}'_\perp$, because the overall amplitude of the foregrounds can be absorbed into $\Gamma(\nu,\nu')$. This means that

$$\Theta(\mathbf{r}_\perp,\mathbf{r}'_\perp) = e^{-\frac{r_\perp^2}{2\sigma_\perp^2}}, \tag{6.27}$$

where the $\sigma_\perp$-parameter is chosen to take the small value ~ 7 arcminutes to reflect the weakness of the clustering.

In contrast to the perpendicular directions, one expects correlations in the LOS

direction to be strong. In other words, $\Gamma(\nu, \nu')$ should be large even for $\nu \neq \nu'$ because of the high spectral coherence of foregrounds. To compute this correlation, consider first a simplified model where we have a population of point sources with an average number density n per steradian, giving an average of $n\Omega_{pix}$ sources per sky pixel of size Ω_{pix} steradians. We imagine that all the point sources in this population have the same flux S_* at frequency $\nu_* = 150\,\mathrm{MHz}$, and a power law frequency spectrum $\propto \nu^{-\alpha}$, with a spectral index α drawn from a Gaussian distribution

$$p(\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left[-\frac{(\alpha - \alpha_0)^2}{2\sigma_\alpha^2}\right], \tag{6.28}$$

with $\sigma_\alpha = 0.5$ (Tegmark et al., 2000) and α_0 to be fixed later. With this foreground model, the average signal along a particular LOS is given by

$$m(\nu) \equiv \langle x(\nu)\rangle = \left(\frac{A_\nu}{\Omega_{pix}}\right)(n\Omega_{pix}S_*)\int \left(\frac{\nu}{\nu_*}\right)^{-\alpha} p(\alpha)d\alpha, \tag{6.29}$$

where A_ν/Ω_{pix} is a conversion factor that converts our expression from flux units to temperature units. The ν subscript serves to remind us that it is a function of frequency, and numerically we have

$$\left(\frac{A_\nu}{\Omega_{pix}}\right) = 1.4\times 10^{-6}\left(\frac{\nu}{\nu_*}\right)^{-2}\left(\frac{\Omega_{pix}}{1\,\mathrm{sr}}\right)^{-1}\mathrm{mJy}^{-1}\,\mathrm{K}. \tag{6.30}$$

Similarly, the covariance (or correlation function) is

$$\begin{aligned}\Gamma(\nu,\nu') &\equiv \langle (x(\nu)-m(\nu))(x(\nu')-m(\nu'))\rangle \\ &= \frac{A_\nu A_{\nu'}}{\Omega_{pix}^2} n\Omega_{pix}S_*^2\int\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha} p(\alpha)d\alpha.\end{aligned} \tag{6.31}$$

Note that the covariance is proportional to n and not n^2 because in a Poisson distribution[11] the mean is equal to the variance. Evaluating the integral gives

$$\Gamma(\nu,\nu') = \frac{A_\nu A_{\nu'}}{\Omega_{pix}} nS_*^2\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_0+\frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)}, \tag{6.32}$$

which shows that the superposition of power law spectra with Gaussian-distributed spectral indices gives exactly a power law with a positive running of the spectral

[11] This is not to be confused with the spatial distribution of the sources. What is being modeled as Poisson distributed is the number of sources in a given pixel.

index.

Now, in reality one of course has multiple populations of point sources with varying brightness S_*. If we treat these populations as independent (in a similar fashion to what was done in Tegmark & Villumsen (1997)), then the total covariance due to all populations is given by the *sum* of the covariances of the individual populations. In the limit of an infinite number of populations, one has

$$\Gamma(\nu,\nu') = \frac{A_\nu A_{\nu'}}{\Omega_{pix}} \int_0^{S_{max}} \frac{dn}{dS_*} S_*^2 dS_* \left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_0 + \frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)}, \tag{6.33}$$

where dn/dS is the differential source count and S_{max} is a maximum point source flux above which we assume the sources can be resolved and removed from the data prior to applying our foreground subtraction and power spectrum estimation scheme. Simulations of so-called "peeling" techniques have suggested that resolved point sources can be reliably removed down to $S_{max} \sim 10$ to $100\,\mathrm{mJy}$ (Pindor et al., 2010). In this chapter, we go a little above the middle of this range and take $S_{max} \sim 60\,\mathrm{mJy}$. For the differential source count we use empirical fit of Di Matteo et al. (2002), which takes the form

$$\frac{dn}{dS_*} = \left(4.0\,\mathrm{mJy}^{-1}\,\mathrm{sr}^{-1}\right)\left(\frac{S_*}{880\,\mathrm{mJy}}\right)^{-1.75}, \tag{6.34}$$

and putting everything together, we obtain

$$\Gamma(\nu,\nu') = (149\,\mathrm{K}^2)\left(\frac{\Omega_{pix}}{10^{-6}\,\mathrm{sr}}\right)^{-1}\left(\frac{\nu\nu'}{\nu_*^2}\right)^{-\alpha_{ps} + \frac{\sigma_\alpha^2}{2}\ln\left(\frac{\nu\nu'}{\nu_*^2}\right)}, \tag{6.35}$$

where $\alpha_{ps} \equiv \alpha_0 + 2 = 2.5$, with this numerical value chosen[12] to match extrapolations from CMB data (Tegmark et al., 2000). Combining this with Equation 6.27 gives us our expression for the unresolved point source foreground covariance matrix.

6.3.2 Galactic Synchrotron Radiation

With the synchrotron contribution, it is more difficult to write down the form of the covariance matrix, because it is unclear how to rigorously define the ensemble averages required in the computation of the covariance. In the frequency direction, we simply use the same expression as we obtained for the unresolved point sources.

[12]For the results presented in Section 6.4, we in fact used larger values for both α_{ps} and α_{sync}. We estimate the difference to be no larger than 1% at any frequency. Moreover, with larger spectral indices the foregrounds become steeper functions of frequency, and are therefore harder to subtract out, making our results more conservative.

We expect this to be a reasonable model for two reasons. Physically, the emission from unresolved point sources is also synchrotron radiation, albeit from distant galaxies. Empirically, the spectrum of Galactic synchrotron foregrounds is well-described by the same parametric fits that work well for the unresolved point sources, except with slightly different parameters (Tegmark et al., 2000; Wang et al., 2006). We thus reuse Equation 6.35 with $\alpha_{ps} \rightarrow \alpha_{sync} = 2.8$, $\sigma_\alpha = 0.4$, and the amplitude of Γ (*including* the Ω_{pix} factor) to be $1.1 \times 10^5\,\mathrm{K}^2$, all of which are values more suitable for Galactic synchrotron radiation (Wang et al., 2006). As for the perpendicular part of the covariance matrix, we once again run into the same problem we faced with the unresolved point sources, namely that there is a lack of empirical spatial correlation data on the fine scales of interest. For simplicity, we use Equation 6.27 for the Galactic synchrotron foreground contribution as well, but with $\sigma_\perp \sim 10\,\mathrm{degrees}$ to reflect the fact that the Galactic synchrotron contribution is expected to be much more spatially correlated than the point sources are (Tegmark et al., 2000).

6.3.3 Instrumental Noise

To model the instrumental noise contribution to the covariance matrix, we use the standard formula for the uncertainty in a measurement of the spatial wavevector $\mathbf{k}_\perp$ component of the sky brightness:

$$\widetilde{\Delta T}(\mathbf{k}_\perp, r^\parallel) = \frac{\lambda^2 T_{sys}}{A_e \sqrt{\tau_k \Delta\nu}}, \tag{6.36}$$

where λ is the observing wavelength, $\Delta\nu$ is the channel bandwidth of the instrument, A_e is the effective area of an antenna, T_{sys} is the system temperature of the instrument, and τ_k is the total integration time that the instrument spends observing Fourier mode $\mathbf{k}_\perp$ (see Morales (2005) for a pedagogical discussion). The dependence on the radial distance $r^\parallel$ enters via λ, whereas the dependence on $\mathbf{k}_\perp$ enters via τ_k. The amount of time τ_k that each Fourier mode is observed for is greatly affected by the layout of one's interferometer array, since each baseline in the array probes a particular Fourier mode $\mathbf{k}_\perp$ at any one instant. Once $\widetilde{\Delta T}$ has been determined, one must take Fourier transforms in the two perpendicular directions to obtain ΔT in position space. One can then form the quantity $\mathbf{N} \equiv \langle \Delta T(\mathbf{r}) \Delta T(\mathbf{r}') \rangle$, with the additional standard assumption that the noise is uncorrelated between different frequencies.

In this chapter, we pick fiducial system parameters that are similar to those of the MWA. Specifically, the results in the following sections are computed for an interfer-

ometer array with 500 antenna tiles randomly distributed with a density varying as r^{-2}, where r is the distance from the center of the array. The length of the maximum baseline is taken to be 1500 m, and we assume 1000 hrs of observing time. Rotation synthesis is performed with the telescope located at the South Pole for simplicity. The channel width of a single frequency channel is set at $\Delta\nu = 30\,\mathrm{kHz}$, the system temperature at $T_{sys} = 440\,\mathrm{K}$, and we take the effective area of each antenna tile to be $A_e = 15\,\mathrm{m}^2$. These quantities can be taken to be roughly constant over the narrow frequency range of 150 to 150.9 MHz (corresponding to 30 adjacent frequency channels) that we take to define the radial boundaries of the survey volume we use for the numerical case study of Section 6.4. In the perpendicular directions we take our field of view to be $1.2\,\mathrm{deg}^2$, which gives 16 pixels along each perpendicular direction if we take the natural pixel size of $\Omega_{pix} \approx 4.5\,\mathrm{arcmin}^2$ that is suggested by our array configuration. This results in a data cube with $16 \times 16 \times 30 = 7680$ voxels[13], which is a tiny fraction of the total data output of a typical 21 cm tomography experiment, but a reasonable amount of data to analyze for several reasons:

1. **Cosmological evolution of the signal.** As we remarked in Section 6.2.1, the power spectrum is only a sensible statistic if translational invariance is assumed, and this in turn rests on the assumption that the cosmological evolution of the signal is negligible over the redshift range of one's data cube. In practice this means that one must measure the power spectrum separately for many data cubes, each of which has a short radial extent. The configuration that we use for the numerical case study in Section 6.4 should be considered *one* such small cube.

2. **Comparisons between inverse variance subtraction and line-of-sight subtraction.** In Liu et al. (2009b), it was shown that line-of-sight foreground subtraction is most effective over narrow frequency ranges. As we will discuss in Section 6.6.2, the inverse variance method suffers no such limitation. However, we select a narrow frequency range in order to demonstrate (see Section 6.4) that the inverse variance method does a better job cleaning foregrounds even in that regime.

3. **Computational cost.** Even with $n_{pix} = 16^2 \times 30 = 7680$ voxels, the com-

[13]Note that while we speak of *organizing* the data into a cube (see the left-most graphic in Figure 6-1), the physical survey volume need not be cubical. In our computations, for instance, we take into account the fact that light rays diverge, thus giving a survey volume that takes the form of a small piece of a spherical shell. The formalism described in Section 6.2 can be applied to *any* survey geometry, even ones that are not contiguous.

putational cost is substantial. This is due to the fact that one must multiply and/or invert many 7680×7680 matrices to use Equations 6.6, 6.8, and 6.10. Because of this, the results in Section 6.4 required more than half a terabyte of disk space for matrix storage and a little over a CPU-year of computation. The problem quickly gets worse with increasing n_{pix}, since a straightforward implementation of the computations involved scales as $\mathcal{O}(n_{pix}^3)$. Fortunately, for large datasets there exist iterative algorithms for power spectrum estimation that scale as $\mathcal{O}(n_{pix} \log n_{pix})$ (Pen, 2003; Padmanabhan et al., 2003), and with a few approximations (such as the flat-sky approximation) these fast algorithms can be suitably adapted to perform unified foreground subtraction and power spectrum estimation for 21 cm tomography as we have done in this chapter (Dillon et al., 2012).

We stress, however, that the formalism described in Section 6.2 is in principle applicable to arbitrarily large survey volumes of any geometry.

6.3.4 Cosmological Signal

For the computations in Section 6.4, we ignore the cosmological signal. This is justified for several reasons. First, the signal is expected to be at least a factor of 10^4 smaller than the foregrounds (de Oliveira-Costa et al., 2008), which means that including the signal would make little difference to the inverse variance foreground cleaning steps, as $\mathbf{C}$ would be essentially unchanged. The line-of-sight methods, of course, remain strictly unchanged as they are blind. Despite this, one might still object that if foreground cleaning is successful, then we should expect the cosmological signal to be the dominant contribution to the power spectrum in at least a portion of the $k_\perp$-$k_\parallel$ plane. This is most certainly true, and the simulation of an actual measurement of a power spectrum would be incorrect without including the cosmological signal as input. However, in this chapter we do not show what a typical measured power spectrum might look like, but simply seek to quantify the errors and biases associated with our methods. In this context, the main effect of a cosmological signal would be to add a cosmic variance contribution to our error bars at low $k_\parallel$ and $k_\perp$, but as we will see in Section 6.4 it is precisely at the large scale modes that foreground subtraction is least successful. The cosmic variance errors would thus be dominated by the errors induced by imperfect foreground subtraction anyway.

In short, then, it is not necessary to include the cosmological signal in our computations, and summing just $\mathbf{C}_{fg}$ and $\mathbf{N}$ yields a fine approximation to the full covariance

matrix of the system $\mathbf{C}$ for our purposes.

6.4 Computations and Results

In this section we compute window functions, covariances, and biases for both the inverse variance and LOS methods described above. Using Equations 6.6, 6.8, and 6.10, our computations can be performed without the input of real data, and depend only on the parameters of the system.

In what follows, we choose the k-bands of our power spectrum parameterization to have equal logarithmic widths. In the $k^{\parallel}$ direction we have 30 bands spanning from $0.1\,\mathrm{Mpc}^{-1}$ to $10\,\mathrm{Mpc}^{-1}$. In the $k^{\perp}$ direction we again have 30 bands, but this time from $0.01\,\mathrm{Mpc}^{-1}$ to $1\,\mathrm{Mpc}^{-1}$. Note that these bands can be chosen independently of the instrumental parameters, in the sense that a quadratic estimator can be formed regardless of what $(k^{\perp}, k^{\parallel})$-values are chosen, although the quality of the power spectrum estimate will of course depend on the instrument.

6.4.1 Window Functions

In evaluating the quality of one's power spectrum estimator, one is ultimately interested in the size of the error bars. This means that the quantity of interest is the covariance matrix $\mathbf{V}$. In this section, however, we first examine the window functions. As we saw from the discussion in Section 6.2.2, the window functions are intimately connected to the covariances, and often provide a more detailed understanding for *why* the error bars behave the way they do. In particular, window functions that are wide and badly behaved tend to come hand in hand with large error bars. Moreover, our expression for the window function matrix (Equation 6.7) remains strictly accurate even in the presence of non-Gaussian signals (see footnote 6), so in many ways the window function matrix $\mathbf{W}$ is a more robust diagnostic tool than the covariance matrix $\mathbf{V}$.

In the top panel of Figure 6-2, we examine a situation where $\mathbf{C} = \mathbf{I}$, corresponding to a scenario where no foregrounds are present and the signal is dominated by white noise. This is of course a highly unrealistic situation, and we include it simply to build intuition. Rather than plotting the raw window functions, we show the normalized quantities

$$\widetilde{\mathbf{W}}_{\alpha\beta} \equiv \frac{\mathbf{W}_{\alpha\beta}}{\sqrt{\mathbf{W}_{\alpha\alpha}\mathbf{W}_{\beta\beta}}}, \tag{6.37}$$

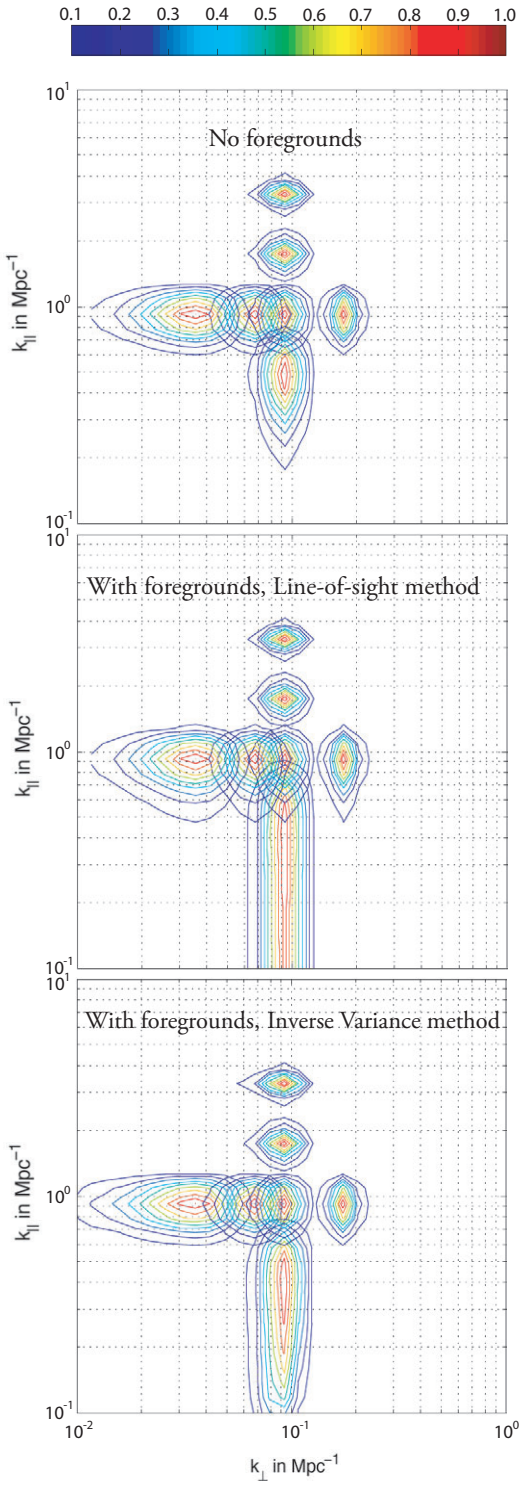

Figure 6-2: Normalized window functions $\widetilde{\mathbf{W}}$ (see Equation 6.37) for an unrealistic situation with no foregrounds (top panel), the line-of-sight foreground subtraction described in Section 6.2.3 (middle panel), and the inverse variance foreground subtraction described in Section 6.2.2 (bottom panel).

where $\mathbf{W}$ is given by Equation 6.7, and like before, each Greek index corresponds to a particular location on the $k_\perp$-$k_\parallel$ plane. Each set of contours shown in Figure 6-2 corresponds to a specific *row* of $\widetilde{\mathbf{W}}$. For example, the set of contours at the center of the "cross" in each panel of Figure 6-2 has a fixed index β that corresponds to $(k_\perp, k_\parallel) = (0.092, 0.92)\,\mathrm{Mpc}^{-1}$ and is plotted as a function of α, or equivalently, as a function of $k_\perp$ and $k_\parallel$. From Equation 6.7, we see that each set of contours describes the weighted average of neighboring points of the true power spectrum p that one is really measuring when forming the power spectrum estimate $\widehat{p}$. With $\mathbf{C}$ set to the identity, the nonzero width of the window functions is due solely to the finite volume of our survey, and any apparent widening of the window functions towards low $k_\perp$ and $k_\parallel$ is due mostly to the logarithmic scale of our plots.

With foregrounds present, the precise shape of the window functions will depend on the foreground subtraction algorithm employed. Shown in the middle panel of Figure 6-2 are window functions for the LOS foreground subtraction described in Section 6.2.3. These window functions have the same $k_\perp$-widths as those in the top panel of Figure 6-2, which is expected since the LOS foreground subtraction scheme operates only in the frequency/radial direction. On the other hand, foreground subtraction increases the $k_\parallel$-widths of the window functions, although this effect is significant only in the lower $k_\parallel$ regions of the $k_\perp$-$k_\parallel$ plane. At high $k_\parallel$, the window functions are very similar to those for the case with no foregrounds, since (by design) the LOS subtraction of low-order polynomials has a negligible effect on high $k_\parallel$ Fourier modes of the signal. As one moves to lower $k_\parallel$, the foreground subtraction has more of an effect, and the window functions widen in the radial direction. This results in large "horizontal" error bars in the $k_\parallel$ direction for the final power spectrum estimate. In addition, the widening of the window functions represents a loss of information at these low $k_\parallel$ values (which is unsurprising, since the very essence of the LOS method is to project out certain modes), and in Section 6.4.2 we will see this manifesting itself as larger "vertical" error bars on the power spectrum estimate.

Shown in the bottom panel of Figure 6-2 are the window functions for the inverse variance foreground subtraction described in Section 6.2.2. The same elongation effect is seen in the $k_\parallel$ direction, but the problem is less severe in the sense that severe elongation does not occur until one reaches much lower values of $k_\parallel$. This is illustrated in Figure 6-3, where we plot the fractional increase (compared to the foreground-free case) in window function width in the $k_\parallel$ direction for both the LOS method (red, solid line) and the inverse variance method (blue, dashed line) as a function of the $k_\parallel$ coordinate of the central peak of the window function. The width is defined to be

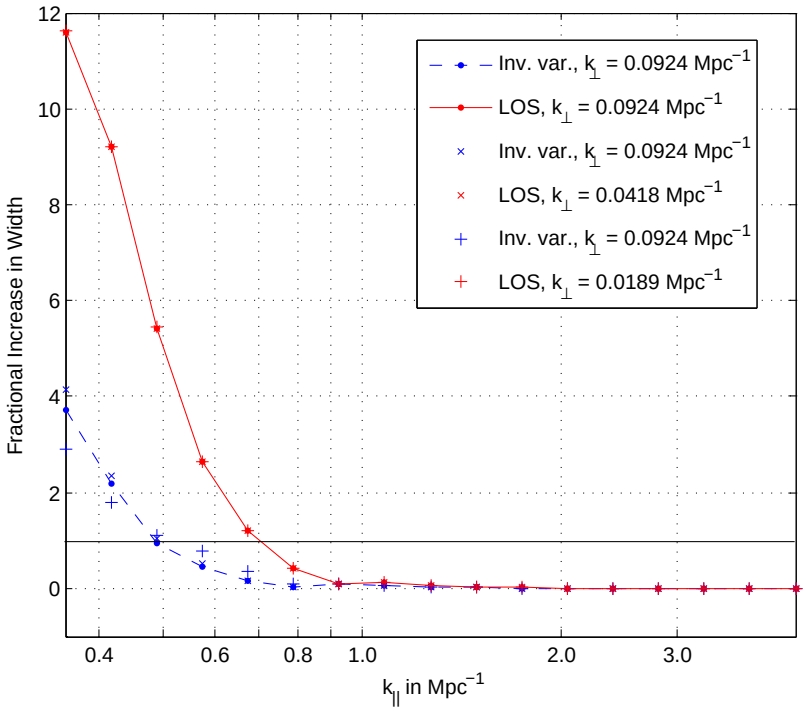

Figure 6-3: Fractional increase (compared to the foreground-free case) in window function width in the $k_\parallel$ direction, plotted as a function of the $k_\parallel$ location (on the $k_\perp$-$k_\parallel$ plane) of the window function peak . The blue, dashed line is for inverse variance foreground subtraction, whereas the red, solid line is for the LOS subtraction. The various data symbols denote widths measured at different values of $k_\perp$, and the fact that they all lie close to the curves suggest that the window function width in the $k_\parallel$ direction is largely insensitive to the $k_\perp$ location of the window function peak. The width is defined as the full-width of the window function in logarithmic k-space at which the function has dropped 1% from its peak value. The solid line is to guide the eye, and marks a fractional increase of unity *i.e.* where the width is double what it would be if there were no foregrounds.

the full-width-99%-max[14] (in logarithmic k-space) of the window in either the $k_\parallel$ or $k_\perp$ direction, and as expected there is essentially no elongation at regions of high $k_\parallel$. Foreground subtraction begins to have an elongating effect on the $k_\parallel$ widths of the window functions at $k_\parallel < 1\,\mathrm{Mpc}^{-1}$, just like we saw from our comparisons of the three panels of Figure 6-2. The effect is much more pronounced for the LOS subtraction, and as an example we can consider the $k_\parallel$ value at which the fractional increase in width is unity (*i.e.* where foregrounds cause the window functions to double in width in the $k_\parallel$ direction). This occurs at $k_\parallel = 0.7\,\mathrm{Mpc}^{-1}$ for the LOS subtraction, but not until $k_\parallel = 0.5\,\mathrm{Mpc}^{-1}$ for the inverse variance subtraction, which means that with the latter scheme, one can push to lower values of $k_\parallel$ before there is significant information loss and the power spectrum band estimates become massively correlated. Note that this conclusion holds for all $k_\perp$, as evidenced by the way the various plot symbols for different $k_\perp$ all lie very close[15] to the curves in Figure 6-3.

Unlike the $k_\parallel$ widths, with the $k_\perp$ widths we find that there are important *qualitative* differences in addition to quantitative differences between the two methods. This is evident in Figure 6-4, where we compare inverse variance and foreground-free window functions (solid and dashed contours respectively) centered at $(k_\perp, k_\parallel) = (0.0259, 0.2593)\,\mathrm{Mpc}^{-1}$. The analogous plot for the LOS method is shown in Figure 6-5. As discussed before, there is an elongation in the $k_\parallel$ direction that is more pronounced in the LOS method than the inverse variance method. For the LOS method, this is all that happens, because the foreground subtraction is performed using only line-of-sight information. In contrast, since the inverse variance method works by de-weighting parts of the data that are heavily contaminated by foregrounds, it is capable of also leveraging information in the transverse direction. This results in the slight widening in the $k_\perp$ direction seen in Figure 6-4.

In Figure 6-6, we show the fractional increase in $k_\perp$ width (again as compared to the foreground-free windows) as a function of $k_\perp$. The width increases towards lower

[14] We define the full-width-99%-max as the full width at which the function has dropped to 99% of its peak value. A more conventional measure of the width like the full-width-half-max (FWHM) is not feasible for our purposes, as the window functions are elongated so much by foreground subtraction that if we defined the widths in such a way, the widths for the windows at the very lowest $k_\parallel$ would run off the edge of our simulation. Our measure of the width allows a quantification of the elongation even for the lowest $k_\parallel$ values. In all regimes where a meaningful measurement of the band powers can be made, the window functions will be compact enough for the width to be defined using the FWHM. With the inverse variance method, for instance, one sees from the bottom panel of Figure 6-2 that the FWHM remains nicely within the simulation box even though it is larger than for the foreground-free case.

[15] For the LOS subtraction, this invariance is in fact exact since the algorithm performs the same polynomial subtraction along the line-of-sight regardless of what happens in the transverse directions.

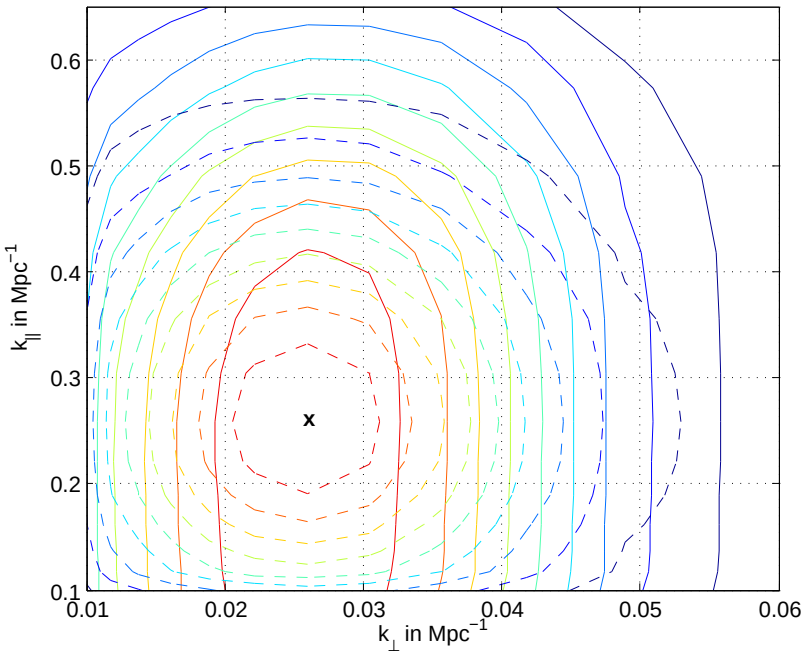

Figure 6-4: A comparison of two normalized window functions $\widetilde{\mathbf{W}}$ centered at $(k_\perp, k_\parallel) = (0.0259, 0.2593)\,\mathrm{Mpc}^{-1}$ (marked by "x"). The dotted contours correspond to a scenario with no foregrounds, while the solid contours correspond to the window functions for the inverse variance foreground subtraction described in Section 6.2.2. Note the linear scale on both axes.

$k_\perp$ because the inverse variance foreground subtraction is taking advantage of the smooth angular dependence of Galactic synchrotron radiation to accomplish some of the foreground cleaning. One can also see that the width decreases with increasing $k_\parallel$, which is due again to the fact that the spectral smoothness of foregrounds mean that they dominate mainly the *low* $k_\parallel$ modes, and so the inverse variance subtraction need not act so aggressively at high $k_\parallel$.

In summary, what we see is that since the inverse variance method downweights data selectively on the basis of expected foreground contamination, its approach to foreground subtraction is more "surgical" than that of the LOS method. For instance, at intermediate $k_\parallel$ (where foregrounds are non-negligible but certainly not at their peak), it subtracts foregrounds in a less aggressive manner than the LOS method does, allowing for final power spectrum estimates in those regions to be less correlated with each other. This is still the case in low $k_\perp$, low $k_\parallel$ regions, but here the algorithm also uses angular information for its foreground subtraction. Put another way, the inverse variance technique automatically "chooses" the optimal way to subtract foregrounds. As a result, the fact that the elongation of window functions in the $k_\parallel$ direction is so

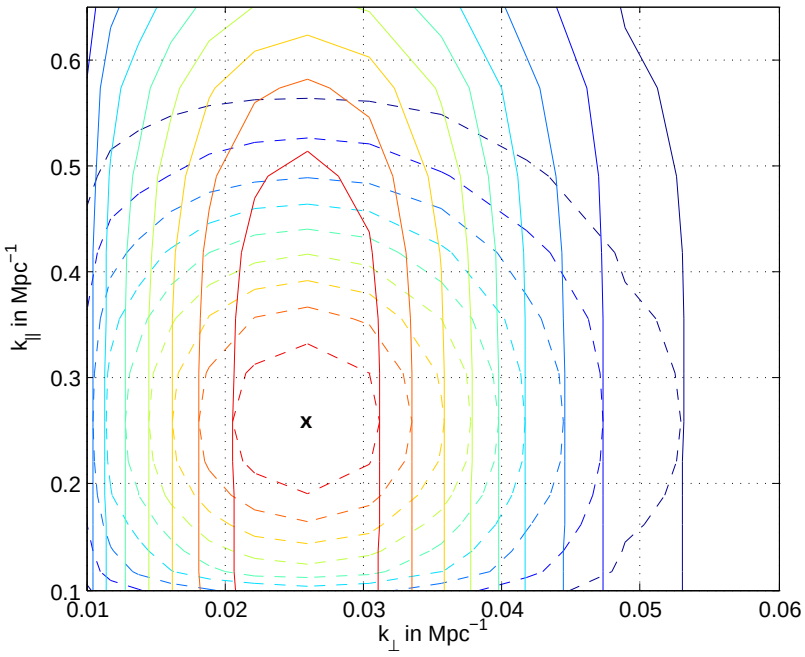

Figure 6-5: A comparison of two normalized window functions $\widetilde{\mathbf{W}}$ centered at $(k_\perp, k_\parallel) = (0.0259, 0.2593)\,\mathrm{Mpc}^{-1}$ (marked by "x"). The dotted contours correspond to a scenario with no foregrounds, while the solid contours correspond to the window functions for the LOS foreground subtraction described in Section 6.2.3. Note the linear scale on both axes.

much greater than the widening in the $k_\perp$ direction (notice the different scales on the vertical axes of Figures 6-3 and 6-6) means that we now have quantitative evidence for what has thus far been a qualitative assumption in the literature — that foreground subtraction for 21-cm tomography is most effectively performed using line-of-sight information rather than angular information.

6.4.2 Fisher Information and Error Estimates

In the top panel of Figure 6-7, we return to the foregroundless scenario where $\mathbf{C} = \mathbf{I}$ and plot the diagonal elements of the Fisher matrix (which should roughly be thought of as being inversely proportional to the square of the power spectrum error bars — see Section 6.2.2). The Fisher information increases from the bottom left to the top right, and in the middle of the $k_\perp$-$k_\parallel$ plane, their contours have a (logarithmic) slope of -2. To understand this, consider the process by which one forms a power spectrum in the $k_\perp$-$k_\parallel$ plane: the three-dimensional real-space data cube is Fourier transformed in all three spatial directions, and then binned and averaged over concentric annuli in

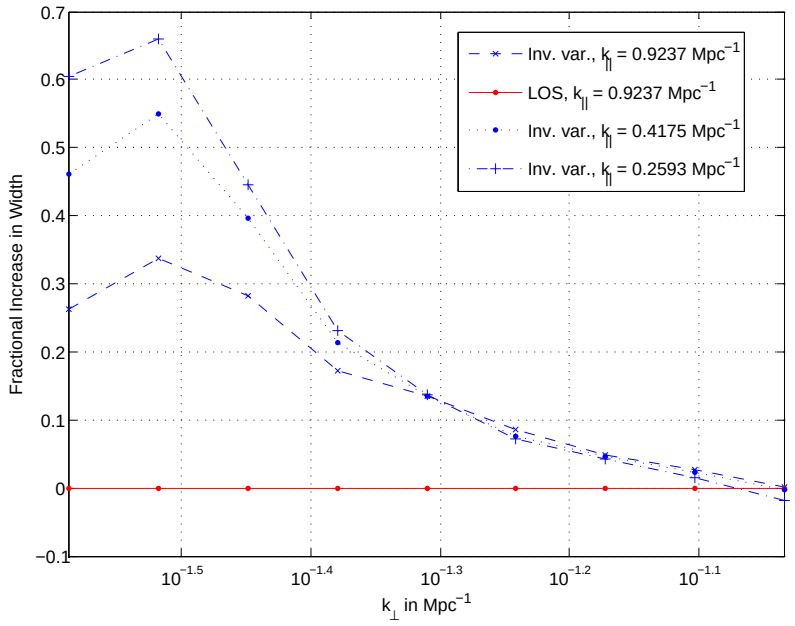

Figure 6-6: Fractional increase (compared to the foreground-free case) in window function width in the $k_\perp$ direction, plotted as a function of the $k_\perp$ location (on the $k_\perp$-$k_\parallel$ plane) of the window function peak . The blue, dashed line is for inverse variance foreground subtraction, whereas the red, solid line is for the LOS subtraction. The dotted and dash-dotted curves display the same quantity as the dashed line, but at lower $k_\parallel$. The width is defined as the full-width of the window function in logarithmic k-space at which the function has dropped 1% from its peak value.

k-space. The averaging procedure has the effect of averaging down noise contributions to the power spectrum estimate, and thus as long as the dominant source of error is instrumental noise, the error bars on a particular part of the power spectrum will decrease with increasing annulus volume. With k-bins of equal *logarithmic* width (see Section 6.4), the volume increases twice as quickly when moving to regions of higher $k_\perp$ in the (logarithmic) $k_\perp$-$k_\parallel$ plane than when moving to regions of higher $k_\parallel$, giving a logarithmic slope of -2 for contours of equal Fisher information. Conversely, Fisher information contours of slope -2 imply that one's error bars are dominated by instrumental noise.

At high $k_\parallel$ or high $k_\perp$, the contours shown in the top panel of Figure 6-7 are seen to deviate from this behavior, although because of the range of the plot, the effects are somewhat visually subtle. At high $k_\parallel$ the contours get *slightly* steeper, implying a lower information content (larger error bars) than one would expect if limited by instrumental noise. This is due to the fact that at very high $k_\parallel$ there is insufficient spectral resolution to resolve the extremely fine small scale modes, resulting in power

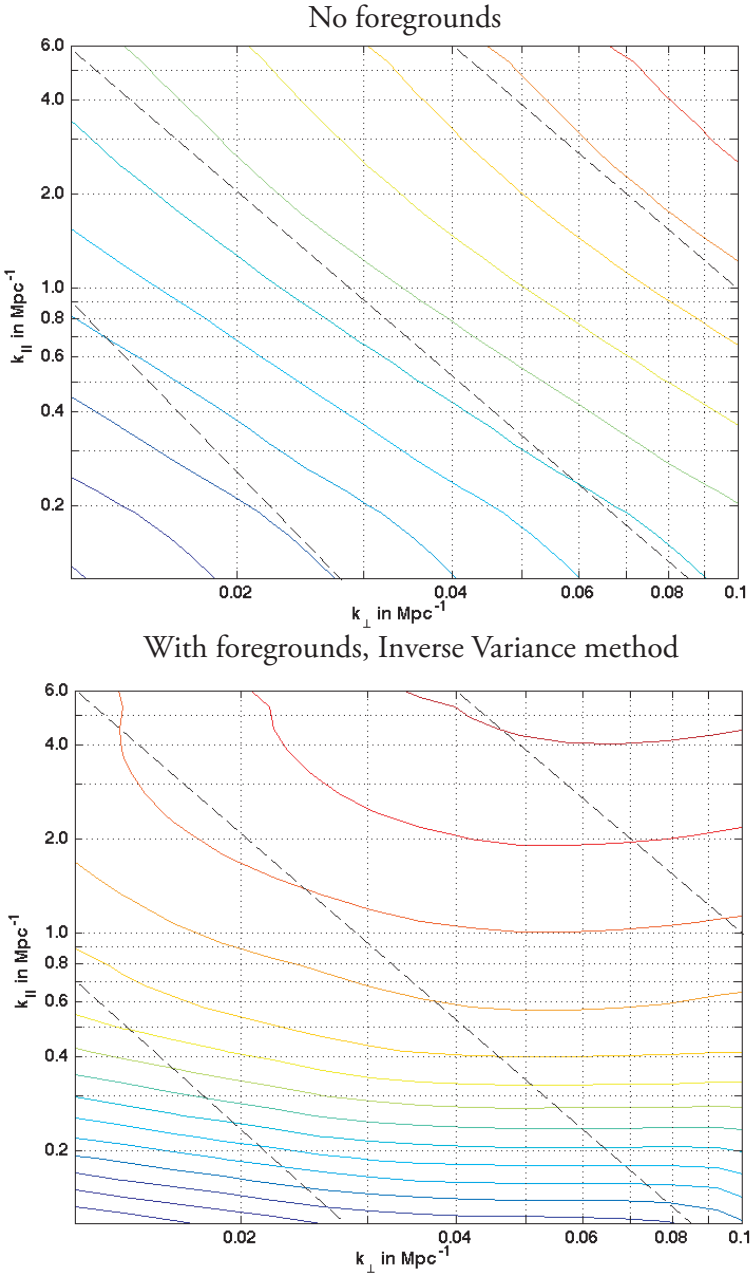

Figure 6-7: A plot of the diagonal elements of the Fisher information matrix for an unrealistic situation with no foregrounds (top panel) and with foregrounds cleaned by the inverse variance method (bottom panel). The (unnormalized) contours increase in value from the bottom left corner to the top right corner of each plot, and are chosen so that crossing every two contours corresponds to an increase by a factor of 10. Dashed lines with logarithmic slopes of -2 are included for reference.

spectrum estimates with large error bars there. Finite angular resolution has a similar effect in the $k_\perp$ direction.

With foregrounds, the Fisher information takes a markedly different form, as seen in the bottom panel of Figure 6-7, where we have plotted the Fisher information after carrying out inverse variance foreground subtraction. One sees that only small regions of the $k_\perp$-$k_\parallel$ plane have contours with a logarithmic slope of -2, suggesting that there are few — if any — parts that are dominated solely by instrumental noise.

The effects of foreground subtraction are more easily interpreted if one instead considers the following quantity[16], which has units of temperature squared:

$$\Delta^2_{21\,\mathrm{cm}}(k_\perp, k_\parallel) \equiv \frac{k_\perp^2 k_\parallel}{2\pi^2} P_T(k_\perp, k_\parallel). \tag{6.38}$$

This is exactly analogous to the quantity

$$\Delta^2(k) \equiv \frac{k^3}{2\pi^2} P(k) \tag{6.39}$$

for galaxy surveys, which quantifies the contribution to the density field variance from a particular logarithmic interval in k, and

$$(\delta T_\ell)^2 \equiv \frac{\ell(\ell+1)}{2\pi} C_\ell \tag{6.40}$$

for CMB measurements, which quantifies the contribution to temperature variance per logarithmic interval in ℓ. The quantity $\Delta^2_{21\,\mathrm{cm}}$ thus measures the contribution to 21 cm brightness temperature variance per logarithmic interval in $(k_\perp, k_\parallel)$-space.

Since Equation 6.17 tells us that the error bar on a particular band of $P_T(k_\perp, k_\parallel)$ is given by $(\mathbf{F}_{\alpha\alpha})^{-1/2}$ for the inverse variance method, we can estimate the error on

[16]The expressions given by Equations 6.38 and 6.40 are often referred to as "dimensionless power spectra". This is of course somewhat of a misnomer, because both expressions carry dimensions of temperature squared, and are "dimensionless" only in the sense that all units of length have been canceled out.

$\Delta_{21\,\mathrm{cm}}$ by computing the quantity[17]

$$\varepsilon(k_\perp, k_\parallel) \sim \left[\frac{k_\perp^2 k_\parallel}{2\pi^2}\sqrt{\frac{1}{F_{\alpha\alpha}}}\right]^{\frac{1}{2}}, \tag{6.41}$$

which, having units of temperature, is convenient for gauging the quality of our foreground cleaning. Note that this expression can also be used for the noiseless case (since $\mathbf{C} \propto \mathbf{I}$ can be considered a special case of the inverse variance method) as well as for the line-of-sight foreground scheme (as long as one uses the pseudoinverse and forms the Fisher matrix using Equation 6.25, as outlined in Section 6.2.3).

Alternatively, we can think of the quantity ε introduced in Equation 6.41 as the *real-world degradation factor*, because it can also be interpreted as the *ratio* (up to an overall dimensionful normalization factor) of the actual error on the power spectrum $P_T(k_\perp, k_\parallel)$ to the error that would be measured by a noisy but otherwise ideal experiment with infinite angular resolution and infinite spectral resolution. In other words, ε tells us how much larger the error bars get because of real-world issues like foregrounds and limited resolution. To see this, note that such an ideal experiment would have a *constant* ε, because the $k_\perp^2 k_\parallel$ factor essentially nulls out the geometric effects seen in the top panel of Figure 6-7 when one is not limited by resolution.

In the top, middle, and bottom panels of Figure 6-8, we show ε for the noise-only scenario, the inverse variance method, and the line-of-sight method, respectively. With the foreground-free case in the top panel, we see that the error bars increase as one goes from the bottom-left corner (low $k_\perp$, low $k_\parallel$) to the top-right corner (high $k_\perp$, high $k_\parallel$). The instrument's finite angular resolution causes the errors to increase in the direction of increasing $k_\perp$, whereas in the direction of increasing $k_\parallel$ this is due to the finite spectral resolution. The smallest error bars on the plot are $\lesssim 10\,\mathrm{mK}$ whereas the largest are $\sim 20\,\mathrm{mK}$.

Introducing foregrounds causes the errors to increase everywhere on the $k_\perp$-$k_\parallel$ plane, even after inverse variance cleaning (middle panel of Figure 6-8). However, in addition to the effects of finite angular resolution and finite spectral resolution, we see that there are now also significant errors at low $k_\parallel$ (indeed, this is where the errors

[17] This expression is in fact an upper limit on the error. That it is not exact arises from our taking of the square root of Equation 6.38 to obtain a quantity with temperature units. In the limit of small errors, for instance, Taylor expanding Equation 6.38 with a perturbation suggests that we ought to have a factor of 1/2 in front of our current form for $\varepsilon(k_\perp, k_\parallel)$. It is only when the errors dominate the signal (such as at low $k_\parallel$ — see Figure 6-8) that ε tends to the expression given in Equation 6.41. For our illustrative purposes, however, Equation 6.41 suffices as a conservative estimate.

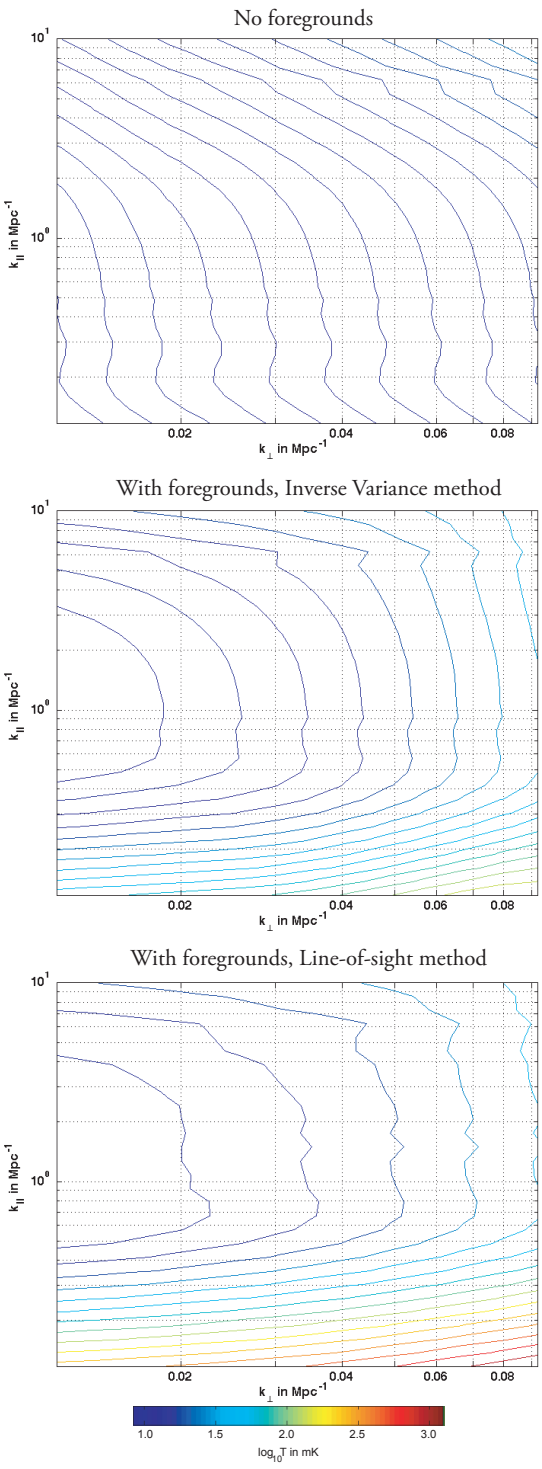

Figure 6-8: Expected power spectrum error bars for a situation with no foregrounds (top panel, Equations 6.13 and 6.41 with $\mathbf{C} \propto \mathbf{I}$, scaled to match the noise-dominated $k_\perp$-$k_\parallel$ regions of the other scenarios); the inverse variance method (middle panel, Equations 6.13 and 6.41); the line of sight method (bottom panel, Equations 6.25 and 6.41). For the last two plots, the errors are high at low $k_\parallel$, high $k_\parallel$, and high $k_\perp$ due to residual foregrounds, limited spectral resolution, and limited angular resolution, respectively.

are largest). This is due to residual foreground contamination — being spectrally smooth, the foregrounds have their greatest effect at low $k_\parallel$, and thus that region of the $k_\perp$-$k_\parallel$ plane is the most susceptible to inaccuracies in foreground cleaning. The "downhill" gradients of the contours there point to higher $k_\parallel$, suggesting that the foregrounds are the dominant source of error.

Considering all the effects together, one sees that the "sweet spot" for measurements of $\Delta_{21\,\mathrm{cm}}$ (or, alternatively, where a realistic experiment compares the most favorably to an ideal experiment for measuring the power spectrum P_T) is at low $k_\perp$ (to avoid being limited by angular resolution) and intermediate $k_\parallel$ (to avoid being limited by spectral resolution or residual foreground contamination). In such a region the typical errors are $\lesssim 10\,\mathrm{mK}$, whereas in the resolution-limited region at the top-right corner of the plot the errors are $\sim 30\,\mathrm{mK}$ and in the foregrounds and angular resolution-limited region in the bottom-right corner the errors are $\sim 130\,\mathrm{mK}$.

Although the features are seen to be qualitatively similar when we move from the inverse variance method to the LOS method (bottom panel of Figure 6-8), quantitatively we see that the errors have grown yet again, regardless of one's location on the $k_\perp$-$k_\parallel$ plane. The smallest errors (found in the low $k_\perp$, intermediate $k_\parallel$ "sweet spot") remain $\lesssim 10\,\mathrm{mK}$, and the errors in the resolution-limited region (high $k_\perp$,$k_\parallel$) remain $\sim 30\,\mathrm{mK}$. However, the foreground contaminated regions take up a larger portion of the $k_\perp$-$k_\parallel$ plane, and the most contaminated parts of the plot have errors of $\sim 900\,\mathrm{mK}$.

As explained in Sections 6.2.2 and 6.2.3, the higher errors are to be expected with the LOS scheme, given that the method does not make full use of the available data, having projected out the low-order polynomial modes. Put another way, the broad window functions in the middle panel of Figure 6-2 indicate that information has been lost and that power has been filtered out of the data. Thus, in order for the power spectrum estimate to not be biased low (because of the power loss), it is necessary to multiply by a large normalization factor. This normalization factor is derived by imposing the $\mathbf{W}_{\alpha\alpha} = 1$ window function constraint (see Tegmark (1997c)), and manifests itself as the $\mathbf{F}_{\alpha\alpha}$ factor in the denominator of Equation 6.16 — if much power has been filtered out of the mode corresponding to the k-space index α, the remaining Fisher information $\mathbf{F}_{\alpha\alpha}$ in that mode will be low, and one must divide by this small number so that the final estimate is of the right amplitude. However, this also has the effect of magnifying the corresponding error bars, leading to the larger errors in the bottom panel of Figure 6-8.

One may be initially puzzled by the trends shown in the last two panels of Figure

6-8, which at first sight appear to differ from plots like those shown in Figure 9 of Bowman et al. (2009). In particular, in Bowman et al. (2009) the post-subtraction foreground residuals are the lowest at low $k_\perp$ and low $k_\parallel$, which seem to imply that the cleanest parts of the final power spectrum ought to be there. This, however, is a misleading comparison. A plot of residuals like that in Bowman et al. (2009) should be viewed as a plot of the expected size of the measured power spectrum (containing both the cosmological signal and the residual foregrounds), not the size of the *error bars* associated with the measurement, which are shown in Figure 6-8. Put another way, the power spectrum is not truly clean at low $k_\perp$ and low $k_\parallel$, for even if the measured values are small, the error bars are large, rendering that part of the power spectrum unusable.

6.4.3 Residual Noise and Foreground Biases

In Figure 6-9, we show the residual noise and foreground bias terms (Equations 6.6 and 6.22, but cast in temperature units in a way similar to Equation 6.41) for the inverse variance (top panel) and LOS (bottom panel) subtraction methods, respectively. The biases that need to be subtracted in both methods are seen to be qualitatively similar, and the trends shown in Figure 6-9 can be readily understood by thinking of the bias subtraction in Equation 6.2 as a foreground subtraction step in its own right. Acting on the $k_\perp$-$k_\parallel$ plane directly, this step seeks to remove any residual foreground power from the power spectrum itself. It therefore has its biggest effect where foreground residuals are expected to be large, and in Section 6.5 we will develop a simple toy model to gain intuition for how this works.

6.5 An Intuitive Toy Model

In this section, we develop a simple toy model for understanding the inverse variance foreground subtraction and power spectrum estimation scheme. The goal here is not to reproduce the exact results of the previous section, but rather, to provide intuitive explanations for how the power spectrum estimation algorithms work.

For our toy model, we consider a one-dimensional survey volume in the line-of-sight direction. This choice is motivated by the fact that it is generally the spectral rather than spatial dependence of foregrounds that is most useful for foreground subtraction, and consequently it is the algorithm's behavior as a function of frequency that is the most interesting. We again take the frequency range of the survey to be small.

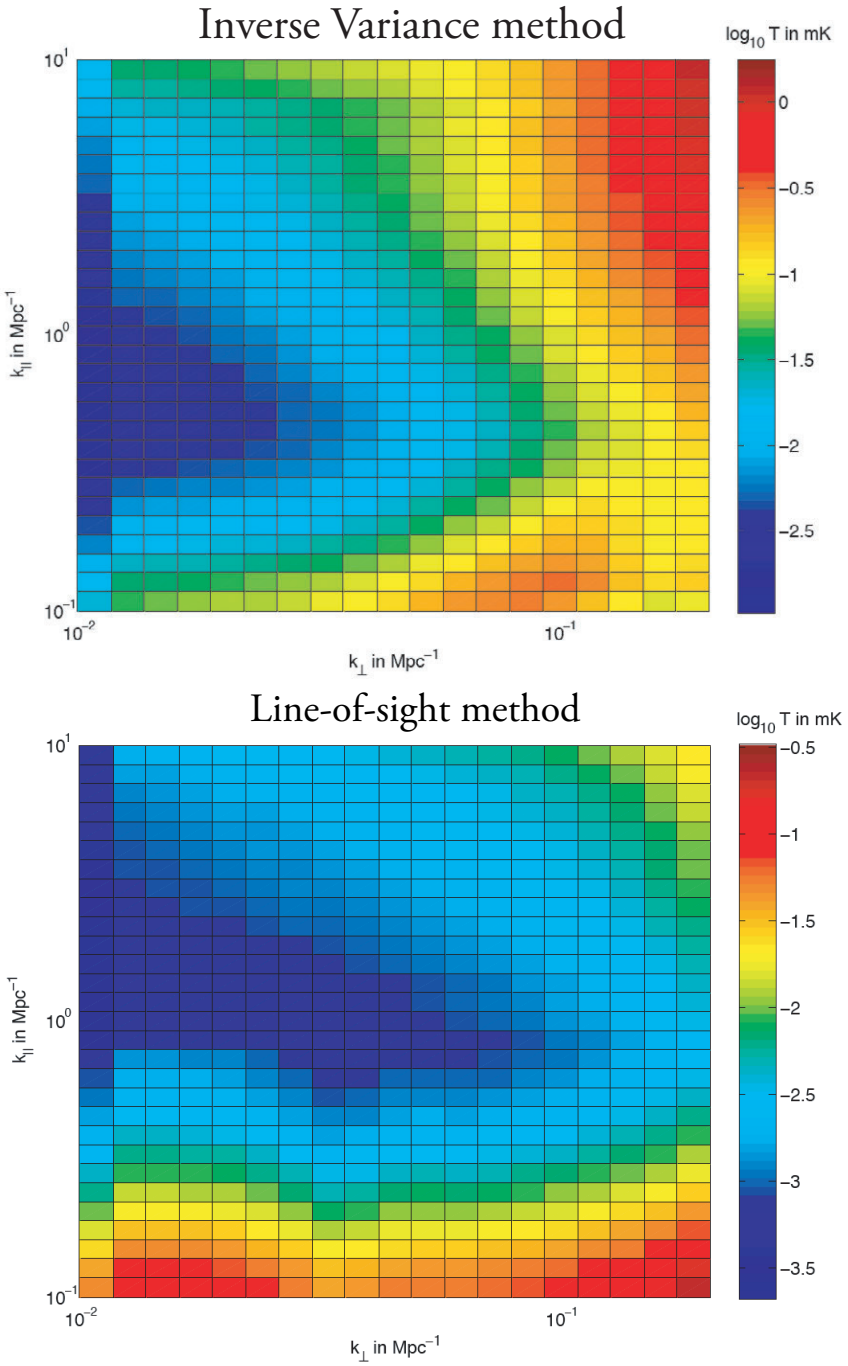

Figure 6-9: Bias term as a function of $k_\perp$ and $k_\parallel$ for the inverse variance method (Equation 6.6, top panel), and for the line-of-sight method (Equation 6.22, bottom panel). The LOS plot has been artificially normalized to match the inverse variance plot at values of medium $k_\perp$ and $k_\parallel$, where we expect the two techniques to be extremely similar.

We employ a one-dimenstional covariance of the form

$$C(\nu, \nu') = \delta(\nu - \nu') + NR(\nu, \nu'), \tag{6.42}$$

where we have written a continuous covariance instead of a discretized covariance matrix because we will be working in the continuous limit in this section in order to eliminate pixelization effects. The foregrounds are modeled by $R(\nu, \nu')$, N is a normalization constant to be determined later, and the noise is modeled by the delta function. Without loss of generality, we can select units such that the r.m.s. foreground covariance is equal to unity at all frequencies *i.e.* $R(\nu, \nu) = 1$ for all ν. This amounts to dividing the line-of-sight covariance $\Gamma(\nu, \nu')$ of Section 6.3 by a frequency-dependent normalization $\sigma(\nu)\sigma(\nu')$, where

$$\sigma(\nu) = \left(\frac{\nu}{\nu_*}\right)^{-\alpha_0 + \sigma_\alpha^2 \ln(\nu/\nu_*)}, \tag{6.43}$$

giving

$$R(\nu, \nu') = \exp\left[-\frac{\sigma_\alpha^2}{2}\left(\ln\nu - \ln\nu'\right)^2\right], \tag{6.44}$$

i.e. a Gaussian in logarithmic frequency. Since we are assuming a narrow frequency range, the logarithmic terms can be expanded about ν_*, giving

$$R(\nu, \nu') \approx R(\nu - \nu') = \exp\left[-\frac{(\nu - \nu')^2}{2\nu_c^2}\right], \tag{6.45}$$

where we have defined the *foreground coherence length*[18]:

$$\nu_c \equiv \frac{1}{\sigma_\alpha}\left(\frac{\nu_*}{\ln\nu_*}\right). \tag{6.46}$$

Note that the foreground contribution now depends on just the *difference* between ν and ν' *i.e.* it is translation invariant in frequency space. The same is true for the delta function noise term, which accurately captures the fact that the instrumental noise is uncorrelated between different frequencies. For simplicity, we model the noise

[18]While this definition of the foreground coherence length is what follows from Taylor expanding $R(\nu, \nu')$, we caution that Equation 6.46 should not be taken too literally as a way to compute ν_c. Doing so gives rather long coherence lengths, over which the linearized version of $R(\nu, \nu')$ becomes a bad approximation. Indeed, it is somewhat unphysical for ν_c to depend on ν_*, which after all was just an arbitrary frequency in the power-law foreground parameterizations of Section 6.3. In what follows, we will choose a fiducial value of $\nu_c = 0.5\,\mathrm{MHz}$ to "anchor" our toy model to the results of Section 6.4.

as being white in our chosen units[19]. Putting everything together, we see that our toy model is one that is defined by the covariance

$$C(\nu - \nu') = \delta(\nu - \nu') + \frac{\gamma}{\sqrt{2\pi\nu_c^2}} \exp\left[-\frac{(\nu - \nu')^2}{2\nu_c^2}\right], \tag{6.47}$$

where we have redefined our normalization constant so that the noise and foreground pieces are individually normalized to unity, allowing the constant γ to be interpreted as a foreground-to-noise ratio.

Let us now apply the inverse-variance power spectrum estimation method of Section 6.2.2 to our toy model. Interpreting $C(\nu - \nu')$ as an integral kernel, an inverse-variance weighting of the data corresponds to applying the inverse kernel $C^{-1}(\nu - \nu')$. By the convolution theorem, a translation invariant kernel acts multiplicatively when written in its dual Fourier basis[20]. This makes the computation of the inverse kernel straightforward — one simply computes the reciprocal of the Fourier transform of C. Defining η to be the Fourier dual of ν, the inverse kernel can be shown to take the form of a logistic curve in η^2:

$$\begin{aligned} \tilde{C}^{-1} &= \left[1 + \gamma \exp\left(-\frac{1}{2}\nu_c^2\eta^2\right)\right]^{-1} \\ &\approx \left[1 + \gamma \exp\left(-\frac{\nu_c^2 c^2 k_\parallel^2}{2\nu_* \nu_0 H_0^2 \Omega_m}\right)\right]^{-1}, \end{aligned} \tag{6.48}$$

where $\nu_0 = 1420\,\mathrm{MHz}$ is the rest frequency of the 21 cm line, and Ω_m, H_0, and c have their usual meanings. This kernel is plotted in Figures 6-10 and 6-11. In Figure 6-10, we keep the foreground coherence length ν_c fixed at 0.5 MHz (chosen to calibrate our toy model against the results of Section 6.4) and vary the foreground-to-noise ratio γ from 0 to 10^5, whereas in Figure 6-11 we keep γ fixed at 10^5 (typical of first-generation 21 cm tomography experiments with 1000 hrs of integration time) and vary ν_c. In both instances the fiducial experimental cases are plotted using solid black curves.

From the plots, ones sees that the action of the foreground cleaning kernel $\tilde{C}^{-1}$ is to suppress low-$k_\parallel$ modes in the final power spectrum estimate. Foreground subtraction

[19] This is of course not strictly correct, but suffices for the illustrative purposes of this toy model, especially given the narrow frequency ranges required for power spectrum estimation in 21 cm tomography.

[20] Strictly speaking, this is only true if the kernels are integrated from $-\infty$ to ∞, which is somewhat unphysical since $\nu > 0$ and moreover must be within the frequency range of one's instrument. However, since the foreground kernel decays quickly away from $\nu = \nu'$, we make this approximation for the purposes of our toy model

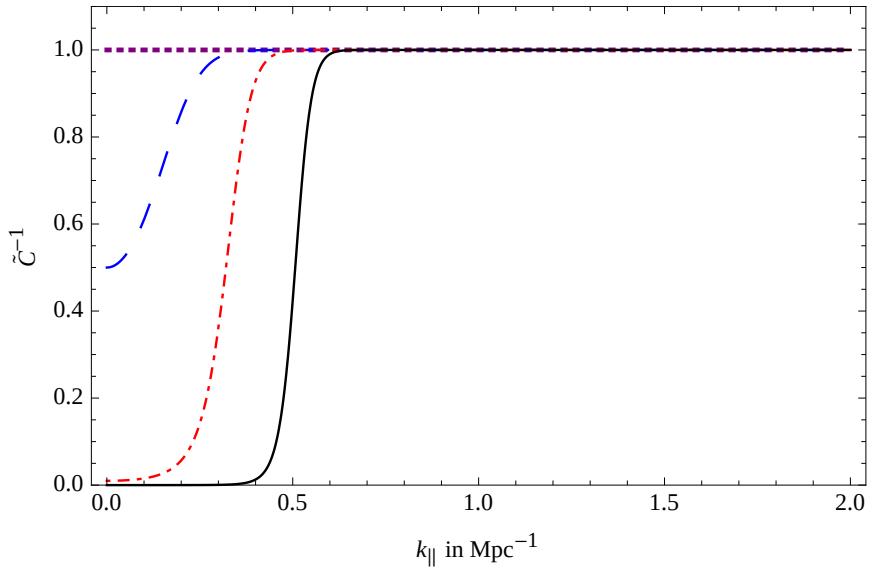

Figure 6-10: A plot of the foreground cleaning kernel $\tilde{C}^{-1}$ (Equation 6.48) for different values of the foreground-to-noise ratio γ, set at $\gamma = 0$ for the purple/dotted curve, at $\gamma = 1$ for the blue/dashed curve, at $\gamma = 100$ for the red/dot-dashed curve, and at $\gamma = 10^5$ for the black/solid curve. In all cases, the foreground coherence length $\nu_c = 0.5\,\mathrm{MHz}$. The black/solid curve is intended to be representative of a first-generation 21 cm tomography experiment.

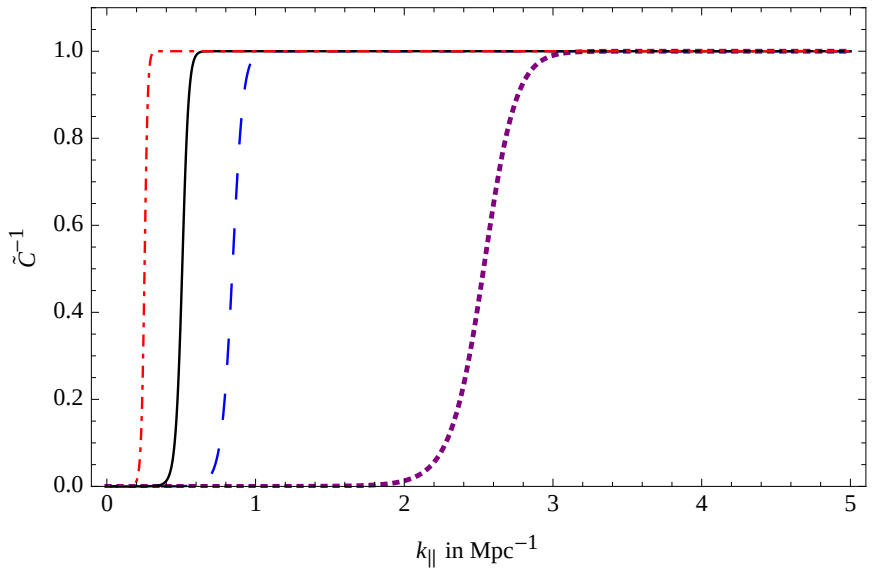

Figure 6-11: A plot of the foreground cleaning kernel $\tilde{C}^{-1}$ (Equation 6.48) for different values of the foreground coherence length ν_c, with $\nu_c = 0.1\,\mathrm{MHz}$ for the purple/dotted curve, at $\nu_c = 0.3\,\mathrm{MHz}$ for the blue/dashed curve, at $\nu_c = 0.5\,\mathrm{MHz}$ for the black/solid curve, and at $\nu_c = 1.0\,\mathrm{MHz}$ for the red/dot-dashed curve. In all cases the foreground-to-noise ratio γ is fixed at 10^5. The black/solid curve is intended to be representative of a first-generation 21 cm tomography experiment.

in the inverse variance scheme thus amounts to applying a high-pass filter in the frequency/line-of-sight direction. The highest $k_\parallel$ modes are left untouched, whereas the lowest $k_\parallel$ modes are suppressed by a factor of $1+\gamma$. This can be seen in Figure 6-10, where there is no suppression when $\gamma = 0$. Physically, this is because with $\gamma = 0$ there are no foregrounds and one only needs to deal with instrumental noise. Since noise contributions from different frequencies are uncorrelated, the effect of noise is best mitigated through simple averaging and binning, which is accomplished by the $\mathbf{C}_{,\alpha}$ piece of Equation 6.16 and not by the inverse variance weighting. The $\tilde{C}^{-1}$ kernel is thus flat when $\gamma = 0$. At high values of γ, the low-$k_\parallel$ modes are essentially taken out completely because those are precisely the modes that are heavily contaminated by foregrounds. In intermediate regimes, there is partial suppression of the low-$k_\parallel$ modes to mitigate the foregrounds, but a complete suppression is unnecessary because the foregrounds are relatively low in amplitude; thus, useful cosmological information may still be extracted from these modes.

To understand the dependence of $\tilde{C}^{-1}$ on the coherence length ν_c, it is useful to define a cutoff scale $k_\parallel^{cut}$. We define this scale to be the scale at which the $k_\parallel$ modes are "half-suppressed", with a weighting that is the arithmetic mean of $(1+\gamma)^{-1}$ (the weighting at $k_\parallel = 0$) and 1 (the weighting at $k_\parallel = \infty$). A little bit of algebra reveals that

$$k_\parallel^{cut} = \left(\frac{c}{H_0}\right)^{-1} \left(\frac{\nu_* \nu_0}{\nu_c^2}\right)^{\frac{1}{2}} \Omega_m^{1/2} \sqrt{2\ln(2+\gamma)}. \tag{6.49}$$

For $k_\parallel \ll k_\parallel^{cut}$, the modes are severely contaminated by foregrounds and are largely thrown out, whereas for $k_\parallel \gg k_\parallel^{cut}$ the foregrounds are minimal and essentially no cleaning is required. Equation 6.49 and Figure 6-10 show that as foregrounds become increasingly important (*i.e.* as γ grows), $k_\parallel^{cut}$ increases. This is because with higher foregrounds, one is forced to clean to higher $k_\parallel$ modes before estimating the power spectrum. Encouragingly, however, $k_\parallel^{cut}$ depends only *logarithmically* on γ, which means that as the foregrounds become increasingly important, only a few more modes need to be thrown out. For instance, as the foreground-to-noise ratio γ goes from 1 to 10^5, $k_\parallel^{cut}$ increases by only a factor of ~ 3. In Figure 6-11 and Equation 6.49 we see that as ν_c increases, $k_\parallel^{cut}$ decreases. Physically, this is because a high ν_c implies foregrounds that are very spectrally coherent, which allows the foreground cleaning to be confined to a handful of large scale (small $k_\parallel$) modes.

As noted in Section 6.2, subtracting foregrounds using the inverse variance method is a two-step process. The first step is a linear one that acts on the data itself, and we have seen in this section that this first step can be roughly understood as a high-pass

filter in the frequency direction. The second step involves foreground mitigation in the power spectrum itself and amounts to the subtraction of the residual noise and foreground bias term $b_\alpha = \mathrm{tr}[\mathbf{C}_{,\alpha}\mathbf{C}^{-1}]$. This term can be approximated in our toy model by replacing the discrete trace by a continuous integral:

$$\begin{aligned} b_\alpha &= \mathrm{tr}[\mathbf{C}_{,\alpha}\mathbf{C}^{-1}] = \sum_{i,j}(\mathbf{C}_{,\alpha})_{ij}\mathbf{C}^{-1}_{ji} \\ &\approx \int C_{,\alpha}(\nu_i,\nu_j)C^{-1}(\nu_j,\nu_i)d\nu_i d\nu_j \\ &\propto \int e^{i\eta_\alpha(\nu_i-\nu_j)}C^{-1}(\nu_j,\nu_i)d\nu_i d\nu_j = \tilde{C}^{-1}(\eta_\alpha), \end{aligned} \tag{6.50}$$

where we used the fact that in one dimension, $C_{,\alpha}(\nu_i,\nu_j) \propto e^{i\eta_\alpha(\nu_i-\nu_j)}$. With the final integral being precisely the Fourier transform of C^{-1}, we see that the bias term that needs to be subtracted off is proportional to the $\tilde{C}^{-1}$ function given by Equation 6.48. Intuitively, this can be thought of as a subtraction of the statistically expected foreground residuals in $k_\parallel$-space. From the form of Equation 6.48, we can see that at low $k_\parallel$ — where the linear foreground subtraction step acts more aggressively — the residuals are expected to be small, and the extra foregrounds that need to be removed in $k_\parallel$-space are small.

Using similar manipulations, one can see that the bias term acts in a similar way for the LOS polynomial subtraction. From Equation 6.22 we have $b_\alpha = \mathrm{tr}[\mathbf{C}_{,\alpha}\mathbf{DCD}]$ for the LOS subtraction, where recall that $\mathbf{D}$ was the projection matrix responsible for projecting out the lowest order polynomials in the frequency direction (which hopefully contain most of the foregrounds). Comparing this to Equation 6.50 tells us that one simply needs to substitute $\mathbf{C}^{-1}$ for $\mathbf{DCD}$. This combination can be rewritten as

$$\mathbf{DCD} = \langle \mathbf{D}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t\mathbf{D}^t\rangle, \tag{6.51}$$

which is the covariance of the residual foregrounds after LOS subtraction. Conceptually, then, the residual bias term works in the same way as it did in the inverse variance subtraction — one is simply subtracting off the statistically expected foreground residuals in power spectrum space.

Finally, we can make a similar estimate of the Fisher information for the inverse

variance method. Starting from Equation 6.13, we have

$$\begin{aligned}
\mathbf{F}_{\alpha\beta} &= \frac{1}{2}\mathbf{tr}\left[\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta}\mathbf{C}^{-1}\right] \\
&\approx \frac{1}{2}\int C_{,\alpha}(\nu_i,\nu_j)C^{-1}(\nu_j,\nu_k)\times \\
&\qquad C_{,\beta}(\nu_k,\nu_m)C^{-1}(\nu_m,\nu_i)d\nu_i d\nu_j d\nu_k d\nu_m \\
&\propto \left|\int e^{-i\eta_\alpha\nu_j}C^{-1}(\nu_j,\nu_k)e^{i\eta_\beta\nu_k}d\nu_j d\nu_k\right|^2 \\
&\propto \left|\tilde{C}^{-1}(\eta_\alpha)\right|^2 \propto \frac{\delta_{\alpha\beta}}{\left[1+\gamma\exp\left(-\frac{1}{2}\nu_c^2\eta_\alpha\right)\right]^2} \\
&\approx \delta_{\alpha\beta}\left[1+\gamma\exp\left(-\frac{\nu_c^2c^2k_\parallel^2}{2\nu_*\nu_0H_0^2\Omega_m}\right)\right]^{-2},
\end{aligned} \tag{6.52}$$

which looks exactly like the curves plotted in Figure 6-10, but squared. This shows that the Fisher information content is much higher at high values of $k_\parallel$ than at low values of $k_\parallel$. In other words, after foreground subtraction removes power from large scales (small $k_\parallel$) there is little information left there and the error bars are large. This is in accordance with the middle panel of Figure 6-8, where the expected errors are seen to decrease as one moves away from the lowest $k_\parallel$ values. Note, however, that the finite spectral resolution effects that we saw in the high $k_\parallel$ regions of Figure 6-8 are *not* captured by Equation 6.52. This is to be expected, since here we have a toy model whose covariance is a *continuous* function of frequency.

The results of this section suggest that the inverse variance foreground subtraction can be qualitatively thought of as a scheme where the data are high-pass filtered in the frequency direction. This can be done by applying a multiplicative filter in $k_\parallel$ space, an example of which is given by Equation 6.48. In fact, such line-of-sight filtering can be considered a computationally cheap approximation to the full inverse variance method, requiring just a simple multiplication instead of the inversion of an $n_{pix}\times n_{pix}$ covariance matrix. The results should be almost as good, since we saw from the bottom panel of Figure 6-2 that even in the full method, the algorithm "chooses" to perform foreground subtraction almost exclusively in the line-of-sight direction.

Filtering in the line-of-sight direction is an approach that is extremely similar to one that was proposed in Petrovic & Oh (2011). There, the authors suggest cleaning foregrounds by excluding all modes with $k_\parallel < k_\parallel^{crit}$, where $k_\parallel^{crit}$ is some critical line-of-sight mode below which foregrounds are expected to dominate. This corresponds to applying a step-function filter, and in Petrovic & Oh (2011) it was found that

this dramatically reduces systematic biases in the spherically averaged 3D power spectrum. However, the authors also caution that some foregrounds may have some small (but nonzero) high $k_\parallel$ component that will not be alleviated by a sharp filter. This problem is solved by instead using the filter $\tilde{C}^{-1}(k_\parallel)$ given in Equation 6.48. From Figures 6-10 and 6-11 one sees that the filter is qualitatively similar to a step function, but the fact that the filter is itself dictated by a foreground model allows it to enact slight suppressions of foregrounds even at high $k_\parallel$ modes.

6.6 Discussion and Conclusions

In this chapter, we have presented a unified matrix-based formalism for 21 cm foreground removal and power spectrum estimation. Using this framework, we compared existing line-of-sight (LOS) polynomial foreground subtraction schemes with a proposed inverse variance method, quantifying their errors and showing how to eliminate their biases. We now review our basic results, and follow with a discussion of their general applicability before summarizing our conclusions.

6.6.1 Basic Results

Through the numerical case study performed in Section 6.4, we have established a number of qualitative results:

1. LOS polynomial foreground subtraction is non-optimal, and gives rise to larger error bars in final power spectrum than inverse-variance subtraction does. This is mostly due to the fact that the LOS method projects out low-order polynomial modes, destroying information. The inverse variance method, on the other hand, preserves all modes, even if it does downweights some of them substantially. As a result, the error bars on the final power spectrum are larger for the LOS method (bottom panel of Figure 6-8) than for the inverse variance method (middle panel of Figure 6-8).

2. Traditional LOS polynomial subtraction methods contain residual noise and foreground biases in estimates of the power spectrum. Fortunately, this bias can be easily quantified and removed using Equation 6.6.

3. In the low $k_\parallel$ regions of the $k_\perp$-$k_\parallel$ plane, the LOS polynomial foreground cleaning gives rise to power spectrum estimates that are highly correlated between neighboring $k_\parallel$ bins, or equivalently, to window functions that are elongated

in the $k_\parallel$ direction. The LOS polynomial subtraction has no effect on the $k_\perp$ direction and exhibits the same qualitative behavior regardless of the value of $k_\perp$ since it does not use angular foreground information in its cleaning.

4. The behavior of the inverse variance window functions, on the other hand, depends strongly on the value of $k_\perp$:

   (a) In regions where neither $k_\perp$ nor $k_\parallel$ is small, the post-inverse variance subtraction power spectrum estimates are no more correlated than what would be expected from the finiteness of the survey geometry alone. In other words, the inverse variance window functions in this region are unchanged by the presence of foregrounds. This is due to the fact that foreground contamination was mild to begin with, resulting in a less aggressive foreground subtraction by the inverse variance weighting in order to keep power spectrum estimates as uncorrelated as possible.

   (b) In regions where both $k_\perp$ and $k_\parallel$ are small, the inverse variance window functions widen in both directions, but the effect in the $k_\perp$ direction is minimal. This indicates that inverse variance subtraction operates mainly in the radial direction, even though this is not an a priori mathematical constraint. Put another way, the nature of the foregrounds and the availability of high spectral resolution instruments makes it much more fruitful to leverage frequency information (rather than angular information) in performing foreground subtraction. This confirms intuitive expectations in the existing literature.

   (c) The inverse variance window functions show less $k_\parallel$ widening than the LOS polynomial window functions do, regardless of the location on the $k_\perp$-$k_\parallel$ plane. This suggests that with the inverse variance scheme, one should be able to push to lower values of $k_\parallel$ before correlations between neighboring $k_\parallel$-bins make the resulting power spectrum estimates scientifically unusable.

In summary, the inverse-variance method that we have presented has several advantages over the LOS polynomial subtraction methods proposed in the literature: they produce power spectrum measurements with smaller and less correlated error bars as well as narrower window functions.

6.6.2 Applicability of Results to general arrays

While the results presented in Section 6.4 were computed for a set of specific MWA-like array parameters and a specific foreground model, we emphasize that the formalism presented in Section 6.2 is completely general and can be applied to any 21 cm tomography experiment and any foreground model. In addition, as we will now argue, our qualitative conclusions from the last subsection should in fact hold quite generally.

Consider first an experiment's location, antenna layout, integration time, rotation synthesis, and thermal noise level. These factors may initially seem to form a high-dimensional parameter space that needs to be thoroughly explored, but from Section 6.2 we know that the only place where they enter our mathematical framework is Equation 6.36, which determined our noise covariance matrix $\mathbf{N}$. Their effects are thus degenerate with each other, and we saw in Section 6.5 that the resulting differences in the power spectrum estimation can be captured by a single parameter: the foreground-to-noise ratio γ.

Changing the frequency range of one's data cube has several effects, one of which is to alter the instrumental noise of the array, depending on the chosen frequency binning of the data. This, too, is captured by the foreground-to-noise parameter. More subtly, an increased frequency range changes the polynomial fit in the LOS scheme and affects foreground subtraction. While this can result in large changes to the final power spectrum, it was shown in Liu et al. (2009b) that for LOS subtraction to be effective in the first place, one *must* perform the fits only over narrow frequency ranges anyway. Moreover, as we have discussed above, cosmological evolution also imposes a constraint on the frequency range of any one data cube (one may of course separately analyze as many data cubes as necessary to cover the full data set). Thus, the small frequency range used in Section 6.4 is a good range to use, regardless of the full bandwidth of one's instrument. For the inverse variance method, our conclusions are even more general than for the LOS method. As we saw in the previous section, inverse variance subtraction can be thought of as a multiplicative filter in $k_\parallel$ space, and increasing the frequency range simply increases the resolution in $k_\parallel$-space without affecting the overall shape of the filter.

Another tunable parameter in our analysis is the flux cut S_{cut} above which the bright point sources can be considered identifiable and removable in an earlier stage of data analysis. In principle, varying S_{cut} can affect the results in a complicated way, but from Liu et al. (2009b), we know that the amplitude of foregrounds essentially scales with S_{cut}, since the foregrounds are dominated by the brightest sources in a population. Therefore, S_{cut} is simply yet another parameter that can be folded into

the foreground-to-noise ratio γ.

6.6.3 Future challenges

In this chapter, we have shown that inverse variance foreground subtraction is able to clean foregrounds more effectively than traditional LOS methods can, resulting in power spectrum estimates with smaller error bars. Perhaps the only disadvantage of the inverse variance method is its computational cost, since it requires the inversion of an $n_{pix} \times n_{pix}$ matrix, where $n_{pix} \sim n_\perp^2 n_\parallel$ is the total number of pixels in one's data cube. An LOS polynomial algorithm, on the other hand, only requires the inversion of an $n_\parallel \times n_\parallel$ matrix. Since matrix inversion naively scales as $\mathcal{O}(n^3)$, the difference is substantial.

Fortunately, by making a number of reasonable approximations one can speed up the method considerably. As long as the fields of view are relatively narrow, a series of basis changes and matrix prewhitening techniques allow the relevant computations to be performed using algorithms that scale as $\mathcal{O}(n_{pix} \log n_{pix})$, and such algorithms are now being tested numerically for 21 cm tomography (Dillon et al., 2012). Moreover, since the inverse variance method operates mostly in the radial direction, one can define an "effective" LOS algorithm that closely mimics the behavior of the full inverse variance subtraction. This was demonstrated in Section 6.5, where we developed an example of such an algorithm and saw that the inverse variance scheme can be thought of as a high-pass filter in the frequency direction, with the detailed properties of the filter dependent on noise and foreground characteristics.

To further improve the quality of foreground subtraction, it will be necessary to construct better foreground models. This is because inverse variance foreground cleaning is not blind, but instead depends on our empirical knowledge of foreground properties. It is important to note, however, that this is an advantage and not a disadvantage of the inverse variance method — whereas the traditional LOS methods (being blind) will not improve with better knowledge of the foregrounds, the inverse variance method is virtually guaranteed to get better as 21 cm tomography experiments begin to take science-quality data that are able to put tight constraints on our foreground models. Combining the inverse variance method with the improved measurements to be expected from the low frequency radio astronomy community should therefore bring us closer to fulfilling the potential of 21 cm tomography to improve our understanding of the Epoch of Reionization, the Dark Ages, and fundamental physics.

Chapter 7

Optimizing global 21 cm signal measurements with spectral and spatial information

7.1 Introduction

In previous chapters, we explored various techniques for extracting tomographic statistics (such as the power spectrum) from noisy measurements of the highly-redshifted 21 cm line. These tomographic statistics use data from fully three dimensional (angular plus spectral) maps and thus contain a wealth of information about our Universe, as we discussed in Chapter 1. However, this information comes at a high cost, requiring instruments that are expensive not only from a hardware perspective, but also from the standpoint of computational resources.

It is for this reason that in recent years there has been growing interest in a cheaper form of hydrogen cosmology, consisting of measuring the *global* 21 cm signal. The global signal is "global" in the sense that one is seeking to measure the 21 cm brightness temperature as a function of frequency (or redshift) after having spatially averaged over the entire sky. As we shall see in the next section, the global signal contains key observational signatures that can be used to constrain reionization. While global signal measurements are in principle not as informative as tomographic experiments are expected to be, they have the advantage that they are cheaper experiments to build. Without the need for spatial information, global signal experiments have typically been built (or envisioned) as single element instruments with no need for hardware correlators. In addition, while tomographic measurements contain cosmo-

logical information at fine frequency resolution, the global signal is theoretically expected to vary much more smoothly with redshift. As a result, larger bandwidths can be averaged into single frequency bins, giving higher signal-to-noise measurements. Global signal experiments thus have the potential to be quicker, cheaper alternatives for EoR science, and as we noted in Chapter 1, the EDGES experiment has already begun to report scientific results (Bowman et al., 2008; Bowman & Rogers, 2010).

In this chapter, we pose the following question—what is the statistically optimal way to measure the global 21 cm signal, both in terms of experimental design and data analysis? To answer this question, we will construct a systematic framework for quantifying error properties, essentially doing for global signal measurements what we did for power spectrum estimation in Chapter 6. Our framework includes the effect of foregrounds, foreground subtraction, and instrumental noise in a statistically consistent fashion.

The rest of this chapter is organized as follows. In Section 7.2 we construct models for both the cosmological signal as well as for the foregrounds and instrumental noise. In Section 7.3 we establish the formalism for extracting the cosmological signal amidst foregrounds and instrumental noise. Our formalism comes in two flavors, one adapted for experiments that have no angular sensitivity (thus automatically integrating over large portions of the sky), and one adapted for experiments with angular sensitivity (thus requiring a "manual" spatial averaging step in the data analysis pipeline). In Sections 7.4 and 7.5 we compute the measurement errors predicted by our formalism, and find that instruments with fine angular sensitivity will be able to mitigate foregrounds more easily, yielding better measurements of the cosmological signal. We summarize our conclusions in Section 7.6.

7.2 The global spectrum and what we're trying to measure

In this section we describe the global spectrum. We begin in Section 7.2.1 with a fiducial model for the cosmological signal. For the purposes of this thesis, we will consider the cosmological signal—what we want to measure—to be fixed. In Section 7.2.2 we construct models for various contaminants in the data, such as foregrounds and instrumental noise. In Section 7.3 we discuss the challenge that one faces when trying to extract the faint cosmological signal from the data.

7.2.1 Model of the signal

In Chapter 1, we saw that the brightness temperature T_b of redshifted 21 cm line is given by

$$T_b(\hat{\mathbf{r}},\nu) = (27\,\mathrm{mK})\bar{x}_H \left(\frac{T_s - T_\gamma}{T_s}\right)\left(\frac{1+z}{10}\right)^{1/2}(1+\delta_b)(1+\delta_x)\left[\frac{\partial v_r/\partial r}{(1+z)H(z)}\right]^{-1}, \quad (7.1)$$

where $\hat{\mathbf{r}}$ is a unit vector specifying a direction in the sky, ν is the measured frequency, $\bar{x}_H$ is the mean neutral hydrogen fraction, T_s is the spin temperature, T_γ is the CMB temperature, z is the redshift, $H(z)$ is the Hubble parameter, v_r is the velocity along the radial (line-of-sight) direction r, δ_b is the fractional baryon overdensity, and δ_x is the neutral fraction overdensity (Furlanetto et al., 2006). Since T_s, δ_b, and δ_x are spatially dependent quantities, the overall brightness temperature is a function of both frequency and angle on the sky. Tomographic measurements seek to measure this full three-dimensional dependence. In a global signal measurement, on the other hand, one seeks to measure the spatially averaged isotropic component of the signal, given by $\bar{T}_b(\nu) \equiv \frac{1}{4\pi}\int T_b(\hat{\mathbf{r}},\nu)d\Omega$. Any anisotropies should therefore be formally thought of as a form of noise in our measurement (see Section 7.2.2).

In our calculations, we take the model of Pritchard & Loeb (2010) to be our fiducial model for the global cosmological signal. The model consists of the following ingredients:

- A spin temperature that is determined by couplings with:
  - The CMB temperature, via direct 21 cm photon absorption or emission.
  - The kinetic/gas temperature of hydrogen clouds, via collisional excitations and de-excitations of the hyperfine transition, as well as the Wouthysen-Field effect (Wouthuysen, 1952). The Wouthysen-Field effect couples the kinetic temperature to the spin temperature via Lyman alpha photons because a Ly-α photon can excite a neutral hydrogen atom from an $n = 1$ state to an $n = 2$ state, only to have the subsequent spontaneous emission be to a different hyperfine state than the one the atom was originally in.
  - The CMB temperature via the kinetic temperature as an intermediary. At high redshifts, CMB photons can Compton-scatter off free electrons that are left over from recombination. The free electrons can then collide with hydrogen gas, coupling the gas temperature to the CMB temperature.
- Further couplings from astrophysical sources:

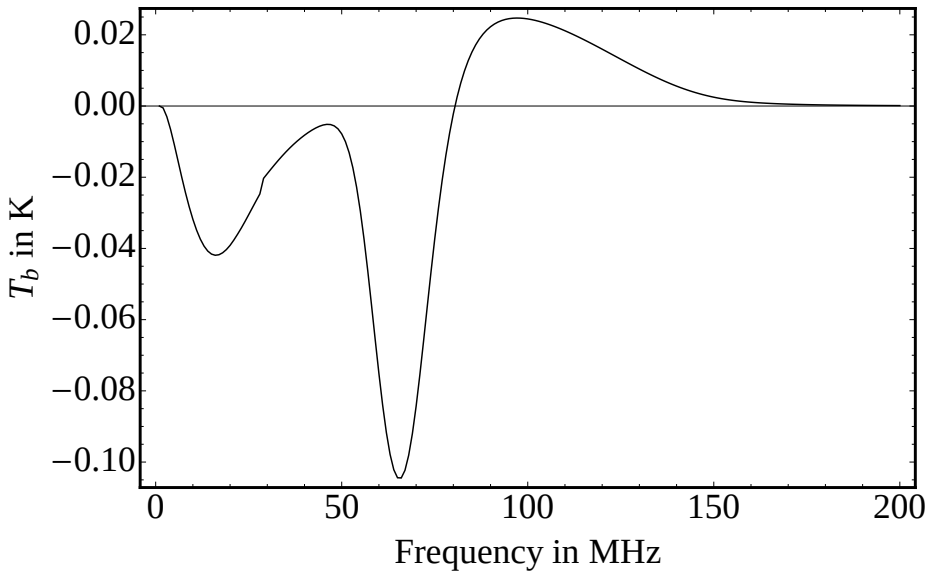

Figure 7-1: The theoretically expected global 21 cm signal.

- The first luminous sources provide a large number of Ly-α photons that affect the spin temperature via the Wouthysen-Field effect mentioned above. Ly-α photons may also heat the IGM, altering the gas temperature. As more and more luminous sources turn on, the Ly-α flux will of course also slowly reduce the mean neutral fraction $\bar{x}_H$ to zero.
- Early active galactic nuclei (AGN) provide X-ray photons which have a strong heating effect on the IGM in late reionization.

Putting all these ingredients together and integrating the relevant rate equations numerically gives the global signal shown in Figure 7-1.

Taken as a whole, the global spectrum succinctly describes much of the important astrophysics that governs reionization. At high redshifts (low frequencies), the 21 cm brightness temperature is zero because it is proportional to the spin temperature's *contrast* against the CMB. As discussed above, the Compton scattering of residual electrons off CMB photons has the effect of driving the spin temperature into equilibrium with the CMB temperature, so the temperature contrast and consequently the brightness temperature are zero. Eventually this coupling becomes negligible, and the spin temperature equilibrates with the gas temperature instead, which by now is cooler than the CMB, since $T_{gas} \propto a^{-2}$ while $T_{CMB} \propto a^{-1}$, where a is the scale factor. This causes the absorption trough at $\sim 10\,\text{MHz}$. As the gas further dilutes, collisional excitations and de-excitations of the hyperfine transition become rare, and the dominant coupling is that of direct absorption and emission of 21 cm

CMB photons, which drives the temperature contrast and T_b towards zero again.

As the first luminous sources turn on and provide Ly-α photons, the aforementioned Wouthysen-Field effect once again drives the spin temperature to the gas temperature, producing the second, deeper absorption trough at $\sim$ 70 MHz. As reionization proceeds, the first X-ray sources reheat the IGM, causing the gas temperature and spin temperature to exceed the CMB temperature. The 21 cm signal thus appears in emission, until reionization rids our Universe of most of its neutral hydrogen content, and the signal vanishes.

Given the large number of physical processes at play during reionization, it is perhaps unsurprising that the quantitative details of Figure 7-1 depend sensitively on a large number of astrophysical parameters. This can be turned around, of course, and a good measurement of the cosmological signal can be used to constrain these parameters. For instance, the second absorption trough "starts" when there are sufficient Ly-α photons for the Wouthysen-Field effect to be important and "ends" when X-ray heating becomes substantial. Its depth and shape therefore allow one to place constraints on the Ly-α and X-ray emissivities from the first luminous objects. This second trough is thought to be particularly promising from an observational standpoint, given its large amplitude. As a result, we focus on a 30 to 70 MHz window for the rest of this chapter.

7.2.2 Generalized foreground and noise model.

We now construct our noise model. We define the noise to be any contribution to the measurement that is not the global 21 cm signal as described in the previous section. Our noise thus contains more than just instrumental noise and foregrounds, and in fact contains a contribution from the *anisotropic* portion of the cosmological signal. In other words, the "signal" in a tomographic measurement (where one measures anisotropic fluctuations on various length scales) is an unwanted contaminant in our case, since we seek to measure the *global* signal (*i.e.* the monopole). Thus, if we imagine performing an experiment that images N_{pix} pixels on the sky over N_{freq} frequency channels, the noise contribution in various pixels at various frequencies can be grouped into a vector $\mathbf{n}$ of length $N_{pix}N_{freq}$ that is comprised of three contaminants:

$$\mathbf{n}_{\alpha i} \equiv \mathbf{n}^{fg}_{\alpha i} + \mathbf{n}^{inst}_{\alpha i} + \mathbf{n}^{s}_{\alpha i}, \tag{7.2}$$

where $\mathbf{n}^{fg}$, $\mathbf{n}^{inst}$, and $\mathbf{n}^{s}$ signify the foregrounds, instrumental noise, and anisotropic cosmological signal, respectively. Throughout this paper, we use Greek indices to sig-

nify the radial/frequency direction, and Latin indices to signify the spatial directions. Note that $\mathbf{n}$ is formally a quantity with a single index (*i.e.* a vector) even though we assign separate spatial and spectral indices to it for clarity. In the following subsections we discuss each contribution to the noise, with an eye towards how each can be mitigated or removed in a real measurement.

Foreground Model

Given that foregrounds are likely to be the largest contaminant in a measurement of the global signal, it is important to have a foreground model that is an accurate reflection of the actual contamination faced by an experiment, as a function of both position and frequency. Having such a model that describes the particular realization of foregrounds contaminating a certain measurement is crucial for optimizing the foreground removal process, as we shall see in Section 7.3. However, constructing such a model is difficult to do from first principles, and is much more difficult than what is typically done, which is to capture only the *statistical* behavior of the foregrounds (*e.g.* by measuring quantities such as the spatial average of a spectral index). It is thus likely that a full foreground model will have to be based at least partially on empirical data.

Constructing an empirically-based foreground model would be trivial if one simply had high angular resolution, full-sky survey data at the relevant frequencies — the model would simply be the survey itself. Unfortunately, full-sky measurements are currently lacking in the low frequency regime of interest to global signal experiments. Foreground models must therefore be constructed via interpolations and extrapolations from measurements that are incomplete both spatially and spectrally. One such effort is the *Global Sky Model* (GSM) of de Oliveira-Costa et al. (2008). In that study, the authors obtained foreground survey data at 11 different frequencies and formed a dimensionless frequency correlation matrix, given by

$$\widetilde{\mathbf{G}}_{\alpha\beta} \equiv \frac{\mathbf{G}^{GSM}_{\alpha\beta}}{\sqrt{\mathbf{G}^{GSM}_{\alpha\alpha}\mathbf{G}^{GSM}_{\beta\beta}}}, \tag{7.3}$$

where $\mathbf{G}^{GSM}_{\alpha\beta} \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_{\alpha i}\mathbf{g}_{\beta i}$, with N being the number of pixels in a spectrally well-sampled region of the sky, and $\mathbf{g}_{\alpha i}$ denoting the measured foregrounds in the i^{th} of the α^{th} frequency channel. By decomposing $\widetilde{\mathbf{G}}$ into principal components, the authors found that the spectral features of the foregrounds were dominated by the first three principal components, which could be used as spectral templates for constructing

empirically-based foreground models. The GSM approach was found to be accurate to $\sim 10\%$.

Being relatively quick and accurate, the GSM has been used as a fiducial foreground model in most studies of the global 21 cm signal to date. However, this may be insufficient for two reasons. First, as mentioned above, the GSM approach predicts the magnitude of foregrounds by forming various linear combinations of three principal component (*i.e.* eigenmode) templates. Thus, if the GSM is considered the "true" foreground contribution in our models, it becomes formally possible to achieve perfect foreground removal simply by projecting out three spectral modes from the data, which is too optimistic an assumption. The other weakness of the GSM is that it does not include bright point sources, which are expected to be quite numerous at the sensitivities of most 21 cm experiments.

In this paper, we use the GSM as a starting point, but add our own extensions to deal with the aforementioned shortcomings. The extensions come partly from the phenomenological model of Liu & Tegmark (2012), which in our notation can be summarized by writing down a matrix $\mathbf{G}^{LT}$ that is analogous to $\mathbf{G}^{GSM}$ defined above:

$$\mathbf{G}^{LT} \equiv \mathbf{G}^{ps} + \mathbf{G}^{sync} + \mathbf{G}^{ff}, \tag{7.4}$$

where $\mathbf{G}^{ps}$, $\mathbf{G}^{sync}$, and $\mathbf{G}^{ff}$ refer to foreground contributions from unresolved extragalactic point sources, Galactic synchrotron radiation, and Galactic free-free emission, respectively. Each contribution takes the generic form

$$\mathbf{G}_{\eta\beta} = A^2 \left(\frac{\nu_\eta \nu_\beta}{\nu_*^2} \right)^{-\alpha + \frac{\Delta\alpha^2}{2} \ln\left(\frac{\nu_\alpha \nu_\beta}{\nu_*^2}\right)}, \tag{7.5}$$

where $A = 335.4\,\mathrm{K}$, $\alpha = 2.8$, and $\Delta\alpha = 0.1$ for the synchrotron contribution, $A = 70.8\,\mathrm{K}$, $\alpha = 2.5$, and $\Delta\alpha = 0.5$ for the unresolved point source contribution, and $A = 33.5\,\mathrm{K}$, $\alpha = 2.15$, and $\Delta\alpha = 0.01$ for the free-free contribution. Our strategy is to perform a principal component decomposition of this model, and to use its higher order principal components to complement the three components provided by the GSM, completing the basis. We check that this is a sensible strategy by forming $\widetilde{\mathbf{G}}$ (Equation 7.3) for both the GSM and the phenomenological model. Each index specifies one of 70 frequency channels, each with bandwidth 1 MHz spanning the 30 MHz to 100 MHz frequency range (relevant to early reionization observations). We then compute the eigenvalue spectrum of both models, and the result is shown in Figure 7-2. The GSM, built from three principal components, contains just three

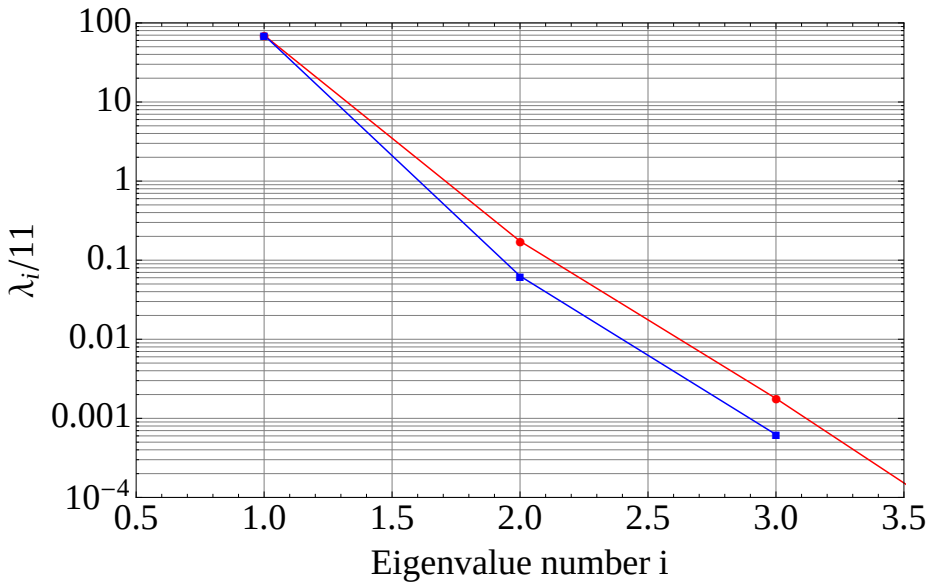

Figure 7-2: Comparison with the global sky model (de Oliveira-Costa et al., 2008). Our eigenvalues are in red, while those of the GSM are in blue.

eigenvalues. The first three eigenvalues of the phenomenological model agree with these values quite well, with the phenomenological model slightly more conservative. It is thus reasonable to use the higher eigenmodes of the phenomenological model to "complete" the foreground model spectrally. From Chapter 3 we expect this to be necessary because there are more than three eigenvalues above the thermal noise limit. Spatially, we form an angular model by averaging the GSM maps over frequency, and give all of the higher eigenmodes this angular dependence. While in general we expect the spatial structure to be different from eigenmode to eigenmode, our simple assumption is a conservative one, since any additional variations in spatial structure provide extra information which can be used to aid foreground subtraction.

Next, we add bright point sources to our model. The brightness of these sources are distributed according to the source count function

$$\frac{dn}{dS} = (4\,\text{sources mJy}^{-1}\,\text{sr}^{-1}) \left(\frac{S}{880\,\text{mJy}}\right)^{-1.75}, \tag{7.6}$$

and we give each source a spectrum of

$$S(\nu) = S_0 \left(\frac{\nu}{150\,\text{MHz}}\right)^{-\alpha}, \tag{7.7}$$

where the S_0 is the brightness drawn from the source count function, and α is the spectral index, drawn from a Gaussian distribution with mean 0.5 and a standard

deviation of 0.25 (Liu et al., 2009b). Spatially, we distribute the bright point sources randomly across the sky. This is in principle an unrealistic assumption, since point sources are known to be clustered. However, a uniform distribution suffices for our purposes because it is a conservative assumption — any spatial clustering would constitute an additional foreground signature to aid with foreground removal.

Putting all these ingredients together, the result is a series of foreground maps like that shown in Figure 7-3, one at every frequency. This collection of maps forms the foreground model vector $\mathbf{n}^{fg}$ of Equation 7.2. Now, while our foreground model is based on empirical constraints, it will naturally have a finite level of accuracy. A robust method for foreground subtraction thus needs to be able to account for possible errors in the error model. To set the stage for this such a method, we define the mean of our foreground model to be the "best guess" model described above, *i.e.*

$$\mathbf{m}^{fg}_{\alpha i} \equiv \langle \mathbf{n}^{fg}_{\alpha i} \rangle. \tag{7.8}$$

The covariance of the foregrounds is defined as the error in our foreground model:

$$\mathbf{N}^{fg}_{\alpha i \beta j} \equiv \langle \mathbf{n}^{fg}_{\alpha i} \mathbf{n}^{fg}_{\beta j} \rangle - \langle \mathbf{n}^{fg}_{\alpha i} \rangle \langle \mathbf{n}^{fg}_{\beta i} \rangle = \varepsilon^2 \mathbf{m}^{fg}_{\alpha i} \mathbf{m}^{fg}_{\beta j} \mathbf{R}_{ij} \mathbf{Q}_{\alpha\beta}, \tag{7.9}$$

where we have assumed that the error in our foreground model is proportional to the model $\mathbf{m}^{fg}$ itself, with a proportionality constant ε between 0 and 1. In other words, we assume that there is some constant percentage error in every pixel and frequency of our foreground model. The matrices $\mathbf{R}$ and $\mathbf{Q}$ encode the spatial and spectral correlations in our foreground model errors respectively, and are constructed to have 1's along their diagonals. If we imagine that our foreground model was based on a foreground survey with some finite angular resolution, the matrix would roughly consist of a near-diagonal band of $\lesssim 1$'s and zeroes elsewhere, with the width of the band corresponding to the width of the survey instrument's beam. Note that this width may or may not be the same as that of the instrument used for the actual global signal measurement.

For computational convenience, we will often treat the spatial correlations in the continuous limit. The matrix $\mathbf{R}$ is then chosen to correspond to a continuous kernel

$$R(\widehat{\mathbf{r}}, \widehat{\mathbf{r}}') \equiv \frac{\exp(\sigma^{-2} \widehat{\mathbf{r}} \cdot \widehat{\mathbf{r}}')}{4\pi\sigma^2 \sinh(\sigma^{-2})}, \tag{7.10}$$

which is known as a Fisher function, the analog of a Gaussian on a sphere. For the spectral correlation matrix $\mathbf{Q}$, we imagine that our foreground model was constructed

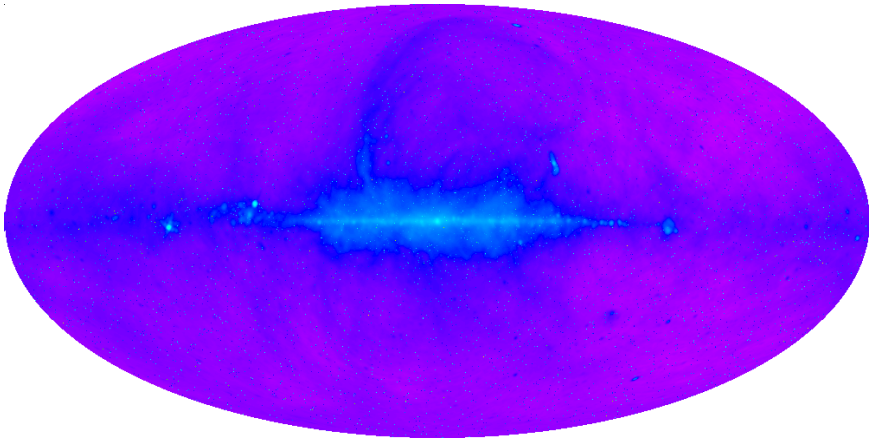

Figure 7-3: Foreground template at 30MHz.

by fitting to the spectrum of every pixel a power law of the form $t(\nu) = A(\nu/\nu_*)^{-\alpha}$, where A is a normalization constant that will later cancel out, $\nu_* = 150\,\mathrm{MHz}$ is a reference frequency, and α is a spectral index. This spectral index will have some error associated with it, due in part to uncertainties in the foreground survey and in part to the fact that foreground spectra are not perfect power laws. We model this error as being Gaussian distributed, so that the probability distribution of spectral indices is given by

$$p(\alpha) = \frac{1}{2\pi\sigma_\alpha^2} \exp\left[-\frac{1}{2}\frac{(\alpha-\alpha_0)^2}{\sigma_\alpha^2}\right], \tag{7.11}$$

where α_0 is a fiducial spectral index for typical foregrounds, which in this paper we take to be 2.5. With this, the mean spectral fit to our foreground survey is

$$m^{survey}(\nu) = A\int\left(\frac{\nu}{\nu_*}\right)^{-\alpha} p(\alpha)d\alpha. \tag{7.12}$$

Sampling this function at a discrete set of frequencies corresponding to the frequency channels of our global signal experiment, we can form a mean vector $\mathbf{m}^{survey}$. A covariance matrix $\mathbf{C}^{survey}$ of the power law fits is then given by

$$\mathbf{C}^{survey} \equiv \langle \mathbf{t}\mathbf{t}^t \rangle - \mathbf{m}^{survey}\mathbf{m}^{survey}, \tag{7.13}$$

where $\mathbf{t}$ is a discretized version of $t(\nu)$, and

$$\langle \mathbf{t}\mathbf{t}^t \rangle_{\beta\eta} = A^2\int\left(\frac{\nu_\beta\nu_\eta}{\nu_*^2}\right)^{-\alpha} p(\alpha)d\alpha. \tag{7.14}$$

Finally, we take the covariance $\mathbf{C}^{survey}$ and insert it into the left hand side of Equation

7.3 (just as we did with $\mathbf{G}_{GSM}$ earlier) to form our spectral correlation matrix $\mathbf{Q}$. Note that the normalization constant A cancels out in the process, as we claimed.

Instrumental Noise

We consider the instrumental noise $\mathbf{n}_{i\alpha}^{inst}$ in every pixel and every frequency channel to be uncorrelated. Additionally, we make the assumption that there are no systematic instrumental effects, so that $\langle \mathbf{n}_{i\alpha}^{inst} \rangle = 0$. What remains is the random contribution to the noise. Assuming a sky-noise dominated instrument, the amplitude of this contribution is given by the radiometer equation. In our notation, this gives rise to covariance of the form

$$\mathbf{N}_{\alpha i \beta j}^{inst} = \langle \mathbf{n}_{\alpha i}^{inst} \mathbf{n}_{\beta j}^{inst} \rangle = \frac{\mathbf{m}_{\alpha i}^{fg} \mathbf{m}_{\beta j}^{fg} \delta_{ij} \delta_{\alpha\beta}}{\Delta t \Delta \nu}, \tag{7.15}$$

where Δt is the integration time per pixel and $\Delta \nu$ is the channel width.

Cosmological Anisotropy Noise

In measuring the global signal, we are measuring the monopole contribution to the sky. As mentioned above, any anisotropic contribution to the cosmological power is therefore a noise contribution as far as a global signal experiment is concerned. By construction, these non-monopole contributions have a zero mean after spatially averaging over the sky, and thus do not result in a systematic bias to a measurement of the global signal. They do, however, have a non-zero variance, and therefore contribute to the error bars.

Despite its *a priori* existence, we can safely ignore cosmological anisotropy noise because it is negligible. Through a combination of analytic theory (Lewis & Challinor, 2007) and simulation analysis (Bittner & Loeb, 2011), the cosmological anisotropies have been shown to be negligible on scales larger than $\sim 1^\circ$ to 2°, which is a regime that we remain in for this chapter.

Generalized noise model summary

In the subsections above, we have outlined the various contributions to the generalized noise that plagues any measurement of the global signal. Of these contributions, only foregrounds have a non-zero mean, so the mean of our generalized noise model is just that of the foregrounds:

$$\mathbf{m}_{\alpha i} \equiv \langle \mathbf{n} \rangle = \mathbf{m}_{\alpha i}^{fg}. \tag{7.16}$$

Foregrounds therefore have a special status amongst the different components of our generalized model, for they are the only contribution with the potential to cause a systematic bias in our global signal measurement. The other contributions appear only in the total noise covariance, taken to be the sum of the foreground and instrumental noise covariances:

$$\mathbf{N}_{\alpha i \beta j} \equiv \mathbf{N}^{fg}_{\alpha i \beta j} + \mathbf{N}^{inst}_{\alpha i \beta j}, \tag{7.17}$$

where as noted above, we are neglecting the cosmological anisotropy noise.

In the foreground subtraction/data analysis scheme that we describe in Section 7.3, we will think of the mean $\mathbf{m}$ as a foreground template that is used to perform a first subtraction. However, there will inevitably be errors in our templates, and thus our scheme also takes into account the covariance $\mathbf{N}$ of our model. In our formalism, the mean term therefore represents our best guess as to what the foreground contamination is, and the covariance quantifies the uncertainty in our guess. We note that this is quite different from many previous approaches in the literature, where either the foreground modeling error is ignored, or the mean is taken to be zero and the covariance is formed by taking the ensemble average of the outer product of the foreground template error. The former approach is clearly unrealistic, while the latter approach has a number of shortcomings. For example, it is difficult to compute the necessary ensemble average, since foregrounds are difficult to model from first principles, and empirically the only sample that we have for taking this average is our Galaxy. As a solution to this, ensemble averages are often replaced by spatial (*i.e.* angular) averages. But this is unsatisfactory for our purposes, since in Section 7.3.3 we will be using the angular structure of foregrounds to aid with foreground subtraction, and this is impossible if the information has already been averaged out. Even if an ensemble average could somehow be taken (perhaps by running a large suite of radiative transfer foreground simulations), a foreground subtraction scheme that involved minimizing the resulting variance would be non-optimal for two reasons. First, in such a scheme one would be guarding against foreground power from a "typical" galaxy, which is irrelevant — all that matters to an experiment are the foregrounds that are seen in our Galaxy, even if it is atypical. In addition, foregrounds are not Gaussian distributed, and thus a minimization of the variance is not necessarily optimal.

Our approach — taking the mean to be an empirical foreground template and the covariance to be the errors in this template — solves these problems. Since the covariance arises from measurement errors (which can usually be modeled to an adequate accuracy), taking the ensemble average is no longer a problem. And

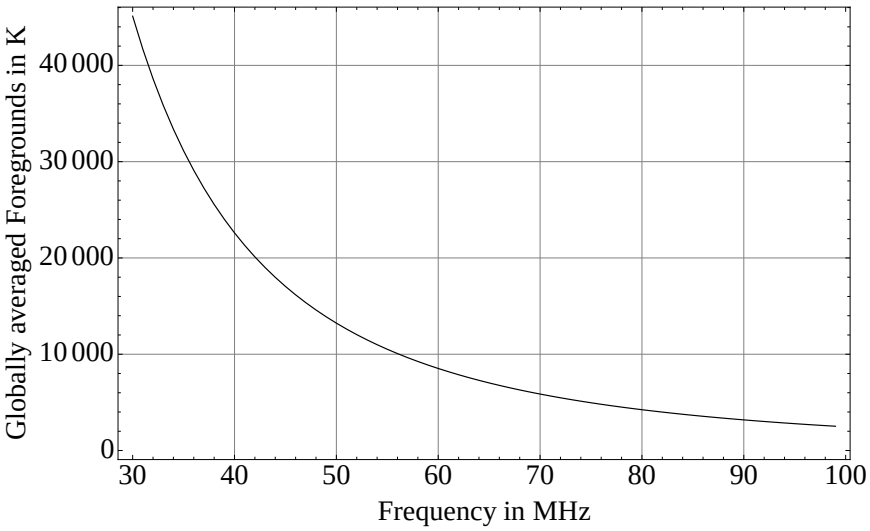

Figure 7-4: Angularly averaged foregrounds

with the mean term being a template for the foregrounds as seen by our experiment, our foreground model is tailored to our Galaxy, even if our Galaxy happens to be atypical. Finally, while the foregrounds themselves are certainly not Gaussian, it is a much better approximation to say that the uncertainties in our model are Gaussian, at least if the uncertainties are relatively small. Constructing our foreground model in this way thus allows us to take advantage of the optimal data analysis techniques that we introduce in Section 7.3.

7.2.3 Why it's hard

Before we proceed to describe how the global 21 cm signal can be optimally extracted, we pause to describe the challenges ahead. As an initial "straightforward" approach, one can imagine measuring the global 21 cm signal by taking a simple spatial average of a measured sky. The corresponding foreground contamination would be obtained by spatially averaging our model, which yields the spectrum shown in Figure 7-4. A quick comparison between the scales on Figure 7-1 (showing the theoretical signal), and the foreground spectrum underscores the magnitude of the problem. In addition, both the foreground contamination and the theoretical signal are smooth as a function of frequency, making it difficult to use proposed techniques for foreground subtraction in tomographic maps, where the cosmological signal is assumed to vary much more rapidly as a function of frequency than the foregrounds. It is therefore crucial that optimal foreground cleaning methods are employed in the analysis, and in the following section we derive such methods.

7.3 How to measure the global signal

In this section, we provide a general mathematical framework for estimating the global signal from data. Section 7.3.1 establishes general notation. Section 7.3.2 considers data analysis algorithms that use only spectral information, and we argue that such methods are unable to robustly deal with errors in one's foreground models. The method presented in Section 7.3.3, on the other hand, guards against such errors by making use of spatial information as well as spectral information, making it our preferred method for the rest of the paper.

7.3.1 The general mathematical framework

We begin with our measurement equation, which is given by

$$\mathbf{y} = \mathbf{A}\mathbf{x}_s + \mathbf{n}, \tag{7.18}$$

where $\mathbf{y}$ is a vector of length $N_{vox} \equiv N_{pix}N_{freq}$ containing the measurement, $\mathbf{n}$ is the generalized noise contribution of Section 7.2.2, $\mathbf{x}_s$ is a vector of length N_{freq} containing the global signal that we wish to measure, and $\mathbf{A}$ is a vertical stack of N_{pix} identity matrices of size $N_{freq} \times N_{freq}$. The effect of multiplying the global signal by $\mathbf{A}$ is to copy the theoretical spectrum to every pixel on the sky before the noise and foregrounds are added. The term $\mathbf{A}\mathbf{x}_s$ therefore represents a data ball[1] that contains ideal (noiseless and foregroundless) data that depends only on the radial distance from the center, and not on the direction. To every voxel of this data volume the combined noise and foreground contribution $\mathbf{n}$ is added, giving the set of measured voxel temperatures that comprise $\mathbf{y}$.

Having obtained $\mathbf{y}$ from our experiment, we wish to construct an estimator for $\mathbf{x}_s$. As a general form for the estimator $\widehat{\mathbf{x}}_\mathbf{s}$, we write

$$\widehat{\mathbf{x}}_s = \mathbf{M}\mathbf{A}^t\mathbf{H}(\mathbf{y} - \mathbf{b}), \tag{7.19}$$

where $\mathbf{b}$ is a vector of length N_{vox}, $\mathbf{H}$ is a matrix of size $N_{pix}N_{freq} \times N_{pix}N_{freq}$, and $\mathbf{M}$ is a matrix of size $N_{freq} \times N_{freq}$. Intuitively, we can understand the various components of this estimator as follows. The subtraction of $\mathbf{b}$ is a direct subtraction of a predetermined signal (for instance, a foreground model) from the data. The multiplication by $\mathbf{H}$ is an operation that potentially mixes voxels, and thus represents

[1] Or perhaps a "data shell", since there is a lower limit to the redshift of the experiment.

a data analysis/foreground cleaning operation that takes into account both angular and spectral information. The summing together of spatial pixels into a spectrum is done by applying $\mathbf{A}^t$, and $\mathbf{M}$ is a final data analysis/foreground cleaning step that uses spectral information only.

In the sections that follow, we will examine various algorithms that are defined by different choices for $\mathbf{M}$, $\mathbf{H}$, and $\mathbf{b}$. We will find that beyond simply subtracting off one's best guess for the foreground contamination, it is formally impossible to mitigate the effects of foregrounds using only spectral information, which is the only information available to instruments that integrate over a large portion of the sky and lose all spatial information. While such a method does remove foreground bias, it does not remove foreground covariance. In other words, foregrounds are removed according to the model, but nothing is done to immunize oneself against *errors* in the model. We will find that the situation is quite different for experiments that take advantage of even just modest angular resolution. The errors in the foreground models can be in some sense averaged down, making it possible to suppress foregrounds to a sufficient level for a statistically significant detection of the global 21 cm signal.

7.3.2 Methods using only spectral information

We begin with methods that seek to distinguish foregrounds from cosmological signal using spectral information only. In other words, these methods attempt to extract the signal using only the spectral information contained in plots such as those in Figures 7-4 and 7-1, and not any spatial information. These methods therefore have $\mathbf{H} = \mathbf{I}$, and we can define $\mathbf{z} \equiv \frac{1}{N_{pix}}\mathbf{A}^t\mathbf{y}$ to be the globally averaged measurement. In this notation, our measurement equation becomes

$$\mathbf{z} = \mathbf{x}_s + \frac{1}{N_{pix}}\mathbf{A}^t\mathbf{n} = \mathbf{x}_s + \mathbf{c}, \tag{7.20}$$

where $\mathbf{c} \equiv \frac{1}{N_{pix}}\mathbf{A}^t\mathbf{n}$ is the angularly averaged noise and foreground contribution, with mean $\langle\mathbf{c}\rangle = \frac{1}{N_{pix}}\mathbf{A}^t\langle\mathbf{n}\rangle = \frac{1}{N_{pix}}\mathbf{A}^t\mathbf{m}$ (where in the last step we used the definition in Equation 7.16) and covariance $\mathbf{C} \equiv \langle\mathbf{c}\mathbf{c}^t\rangle - \langle\mathbf{c}\rangle\langle\mathbf{c}^t\rangle = \frac{1}{N_{pix}^2}\mathbf{A}^t\mathbf{N}\mathbf{A}$. Correspondingly, our estimator $\widehat{\mathbf{x}}_s$ simplifies to:

$$\widehat{\mathbf{x}}_s = \mathbf{M}\mathbf{z} - \frac{1}{N_{pix}}\mathbf{M}\mathbf{A}^t\mathbf{b} \equiv \mathbf{M}\mathbf{z} - \mathbf{d}, \tag{7.21}$$

where we have made the definition $\mathbf{d} \equiv \frac{1}{N_{pix}}\mathbf{M}\mathbf{A}^t\mathbf{b}$. This is the general form of the estimator that must be used when analyzing data from experiments with no angular resolution, or if one deliberately chooses not use the angular information. We now seek to minimize the variance of this estimator by selecting $\mathbf{M}$ and $\mathbf{d}$ appropriately.

Taking the ensemble average of our estimator and inserting Equation 7.20 gives

$$\langle \widehat{\mathbf{x}}_s \rangle = \mathbf{M}\mathbf{x}_s + \mathbf{M}\langle \mathbf{c} \rangle - \mathbf{d}, \tag{7.22}$$

which shows that in order for our estimator to avoid having a systematic additive bias, one should select

$$\mathbf{d} \equiv \mathbf{M}\langle \mathbf{c} \rangle. \tag{7.23}$$

With this choice, we have $\langle \widehat{\mathbf{x}}_s \rangle = \mathbf{M}\mathbf{x}_s$, and in what follows we will impose the normalization constraint $\mathbf{M}_{\alpha\alpha} = 1$. The variance of this estimator can be similarly computed, yielding

$$\mathbf{\Sigma} = \langle \widehat{\mathbf{x}}_s \widehat{\mathbf{x}}_s^t \rangle - \langle \widehat{\mathbf{x}}_s \rangle \langle \widehat{\mathbf{x}}_s^t \rangle = \mathbf{M}\mathbf{C}\mathbf{M}^t. \tag{7.24}$$

We can minimize this variance subject to the normalization constraint by using Lagrange multipliers, minimizing the function

$$(\mathbf{M}\mathbf{C}\mathbf{M}^t)_{\alpha\alpha} - \lambda_\alpha \mathbf{M}_{\alpha\alpha} \tag{7.25}$$

with respect to the elements of $\mathbf{M}$. Taking the necessary derivatives and solving for the Lagrange multiplier λ that satisfies the normalization constraint, one obtains

$$\mathbf{M}_{\alpha\beta} = \frac{\mathbf{C}^{-1}_{\alpha\beta}}{\mathbf{C}^{-1}_{\alpha\alpha}}. \tag{7.26}$$

Inserting this into our general form for the estimator, we find

$$\widehat{\mathbf{x}}_s^\alpha = \frac{1}{\mathbf{C}^{-1}_{\alpha\alpha}} \sum_\beta \mathbf{C}^{-1}_{\alpha\beta} \left(\mathbf{z} - \langle \mathbf{c} \rangle \right)_\beta . \tag{7.27}$$

Intuitively, this prescription states that one should take the data, subtract off the known foreground contamination, and then to inverse variance weight the result before re-weighting to form the final estimator. The inverse variance weighting performs a statistical suppression/subtraction of instrumental noise and foreground contamination. Loosely speaking, this step corresponds to the subtraction of polynomial modes in the spectra as implemented in Bowman & Rogers (2010). The final re-weighting by $\mathbf{C}^{-1}_{\alpha\alpha}$ rescales the modes so that the previous subtraction step does not cause the

estimated modes to be biased high or low. For instance, if a certain mode is highly contaminated by foregrounds, it will be strongly down-weighted by the inverse variance weighting, and thus give an artificially low estimate of the mode unless it is scaled back up.

The corresponding error covariance for this estimator is given by

$$\mathbf{\Sigma}_{\alpha\beta} = \frac{\mathbf{C}^{-1}_{\alpha\beta}}{\mathbf{C}^{-1}_{\alpha\alpha}\mathbf{C}^{-1}_{\beta\beta}}. \tag{7.28}$$

It is instructive to compare this with the errors predicted by the Fisher matrix formalism. By the Cramer-Rao inequality, an estimator that is unwindowed (*i.e.* one that has $\langle \widehat{\mathbf{x}}_s \rangle = \mathbf{x}_s$) will have a covariance that is at least as large as the inverse of the Fisher matrix. Computing the Fisher matrix thus allows one to estimate the best possible errors bars that can be obtained from a measurement. In the approximation that fluctuations about the mean are Gaussian, the Fisher matrix takes the form

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta}\right] + \langle \mathbf{z} \rangle^{\dagger}_{,\alpha}\mathbf{C}^{-1}\langle \mathbf{z} \rangle_{,\beta}, \tag{7.29}$$

where commas denote derivatives with respect to the parameters that one wishes to measure. In our case, the goal is to measure the global 21 cm spectrum, so the parameters are the values of the spectrum at various frequencies. Put another way, we can write our mean measurement equation as

$$\langle \mathbf{z} \rangle = \sum_{\alpha} \mathbf{x}^{\alpha}_{s}\mathbf{e}_{\alpha} + \langle \mathbf{c} \rangle, \tag{7.30}$$

where $\mathbf{e}_{\alpha}$ is a unit vector with 0's everywhere except for a 1 at the α^{th} frequency channel. The derivative $\langle \mathbf{z} \rangle_{,\alpha}$ with respect to the α^{th} parameter (*i.e.* the derivative with respect to the mean measured spectrum $\mathbf{x}^{\alpha}_{s}$ in the α^{th} frequency channel) is therefore simply equal to $\mathbf{e}_{\alpha}$. Since the measurement covariance $\mathbf{C}$ does not depend on the cosmological signal $\mathbf{x}_s$, our Fisher matrix reduces to

$$F_{\alpha\beta} = \mathbf{e}_{\alpha}\mathbf{C}^{-1}\mathbf{e}_{\beta}, \tag{7.31}$$

which shows that the Fisher matrix is simply the inverse covariance *i.e.* $\mathbf{F} = \mathbf{C}^{-1}$. This implies that the covariance of the optimal unwindowed method is equal to the original noise and foreground covariance, which means that the error bars on the estimated spectrum are no smaller than if no line-of-sight foreground subtraction

were attempted beyond the initial removal of foreground bias (Equations 7.22 and 7.23). In our notation, an unwindowed estimator would be one with $\mathbf{M} = \mathbf{I}$ (from Equation 7.22 plus the requirement that the estimator be unbiased, *i.e.* Equation 7.23). But this means our optimal unwindowed estimator is

$$\widehat{\mathbf{x}}_s = \mathbf{z} - \langle \mathbf{c} \rangle, \tag{7.32}$$

which says to subtract our foreground spectrum model but to do nothing more!

Intuitively, our optimizations yielded a "do nothing" algorithm because the global signal that we seek to measure is itself a spectrum (with unique cosmological information in every frequency channel), so a spectral-only measurement provides no redundancy. With no redundancy, one can only subtract a best-guess foreground model, and it is impossible to minimize statistical errors in the model itself.

The key, then, is to have multiple, redundant measurements that all containt the same cosmological signal. This can be achieved by designing experiments with angular information (*i.e.* ones that do not integrate over the entire sky automatically). Because the global signal is formally a spatial monopole, measurements in different pixels on the sky have identical cosmological contributions, but different foreground contributions, allowing us to further distinguish between foregrounds and cosmological signal. We now develop optimal algorithms for processing data from such experiments.

7.3.3 Methods using both spectral and angular information

We now turn to data analysis methods for experiments with both angular and spectral information. These experiments provide us with $\mathbf{y}$ rather than $\mathbf{z} \equiv \frac{1}{N_{pix}}\mathbf{A}^t\mathbf{y}$, the already-spatially-averaged signal. Such experiments give us the freedom of processing the data in some way before the $\frac{1}{N_{pix}}\mathbf{A}^t$ spatial averaging step. In the notation of Equation 7.19, they provide us with the freedom of setting $\mathbf{H}$ to be something other than $\mathbf{I}$.

With this freedom, the problem of selecting $\mathbf{M}$, $\mathbf{H}$, and $\mathbf{b}$ to minimize the error bars on the estimator $\widehat{\mathbf{x}}_s$ becomes a solved problem (Tegmark, 1997b). The answer is to pick $\mathbf{H} = \mathbf{N}^{-1}$, $\mathbf{b} = \mathbf{n}$, and $\mathbf{M} = [\mathbf{A}^t\mathbf{N}^{-1}\mathbf{A}]^{-1}$, so that

$$\widehat{\mathbf{x}}_s = \left[\mathbf{A}^t\mathbf{N}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}^t\mathbf{N}^{-1}(\mathbf{y} - \mathbf{m}). \tag{7.33}$$

With an estimator of this form, the error covariance can be shown to be

$$\mathbf{\Sigma} \equiv \langle \widehat{\mathbf{x}}_s \widehat{\mathbf{x}}_s^t \rangle - \langle \widehat{\mathbf{x}}_s \rangle \langle \widehat{\mathbf{x}}_s^t \rangle = \left[\mathbf{A}^t \mathbf{N}^{-1} \mathbf{A} \right]^{-1}. \tag{7.34}$$

Although these equations in principle encode all we need to know for data analysis, it is necessary to manipulate them a little further. Doing so not only allows us to build intuition for what the matrix algebra is doing, but also makes the algorithm computationally feasible. In its current form, Equation 7.33 is too computationally expensive to be practical because it involves the inversion of $\mathbf{N}$, which is an $N_{vox} \times N_{vox}$ matrix. Given that N_{vox} is likely to be large for a 21 cm with full spectral and angular information, it would be wise to avoid direct matrix inversions[2].

We thus seek to derive analytic results for $\mathbf{\Sigma} \equiv [\mathbf{A}^t \mathbf{N}^{-1} \mathbf{A}]^{-1}$ and $\mathbf{A}^t \mathbf{N}^{-1}$, where $\mathbf{N}$ is given by Equation 7.17. We begin by factoring out the foreground templates from our noise covariance, so that $\mathbf{N} = \mathbf{DWD}$, where $\mathbf{D}_{\alpha i \beta j} \equiv \mathbf{m}_{\alpha i} \delta_{\alpha\beta} \delta_{ij}$. This is essentially a whitening procedure, making the noise term independent of frequency or sky pixel. Since both $\mathbf{\Sigma} \equiv [\mathbf{A}^t \mathbf{N}^{-1} \mathbf{A}]^{-1}$ and $\mathbf{A}^t \mathbf{N}^{-1}$ involve $\mathbf{N}^{-1}$, we proceed by finding $\mathbf{N}^{-1} = \mathbf{D}^{-1} \mathbf{W}^{-1} \mathbf{D}^{-1}$, and to do so we move into a diagonal basis. The matrix $\mathbf{D}$ is already diagonal and easily invertible, so our first step is to perform a similarity transformation in the frequency-frequency components of $\mathbf{W}$ in order to diagonalize $\mathbf{Q}$. In other words, we can write $\mathbf{W}$ as

$$\mathbf{W}_{\alpha i \beta j} = (\mathbf{V} \overline{\mathbf{W}} \mathbf{V}^t)_{\alpha i \beta j} = \sum_{\eta \lambda k m} \mathbf{V}_{\alpha i \eta k} \left(\varepsilon^2 \lambda_\eta \mathbf{R}_{km} + \frac{\delta_{km}}{\Delta t \Delta \nu} \right) \delta_{\eta\lambda} \mathbf{V}_{\beta j \lambda m}, \tag{7.35}$$

where $\mathbf{V}_{\alpha i \eta k} \equiv (\mathbf{v}_\eta)_\alpha \delta_{ik}$, with $\mathbf{v}_\eta \alpha$ signifying the value of the α^{th} component (frequency) of the η^{th} eigenvector of $\mathbf{Q}$. The η^{th} eigenvalue as given by λ_η. Note also that $\mathbf{V}^{-1} = \mathbf{V}^t$.

Our next step is to diagonalize $\mathbf{R}$, the spatial correlations. If we assume that the correlations are rotationally invariant[3], the matrix will be diagonal in a spherical harmonic basis. For computational convenience, we work in the continuous limit, so that we have

$$R(\widehat{\mathbf{r}}, \widehat{\mathbf{r}}') \equiv \frac{\exp(\sigma^{-2} \widehat{\mathbf{r}} \cdot \widehat{\mathbf{r}}')}{4\pi\sigma^2 \sinh(\sigma^{-2})}, \tag{7.36}$$

[2]Although sometimes they are inevitable, as we saw in Chapter 6.

[3]Recall that $\mathbf{R}$ encodes the spatial correlations in the errors of our foreground model. It is thus entirely possible to break rotation invariance, for instance by using a foreground model that is constructed from a number of different surveys, each possessing different error characteristics and different sources of error. For this paper we ignore this possibility in order to gain analytical intuition, but we note that it is easy to correct for — one simply finds the eigenvectors and eigenvalues of $\mathbf{R}$, just as we did with $\mathbf{Q}$.

which is known as a Fisher function, the analog of a Gaussian on a sphere. For $\sigma \ll 1\,\mathrm{rad}$, this reduces to the familiar Gaussian:

$$R(\widehat{\mathbf{r}},\widehat{\mathbf{r}}') \approx \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2}\frac{\theta^2}{\sigma^2}\right], \tag{7.37}$$

where $\theta \equiv \arccos(\widehat{\mathbf{r}}\cdot\widehat{\mathbf{r}}')$ is the angle between the two locations on the sphere. Switching to a spherical harmonic basis involves the usual integrals, but in order to be dimensionally correct when moving back and forth between the continuous limit and the discrete, pixelized sky, each integral receives an extra factor of N_{pix}/Ω_{sky}, where Ω_{sky} is the sky solid angle surveyed. Similarly, our Fisher function is augmented by a factor of Ω_{sky}/N_{pix} so that it converges to Kronecker delta function in the limit $\sigma \to 0$. We thus have

$$\overline{R}_{\ell m\ell' m'} \equiv \frac{N_{pix}}{\Omega_{sky}}\int d\Omega d\Omega' Y^*_{\ell m}(\widehat{\mathbf{r}})R(\widehat{\mathbf{r}},\widehat{\mathbf{r}}')Y_{\ell' m'}(\widehat{\mathbf{r}}') \approx \frac{N_{pix}}{\Omega_{sky}}\exp\left[-\frac{1}{2}\sigma^2\ell(\ell+1)\right]\delta_{\ell\ell'}\delta_{mm'}, \tag{7.38}$$

where $Y_{\ell m}$ denotes the spherical harmonic with azimuthal quantum number ℓ and magnetic quantum number m, and the last approximation holds if $\sigma \ll 1\,\mathrm{rad}$. Doing the same for the instrumental noise term in Equation 7.35 and putting everything together gives

$$\widehat{\mathbf{W}}_{\ell m\eta\ell' m'\lambda} = \frac{N_{pix}}{\Omega_{sky}}\left[\varepsilon^2\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{\Delta t\Delta\nu}\right]\delta_{\ell\ell'}\delta_{mm'}\delta_{\eta\lambda}. \tag{7.39}$$

The diagonal nature of $\widehat{\mathbf{W}}$ in this equation allows a straightforward inversion:

$$\widehat{\mathbf{W}}^{-1}_{\ell m\eta\ell' m'\lambda} = \frac{N_{pix}}{\Omega_{sky}}\left[\varepsilon^2\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{\Delta t\Delta\nu}\right]^{-1}\delta_{\ell\ell'}\delta_{mm'}\delta_{\eta\lambda}, \tag{7.40}$$

thus allowing us to write the inverse matrix in the original (frequency-frequency) basis as

$$(\overline{\mathbf{W}})^{-1}_{\alpha i\beta j} = (\mathbf{F}^\dagger\widehat{\mathbf{W}}^{-1}\mathbf{F})_{\alpha i\beta j} = \frac{N_{pix}}{\Omega_{sky}}\sum_{\eta\lambda\ell\ell' mm'}\mathbf{F}^\dagger_{\alpha i\eta\ell m}\left[\varepsilon^2\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{\Delta t\Delta\nu}\right]^{-1}\delta_{\ell\ell'}\delta_{mm'}\delta_{\eta\lambda}\mathbf{F}_{\lambda\ell' m'\beta j}, \tag{7.41}$$

where $\mathbf{F}_{\beta\ell m\alpha i} \equiv \delta_{\alpha\beta}\mathbf{f}_{\ell mi}$ and $\mathbf{f}$ is a unitary matrix that transforms from a real-space angular basis to a spherical harmonic basis. Obtaining the inverse of $\mathbf{W}$ from here is done by evaluating $\mathbf{W}^{-1} = \mathbf{V}\overline{\mathbf{W}}^{-1}\mathbf{V}^t$.

We are now ready to assemble the pieces to form $\mathbf{\Sigma} \equiv [\mathbf{A}^t\mathbf{N}^{-1}\mathbf{A}]^{-1}$ which, re-

member, is equal to $[\mathbf{A}^t\mathbf{D}^{-1}\mathbf{F}^\dagger\mathbf{V}\overline{\mathbf{W}}^{-1}\mathbf{V}^t\mathbf{F}\mathbf{D}^{-1}\mathbf{A}]^{-1}$. We first compute

$$(\mathbf{V}^t\mathbf{F}\mathbf{D}^{-1})_{\alpha\ell m\beta j} = \mathbf{f}_{\ell mj}(\mathbf{v}_\alpha)_\beta(\mathbf{m}_{\beta\mathbf{j}})^{-1} \equiv \mathbf{f}_{\ell mj}(\mathbf{v}_\alpha)_\beta\mathbf{u}_{\beta j}, \tag{7.42}$$

where $\mathbf{u}_{\alpha i} \equiv 1/\mathbf{m}_{\alpha i}$ is the reciprocal map (at all frequencies) of our foreground templates. Note that in the above expression, there is no sum over j yet. This is accomplished by the angular summation matrix $\mathbf{A}_{\alpha i\beta} = \delta_{\alpha\beta}$, giving

$$(\mathbf{V}^t\mathbf{F}\mathbf{D}^{-1}\mathbf{A})_{\alpha\ell m\beta} = (\mathbf{v}_\alpha)_\beta \sum_j \mathbf{f}_{\ell mj}\mathbf{u}_{\beta j} = (\mathbf{v}_\alpha)_\beta\widehat{\mathbf{u}}_{\beta\ell m}, \tag{7.43}$$

where there is similarly no sum over β, and $\widehat{\mathbf{u}}$ signifies our reciprocal foreground templates in spherical harmonic space. The inverse covariance $\mathbf{\Sigma}^{-1}$ is thus given by

$$\mathbf{\Sigma}^{-1}_{\alpha\beta} = \frac{N_{pix}}{\Omega_{sky}}\sum_\eta(\mathbf{v}_\eta)_\alpha(\mathbf{v}_\eta)_\beta\sum_{\ell m}\frac{\widehat{\mathbf{u}}_{\alpha\ell m}\widehat{\mathbf{u}}_{\beta\ell m}}{\varepsilon^2\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{\Delta t\Delta\nu}}. \tag{7.44}$$

At this point, we note that ε itself may be a function of N_{pix} — our degree of confidence in our foreground model will likely depend on the angular resolution that we are dealing with. In particular, a foreground model is likely to be more accurate at low resolutions, since a low resolution map can be thought of as being derived from combining separate independent measurements at high resolution. We may thus crudely say that $\varepsilon(N_{pix}) = \varepsilon_0\sqrt{N_{pix}/N_{ref}}$, where ε is the error parameter for our foreground model at its "native" angular resolution, with N_{ref} pixels. Using this form for ε and letting $t_{int} \equiv N_{pix}\Delta t$ be the total integration time over the survey, we have

$$\mathbf{\Sigma}^{-1}_{\alpha\beta} = \frac{1}{\Omega_{sky}}\sum_\eta(\mathbf{v}_\eta)_\alpha(\mathbf{v}_\eta)_\beta\sum_{\ell m}\frac{\widehat{\mathbf{u}}_{\alpha\ell m}\widehat{\mathbf{u}}_{\beta\ell m}}{\frac{\varepsilon_0^2}{N_{ref}}\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{t_{int}\Delta\nu}}. \tag{7.45}$$

At this point, notice that the angular cross-power spectrum $C_\ell^{\alpha\beta}$ between two reciprocal maps $\mathbf{u}_{\alpha i}$ and $\mathbf{u}_{\beta i}$ at frequencies α and β respectively is given by

$$C_\ell^{\alpha\beta} \equiv \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell}\widehat{\mathbf{u}}_{\alpha\ell m}\widehat{\mathbf{u}}_{\beta\ell m}, \tag{7.46}$$

so our expression for the inverse measurement covariance $\mathbf{\Sigma}^{-1}_{\alpha\beta}$ can be written as

$$\mathbf{\Sigma}^{-1}_{\alpha\beta} = \frac{1}{\Omega_{sky}}\sum_\eta(\mathbf{v}_\eta)_\alpha(\mathbf{v}_\eta)_\beta\sum_\ell\frac{2\ell+1}{\frac{\varepsilon_0^2}{N_{ref}}\lambda_\eta e^{-\frac{1}{2}\sigma^2\ell(\ell+1)} + \frac{1}{t_{int}\Delta\nu}}C_\ell^{\alpha\beta}. \tag{7.47}$$

This provides a fast prescription for computing $\mathbf{\Sigma} \equiv [\mathbf{A}^t \mathbf{N}^{-1} \mathbf{A}]^{-1}$. One simply uses publicly available fast routines for computing the angular cross-power spectrum, multiplies by a function of ℓ, sums over all ℓ's, and performs some sums over eigenvectors $\mathbf{v}$ of $\mathbf{Q}$, which can be precomputed. Note that the matrix inversion to go from our inverse measurement covariance $\mathbf{\Sigma}^{-1}$ to the measurement covariance $\mathbf{\Sigma}$ is not computationally expensive because in a global signal experiment our measurement covariance has dimensions $N_{freq} \times N_{freq}$, and not $N_{vox} \times N_{vox}$ as is the case for the full data set.

While Equation 7.47 is convenient for computational purposes, for purposes of intuition it useful to rewrite $\mathbf{\Sigma}_{\alpha\beta}$ in the following manner. Incorporating the denominator of Equation 7.45 into one of the copies of $\widehat{\mathbf{u}}$, we can recast the equation as

$$\mathbf{\Sigma}^{-1}_{\alpha\beta} = \frac{1}{\Omega_{sky}} \sum_{\eta} (\mathbf{v}_\eta)_\alpha (\mathbf{v}_\eta)_\beta \sum_{\ell m} \widehat{\mathbf{u}}^{\eta}_{\alpha\ell m} \widetilde{\mathbf{u}}^{\eta}_{\beta\ell m}, \tag{7.48}$$

where $\widetilde{\mathbf{u}}^{\eta}_{\alpha\ell m}$ is a weighted version of the reciprocal foreground template in spherical harmonic space:

$$\widetilde{\mathbf{u}}^{\eta}_{\alpha\ell m} \equiv \widehat{\mathbf{u}}_{\alpha\ell m} \left[\frac{\varepsilon_0^2}{N_{ref}} \lambda_\eta e^{-\frac{1}{2}\sigma^2 \ell(\ell+1)} + \frac{1}{t_{int} \Delta\nu} \right]^{-1}. \tag{7.49}$$

Using the spherical harmonic generalization of Parseval's theorem, we can rewrite our inverse error covariance as

$$\begin{aligned} \mathbf{\Sigma}^{-1}_{\alpha\beta} &= \sum_{\eta} (\mathbf{v}_\eta)_\alpha (\mathbf{v}_\eta)_\beta \left(\frac{1}{\Omega_{sky}} \int \frac{d\Omega}{m(\hat{\mathbf{r}}, \nu_\alpha) m^{\eta}_{hp}(\hat{\mathbf{r}}, \nu_\beta)} \right) \\ &\rightarrow \sum_{\eta} (\mathbf{v}_\eta)_\alpha (\mathbf{v}_\eta)_\beta \left(\frac{1}{N_{pix}} \sum_i \frac{1}{\mathbf{m}_{i\alpha} (\mathbf{m}_{hp})^{\eta}_{i\beta}} \right), \end{aligned} \tag{7.50}$$

where in the last step we moved back from the continuous limit to a discretized (*i.e.* pixelized) system, and defined the filtered foreground template

$$m^{\eta}_{hp}(\hat{\mathbf{r}}, \nu_\alpha) = \left[\sum_{\ell m} \frac{\widehat{\mathbf{u}}_{\alpha\ell m} Y_{\ell m}(\hat{\mathbf{r}})}{\frac{\varepsilon_0^2}{N_{ref}} \lambda_\eta e^{-\frac{1}{2}\sigma^2 \ell(\ell+1)} + \frac{1}{t_{int}\Delta\nu}} \right]^{-1}, \tag{7.51}$$

which is essentially a high-pass filtered version of our original foreground template.

We now turn to computing $\mathbf{A}^t \mathbf{N}^{-1}$, which we also need for Equation 7.33. Instead of computing this quantity directly, we compute $\mathbf{w} \equiv \mathbf{N}^{-1} \mathbf{A}$, which can be interpreted

as an optimal set of spatial weights to be used when summing the measured sky maps into a global spectrum. To see this, note that Equation 7.33 can be written as

$$(\widehat{\mathbf{x}}_s)_\alpha = \sum_{\beta\gamma} \left[\mathbf{A}^t\mathbf{N}^{-1}\mathbf{A}\right]^{-1}_{\alpha\beta} \sum_i \mathbf{w}_{\beta i\gamma}(\mathbf{y}-\mathbf{m})_{i\gamma}. \tag{7.52}$$

Following manipulations similar to what we had above, we obtain

$$\mathbf{w}_{\alpha i\beta} = \left[\mathbf{D}^{-1}\mathbf{F}^\dagger\mathbf{V}\overline{\mathbf{W}}^{-1}\mathbf{V}^t\mathbf{F}\mathbf{D}^{-1}\mathbf{A}\right]_{\alpha i\beta} = \frac{1}{N_{pix}}\sum_\eta (\mathbf{v}_\eta)_\alpha (\mathbf{v}_\eta)_\beta \frac{1}{\mathbf{m}_{i\alpha}(\mathbf{m}_{hp})^\eta_{i\beta}}. \tag{7.53}$$

Armed with Equations 7.52 and 7.53, we can interpret our data analysis algorithm as the following recipe:

1. Take the measured sky map at every frequency and subtract our best guess foreground model. This corresponds to the $\mathbf{y}-\mathbf{m}$ part of Equation 7.52.

2. Perform a weighted spatial average of the resulting sky maps to form a spectrum. The weights are given by Equation 7.53 and roughly speaking correspond to downweighting each map first by a factor of our best-guess sky model $\mathbf{m}$, and then downweighting by a factor of a high-pass filtered version $\mathbf{m}_{hp}$ of the same model. Both steps have the effect of suppressing the most foreground-contaminated part of the sky when forming our spatial average. The best-guess sky model simply downweights in proportion to expected contamination (Figure 7-5), while the high pass filtered version (Figure 7-6) singles out the "worst offenders" and suppresses them further[4].

3. Multiply by the appropriate foreground eigenmodes $\mathbf{v}_\eta$ and sum over all foreground eigenmodes. This yields a spectrum in ordinary frequency space.

4. Normalize the spectrum using $[\mathbf{A}^t\mathbf{N}^{-1}\mathbf{A}]^{-1}$. This is necessary because our procedures may have resulted in uneven weightings at different frequencies. This step also gets rid of unwanted correlations between frequency channels. Note that in a strict sense, this final matrix multiplication has no effect on how well one can constrain theoretical models. The reason for this is that the matrix in question is invertible, and there is thus no change in information content. Different choices for the final multiplication result in estimators with different

[4] A close examination of the numbers reveals that the high-pass filtered map in fact does a little more—the map is negative in some regions, suggesting that there is also some differencing between pixels to subtract out certain contaminants

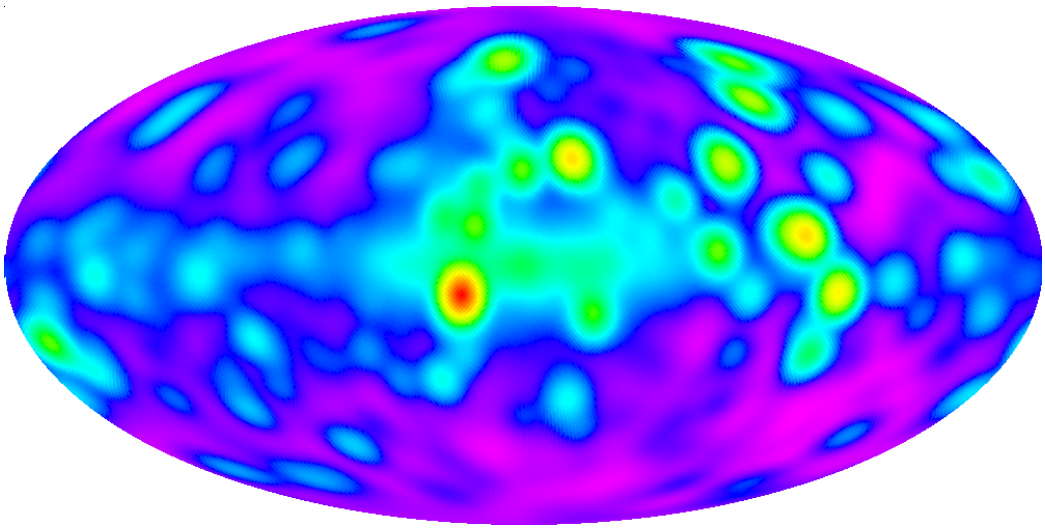

Figure 7-5: Best-guess sky model at 30 MHz, smoothed to approximately 1° resolution.

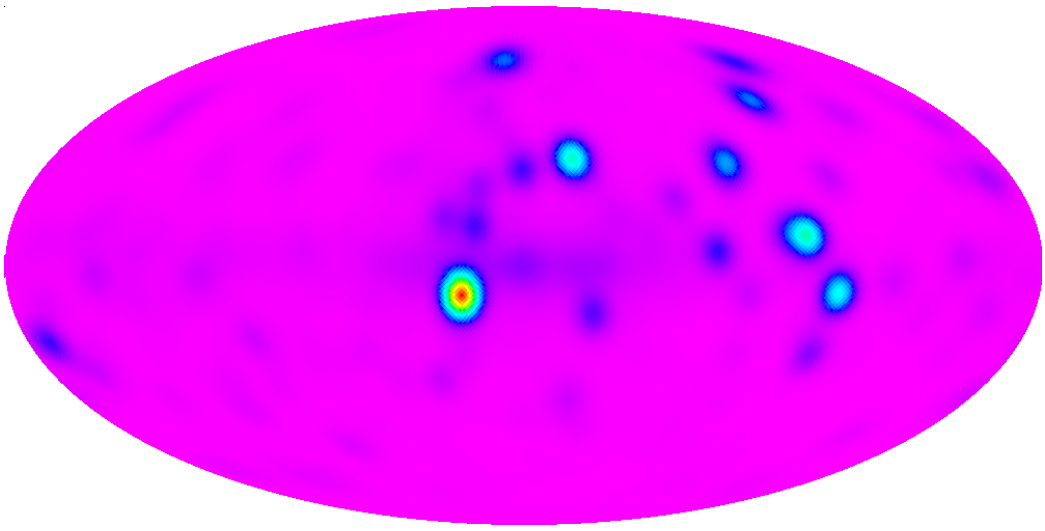

Figure 7-6: High-pass filtered sky model (given by Equation 7.51, with the parameters of the model given by the MID fiducial scenario in Table 7.1) at 30 MHz, smoothed to approximately 1° resolution.

statistical properties, and our particular choice here has the desirable property that it produces the smallest possible error bars (square root of the diagonal elements of $\mathbf{\Sigma}$) subject to the constraint that the final estimate of the global signal spectrum is unwindowed.

7.4 Designing an experiment — an exploration of parameter space

Having established an optimal procedure for the analysis of data from global signal experiments, we now turn to the design of global signal experiments. In particular, we consider the various parameters that go into determining our measurement covariance $\mathbf{\Sigma}$ (our equivalently the error bars $\sqrt{\mathbf{\Sigma}_{\alpha\alpha}}$), and ask how well different experimental designs can perform *given* our data analysis algorithm.

7.4.1 Properties of the foreground model

We first consider parameters that deal with the foreground model. Some of these, such as the "true" amplitude of Galactic synchrotron radiation, are "determined by our Universe" and in our study we will simply pick fiducial values for them. The remaining parameters—the foreground model error ε_0, the spatial coherence of foreground errors σ, and the spectral coherence of foregrounds σ_α—are determined by the previous foreground surveys that were used to construct our model. These parameters quantify, in a loose sense, the quality of our survey, and therefore affect our ability to mitigate foregrounds in our global signal measurement.

Consider first the fractional foreground model error ε_0. In Figure 7-7 we show the expected error bars $\sqrt{\mathbf{\Sigma}_{\alpha\alpha}}$ as a function of frequency. The absolute value of the theoretically expected cosmological signal is shown as a black dashed line. The parameter ε is varied about a fiducial value of 0.01, while all other parameters are fixed at the fiducial values of the MID scenario in Table 7.1. As expected, the lower the error in our foreground model, the better we can subtract off foregrounds, and the lower the error bars on our final measurement of the power spectrum. Importantly, we see that in our fiducial scenario we can achieve error bars that are much smaller than the size of the expected cosmological signal. While this fiducial scenario requires a (currently non-existing) foreground model that is accurate to 1%, such a model is not inconceivable, for at these low frequencies it only requires a survey of the radio sky that is good to ~ 10's of Kelvin.

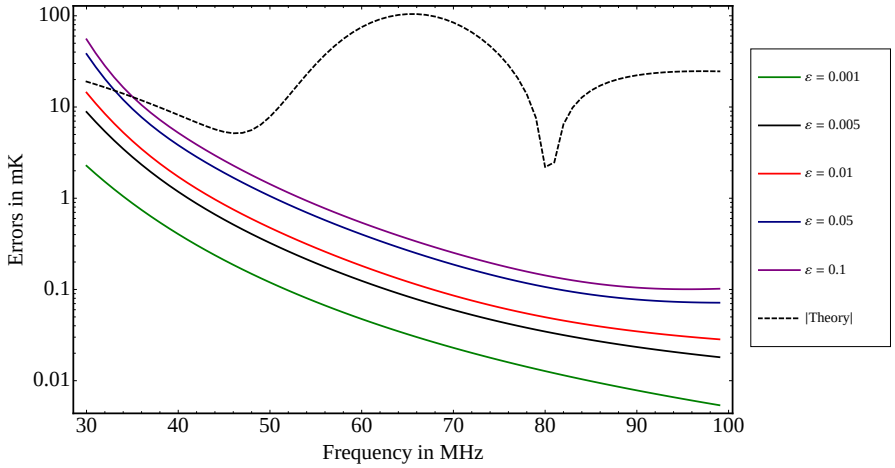

Figure 7-7: Expected error bars as a function of frequency, varying the fractional foreground model error ε_0.

Now consider the spatial coherence σ of the errors in our foreground model. In Figure 7-8 we vary the value of σ, once again showing the expected error bars as a function of frequency. Since this is a spatial scale of error correlations, it is intimately connected to the properties of the foreground survey instrument that was used to construct the model. A survey instrument with a very wide beam, for instance, will tend to produce foreground maps that are very spatially correlated, giving a large σ. From Figure 7-8 we see that the larger the spatial coherence, the smaller the error bars on our measured global signal. This is because a spatially correlated error is one that can be potentially removed using the patterns of the correlation. Put another way, if the errors in a large number of pixels are strongly correlated, the errors can be described using a small number of modes, making them easier to marginalize over. On the other hand, uncorrelated errors can only be averaged down. Note that this does *not* imply that one ought to produce foreground models using instruments with the worst possible angular resolution! Our final measurement errors on the global signal are smallest with high σ *if all other parameters are kept constant.* In practice, a foreground model constructed from surveys with very little angular sensitivity will be quite inaccurate, resulting in a large ε_0.

In Figure 7-9 we vary the spectral coherence of our foreground model errors. Two competing influences are at work here. On one hand, we can make a similar argument to the one we made for the spatial coherence: higher coherence means there are fewer modes to be taken out in our subtraction. This is the behavior shown in Figure 7-9, where the measurement errors on the global signal go down with greater coherence. On the other hand, the situation is not strictly analogous to the spatial case because

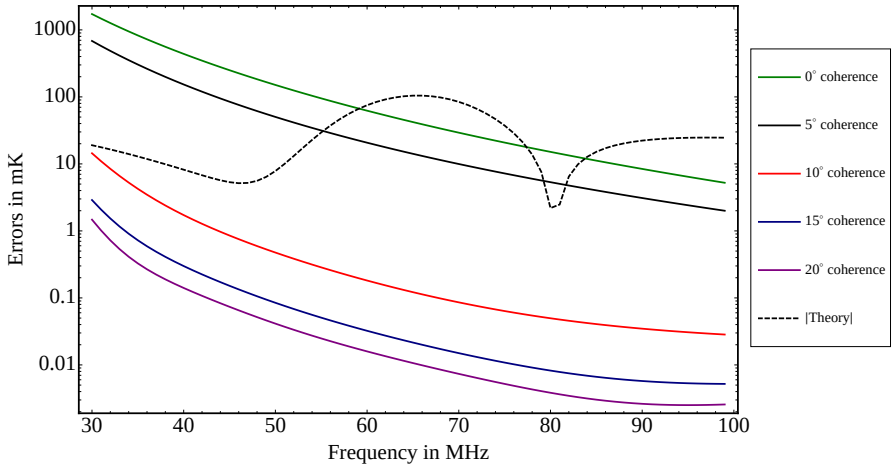

Figure 7-8: Expected error bars as a function of frequency, varying the spatial coherence length scale σ.

our goal is to measure a global *spectrum*. With high spectral coherence, the errors in our foreground model are highly correlated, and this correlation propagates through to our final measurement of the global signal. The measured spectrum thus contains fewer independent data points, representing a degraded constraint. However, the extent to which this affects one's ability to learn about the EoR depends on the expected cosmological signal, and a detailed study of this is beyond the scope of this work. In general, one should aim to produce a foreground model with errors that are as spectrally uncorrelated as possible. The reason is the same as for the spatial correlations—a foreground survey that is unable to make independent measurements at different frequencies will give rise to a rather inaccurate foreground model, one that has a high ε_0.

7.4.2 Properties of the instrument

We now turn to properties of the instrument that we use to measure the global signal. To keep the discussion general, we do not assume a specific instrument or even a specific instrumental design. Instead, we consider a generic radio telescope, characterized by its integration time and its beam pattern.

In general, a radio telescope may have a complicated beam pattern. Fortunately, as we will now prove, complicated beams have no effect on our data analysis algorithms. To take into account a (possibly complicated) beam shape, we can rewrite

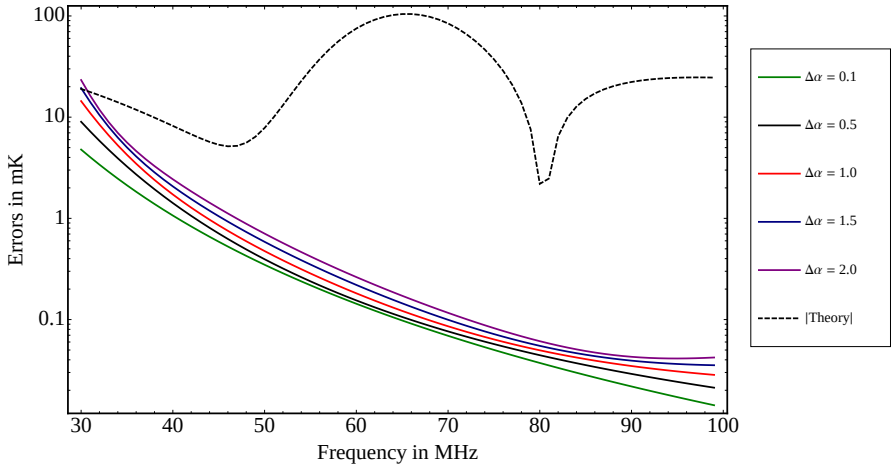

Figure 7-9: Expected error bars as a function of frequency, varying the spectral coherence length scale.

our measurement equation (Equation 7.18) as

$$\mathbf{y} = \mathbf{BAx}_s + \mathbf{Bn}^{fg} + \mathbf{n}^{inst}, \tag{7.54}$$

where we have elected to separate the foreground noise and instrumental noise contributions because the former is a sky signal and is thus multiplied by the beam, whereas the latter is added by the instrument and is not affected by the beam. The matrix $\mathbf{B}_{\alpha i \beta j} = \mathbf{b}^{\alpha}_{ij}\delta_{\alpha\beta}$, and $\mathbf{b}^{\alpha}_{ij}$ encodes the beam response in the i^{th} pixel from the j^{th} direction, at the α^{th} frequency. Note that the index j will in general run over a wider range of values than the index i. In fact, the former should really be continuous whereas the latter range is determined by the angular resolution of one's instrument. Now, we can say without loss of generality that $\sum_j \mathbf{b}^{\alpha}_{ij} = 1$, *i.e.* the beam profiles are normalized to unity. Using this fact, we have

$$(\mathbf{BA})_{\alpha i \beta} = \delta_{\alpha\beta} \sum_k \mathbf{b}^{\alpha}_{ik} = \delta_{\alpha\beta} = \mathbf{A}_{\alpha i \beta}, \tag{7.55}$$

with the only subtlety being that the index i on the new copy of the $\mathbf{A}$ matrix runs over a smaller set of values than before. The effect of multiplying by $\mathbf{B}$, then, is to reduce the number of rows of $\mathbf{A}$ while preserving its structure. Thus, as long as we define binned and smoothed noise vector $\mathbf{n}' \equiv \mathbf{Bn}^{fg} + \mathbf{n}^{inst}$, our formalism from previous sections remains intact. Put another way, having a complicated beam does not complicate our analysis algorithm; it simply requires that we take this beam into

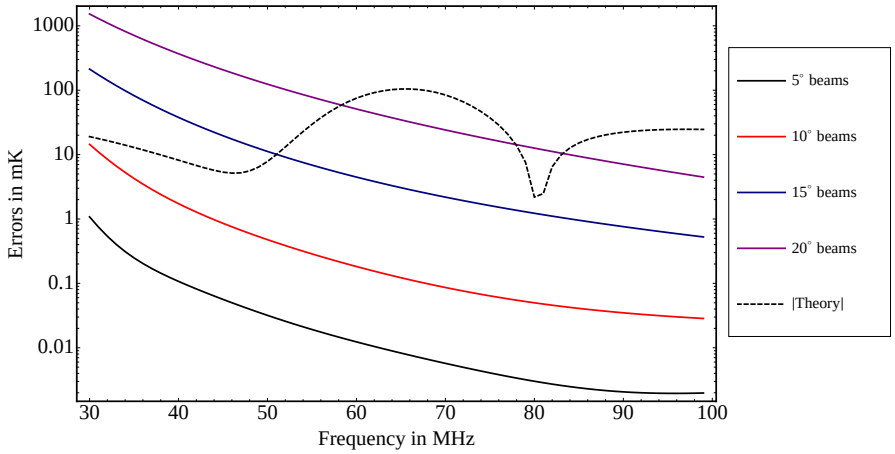

Figure 7-10: Expected error bars as a function of frequency, varying the angular resolution.

account when we construct our noise model[5]. Intuitively, the detailed shape of the beam is unimportant because we are interested in the spatially averaged global signal, and so we would integrate over the beam profile in our analysis anyway.

While the shape of the beam may not have a large impact on our analysis, its *width* certainly does, for it controls the amount of spatial information that is available for mitigating foreground errors (as discussed in Section 7.3.3). On one extreme, for instance, an instrument with a beam that encompasses the entire sky would yield no spatial information whatsoever, and one would have to resort to the spectral-only analysis techniques of Section 7.3.2. To parameterize the dependence on beam width (*i.e.* angular resolution), we perform calculations with Gaussian beams of various widths (quantified by each beam's full-width-half-max). In Figure 7-10, we show the expected errors as a function of angular resolution. As expected, finer angular resolution yields lower measurement errors, as a smaller beam allows one to better isolate highly foreground-contaminated regions and to downweight them, as well as to take advantage of any small scale correlations in the foreground model errors that may be used to subtract foregrounds.

In Figure 7-11 we vary the integration time. We see that while longer integrations do reduce the errors, the improvement is not dramatic as long as one is within the

[5]While it is only the foreground noise that appears explicitly in the expression $\mathbf{n}' \equiv \mathbf{B}\mathbf{n}^{fg} + \mathbf{n}^{inst}$, the instrumental noise is in fact also affected. This is because our instrument is sky-noise dominated, and thus the measured foregrounds affect the *amplitude* of the instrumental noise. However, the beam does not result in any noise *correlations*, which is what the matrix $\mathbf{B}$ encodes. In our calculations, we take into account the modifications to both the foreground noise and instrumental noise.

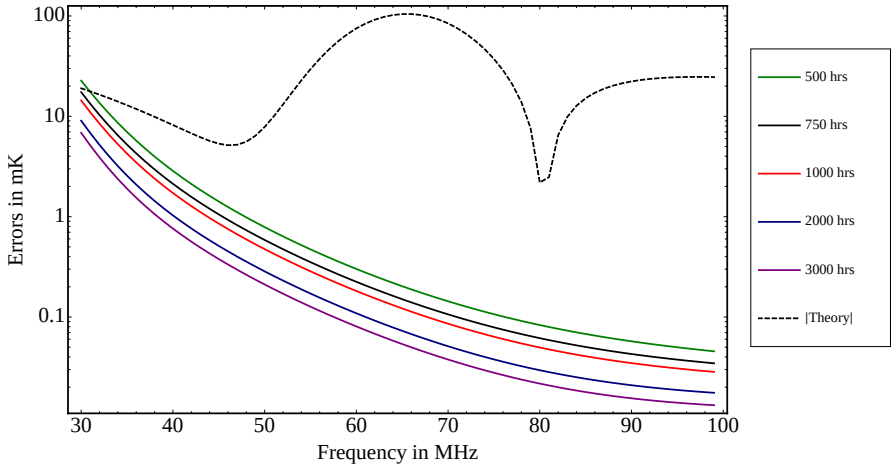

Figure 7-11: Expected error bars as a function of frequency, varying the integration time.

	Pessmistic scenario (PESS)	Middle-of-the-road scenario (MID)	Optimistic scenario (OPT)
ε_0	0.1	0.01	0.001
σ (spatial coherence)	0 degrees	10 degrees	20 degrees
σ_α (spec. coherence)	2	1	0.1
Angular Resolution	20 degrees	10 degrees	5 degrees
t_{int}	500 hrs	1000 hrs	3000 hrs

Table 7.1: Fiducial scenarios examined in Section 7.5.

typical range of integration times for proposed measurements (roughly $\sim$ 100's to 1000's of hours). In this regime, instrumental noise (which is the only source of noise that would average down with further integration) is a subdominant contribution to the errors, and larger improvements can be obtained by instead reducing residual foreground contamination.

7.5 Fiducial Results

We now present results from a set of three fiducial scenarios: a pessimistic scenario (PESS), a middle-of-the-road scenario (MID), and an optimistic (OPT) model. Their parameters are listed in Table 7.1. The error bars for them are shown in Figure 7-12

As we can see from Figure 7-12, the chances of a statistical significant detection of the cosmological signal are high for both the MID and OPT scenarios. We can crudely quantify the significance of a detection by computing the "number of sigmas"

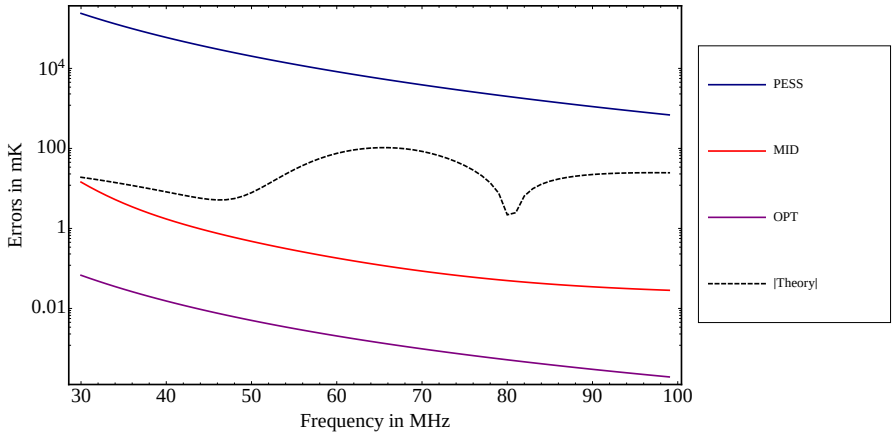

Figure 7-12: Some fiducial models

ν over which a cosmological signal can be seen over pure noise and foregrounds. This is given by $\nu \equiv \sqrt{\mathbf{s}^t \boldsymbol{\Sigma} \mathbf{s}}$, where $\mathbf{s}$ is the cosmological global signal spectrum packaged into a vector, and $\boldsymbol{\Sigma}$ is the measurement covariance. For the MID scenario, one obtains $\nu = 90$, suggesting a relatively easy detection.

7.6 Conclusion

In this chapter, we have considered the prospects for a detection of the global, spatially-averaged 21 cm signal between 30 and 70 MHz. We focused on this particular frequency range because theoretical expectations suggest the existence of a relatively large (~ 100 mK) absorption signal. This arises from the coupling of the 21 cm spin temperature to the hydrogen gas temperature via the Wouthysen-Feld effect during early reionization, and the signal becomes even more pronounced if X-ray heating of the IGM is small (Pritchard & Loeb, 2010).

We have also developed formalism for analyzing data from global signal experiments that seek to measure the aforementioned absorption signature. This formalism uses full spectral and spatial information, and allows one to use the spatial information to correct for errors in one's foreground model, yielding smaller error bars than would be expected if only spectral information were used. Computing these error bars for a wide range of instrumental and foreground parameters suggests that a statistically significant detection of the cosmological global signal should be achievable.

Chapter 8

A Measurement of the Redshifted 21 cm Power Spectrum with the Murchison Widefield Array

In previous chapters, we have investigated various analysis techniques and algorithms for extracting cosmological signals from redshifted 21 cm data. We now apply many of these methods to real data. In particular, we use data taken from the Murchison Widefield Array (MWA) to place an upper limit on the EoR power spectrum. Such an upper limit is of course interesting in its own right, but also allows us to demonstrate the feasibility of the methods that we have proposed in this thesis.

The rest of this chapter is organized as follows. In Section 8.1 we describe the observations that went into the power spectrum measurement. Section 8.2 focuses on the imaging portion of the data analysis pipeline. After having gone through this pipeline, the data is in the form of images of the sky at various frequencies, which is precisely the format needed for the quadratic estimator power spectrum formalism of Chapter 6. In Section 8.3 we describe the minor modifications that were made to the formalism for the purposes of this chapter. Our results are presented in Section 8.4, and we summarize them in Section 8.5.

8.1 Observations

Data was taken[1] using the MWA in its 32-tile prototype form. As mentioned in Chapter 1, each tile of the MWA consists of 16 dual-polarization dipole antennas whose signals are combined ("beam-formed") by analog means. The resulting primary beam can be coarsely steered by a set of discrete delay-lines, and has full-width-half-maximum of approximately $25^\circ \left(\frac{\nu}{150\,\mathrm{MHz}}\right)^{-1}$. Importantly, different beam-former settings (*i.e.* primary beam pointings in different directions) give rise to different primary beam shapes, which must be taken into account in the data analysis.

Over two-weeks in March 2010, a series of short drift scans were performed to image a field centered at RA(J2000) $= 10^h\,20^m\,0^s$ and Dec.(J2000) $= -10^\circ\,0'\,0''$. Typically, a single beam-former setting would be used to image the field in a series of ~ 1 second-long snapshots. These snapshots would be taken over a 5 minute interval, at which point the beam-former settings would be changed to coarsely track the field. For the results in this chapter, data from a total of 60 such intervals were used, giving a total of 5 hours of integration on the field.

The MWA has an instantaneous bandwidth of 30.72 MHz, consisting of 768 channels each with a channel width of 40 kHz. For the observations relevant to this chapter, the observing band was centered at 154.24 MHz. Having a relatively wide frequency range of ~ 30 MHz is useful for the calibration purposes of the next section, but in the actual power spectrum estimation, we use only data from 139.065 MHz to 140.665 MHz for reasons pertaining to computational cost.

8.2 Calibration and Imaging

Since the power spectrum algorithms described in Chapter 6 require sky maps as input, the first step in our data analysis pipeline is to calibrate our data and to make maps at various frequencies.

For every frequency channel, the dataset consists of 2080 correlations per second ($64 \times (64 + 1)/2$ correlations between antenna tiles, where $64 = 2$ polarizations $\times$ 32 tiles). As an initial calibration step, auto-correlations were examined separately for each time and each channel. Any antenna tile whose auto-correlation fluctuated by more than $\sim 3\sigma$ would have all of its correlations discarded for the particular time

[1]The hard work of taking the data with the instrument and the subsequent reduction of data from its correlator output to sky images was performed by Christopher Williams. We include Sections 8.1 and 8.2 for completeness only, and the contributions of this author are limited to the data analysis in Section 8.3 and beyond.

and frequency. This procedure was imposed to guard against possible systematics such as radio frequency interference.

Following this, a bright point source calibration of the array was performed[2]. The bright radio source Hydra A ($\sim 600\,\mathrm{Jy}$ for the MWA) conveniently lies within the field at a location 15° from the field center, facilitating the process. Based on cataloged values for the brightness of Hydra A, ideal complex visibilities (*i.e.* correlations between antenna tiles) were theoretically predicted for the MWA. Calibration parameters were then computed using a least-squares fit of the observed visibilities to the theoretical ones. In addition to the calibration parameters, this process also yields a prediction for the contribution of Hydra A to our data. Hydra A can then be directly subtracted, and this was performed for all the data used in this chapter.

Following Hydra A's subtraction, a time series of sky maps (60 snapshots, each integrated for 5 minutes) of both polarizations were constructed from the calibrated visibilities for all 768 frequency channels. This was done by gridding the visibilities onto the uv-plane with inverse variance weighting[3], then Fourier transforming and adjusting for wide-field deviations (from a pure Fourier transform) using the w-projection algorithm (Cornwell et al., 2005). The separate polarization maps were then combined to give total intensity maps.

Following the production of time-ordered data detailed above, different snapshot maps $D_i(\widehat{\mathbf{r}})$ were combined by appropriately weighting each map according to its calculated primary beam $B_i^{prim}(\widehat{\mathbf{r}})$ and the spatial variance σ_i:

$$T(\widehat{\mathbf{r}}) = \frac{\sum_i \frac{D_i(\widehat{\mathbf{r}})B_i^{prim}(\widehat{\mathbf{r}})}{\sigma_i^2(\widehat{\mathbf{r}})}}{\sum_i \frac{[B_i^{prim}(\widehat{\mathbf{r}})]^2}{\sigma_i^2(\widehat{\mathbf{r}})}}. \tag{8.1}$$

Combining the maps in this manner has the property of maximizing the signal-to-noise ratio of the final map (Williams et al., 2012). From the time series maps, two separate maps were constructed, one using odd data samples and one using even data samples. Having these separate maps allows for cross-correlation studies, which we discuss in more detail in the next section. It should be noted that at this point the maps were dirty maps, in the sense that they represented an estimate not of the true sky, but the sky convolved with the synthesized beam of the instrument.

[2]The MWA, at least in its 32 tile configuration, is not designed to give a particularly redundant set of baselines. While we showed in Chapter 2 that slight off-redundancies can be accounted for in redundant calibration algorithms, we choose in this chapter to be conservative and use traditional calibration techniques instead.

[3]As conventionally defined in the Common Astronomy Software Applications (CASA) package provided by the National Radio Astronomical Observatory (NRAO) http://casa.nrao.edu/

Following mapmaking, another round of calibration and flagging was performed. Three bright point sources were identified and matched to catalog data. The spectra of the sources were simultaneously fit to power laws, and the maps at each frequency were scaled to yield best fits. The spatial variances of the maps were computed frequency-by-frequency, and frequencies with large variances (~ 3 to 5σ) were discarded.

As a final step, the maps were reduced to a computationally manageable size by spatially averaging pixels. In addition, in the following sections we analyze only a subset of the frequencies, from 139.065 MHz to 140.665 MHz.

8.3 Power Spectrum Estimation

Having obtained a series of sky maps at various frequencies (a "data cube", with the sides of the cube being θ_x, θ_y, and ν), we are now in a position to estimate power spectra. To do so, we will employ the quadratic estimator formalism of Chapter 6, which we now briefly review. Following the general review, we will highlight the minor changes that were implemented when apply the formalism to real data.

The data cube can be thought of abstractly as a series of sky temperatures stored in a single data vector $\mathbf{x}$. From this vector, a series of bandpower estimators $\widehat{p}_\alpha$ can be constructed according to the prescription

$$\widehat{p}_\alpha = \mathbf{x}^t \mathbf{E}^\alpha \mathbf{x} - b_\alpha, \tag{8.2}$$

where each value of α corresponds to a different bin in Fourier space, $\mathbf{E}^\alpha$ denotes a series of matrices (one for each value of α) that encode the foreground subtraction, Fourier transform, and binning steps. The b_α term is included to ensure that our bandpowers are unbiased, and takes the form

$$b_\alpha = \mathrm{tr}\left[\mathbf{N}\mathbf{E}^\alpha\right], \tag{8.3}$$

where $\mathbf{N}$ is the total noise covariance (*i.e.* the covariance of everything but the signal). The resulting bandpower estimates $\widehat{p}_\alpha$ are related to the true bandpowers p_α by a window function matrix:

$$\widehat{p}_\alpha = \sum_{\alpha\beta} \mathbf{W}_{\alpha\beta} p_\alpha, \tag{8.4}$$

where

$$\mathbf{W}_{\alpha\beta} = \mathrm{tr}[\mathbf{C}_{,\beta}\mathbf{E}^{\alpha}], \tag{8.5}$$

with $\mathbf{C}_{,\beta}$ being the Fourier transform and binning matrix (see Section 8.3.6 for explicit expressions).

Different choices of $\mathbf{E}^{\alpha}$ give power spectrum estimates with different statistical properties. This can be quantified by computing the measurement covariance of the bandpower estimates:

$$\mathbf{\Sigma}_{\alpha\beta} = 2\mathrm{tr}\left[\mathbf{C}_{,\alpha}\mathbf{C}\mathbf{C}_{,\beta}\mathbf{C}\right], \tag{8.6}$$

where $\mathbf{C}$ is the total covariance matrix of the data.

So far in this section, our methods have been identical to those of Chapter 6. In the subsections that follow, we will discuss the minor variations that we implement in this chapter.

8.3.1 Choice of $\mathbf{E}^{\alpha}$

In a similar manner to what we did in Chapter 6, we select

$$\mathbf{E}^{\alpha} = \frac{1}{2}\sum_{\beta} M_{\alpha\beta}\mathbf{C}^{-1}\mathbf{C}_{,\beta}\mathbf{C}^{-1}, \tag{8.7}$$

where $\mathbf{M}$ is an $n_{bands} \times n_{bands}$ matrix (where n_{bands} is the number of bins that k-space is partitioned into for our power spectrum estimation). If we define the Fisher matrix $\mathbf{F}$:

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{tr}\left[\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta}\mathbf{C}^{-1}\right], \tag{8.8}$$

then with Equation 8.7, the expressions for the window function matrix and the measurement covariance matrix become

$$\mathbf{W} = \mathbf{M}\mathbf{F} \tag{8.9}$$

$$\mathbf{\Sigma} = \mathbf{M}\mathbf{F}\mathbf{M}^{t}. \tag{8.10}$$

In Chapter 6 we saw that if $\mathbf{M}$ is taken to be diagonal and proportional to the diagonal elements of $\mathbf{F}$, the vertical error bars on $\widehat{p}_{\alpha}$ are minimized. There we took the constant of proportionality to be unity, so that $\mathbf{M}_{\alpha\beta}$ was exactly $\delta_{\alpha\beta}\mathbf{F}_{\alpha\alpha}$. In this chapter, we choose the constant differently, so that the rows of the window function

sum to unity:

$$1 = \sum_j \mathbf{W}_{ij} = \sum_j (\mathbf{MF})_{ij}. \tag{8.11}$$

Doing so facilitates our interpretation of the band power estimates, for then Equation 8.4 says that our estimates can be considered weighted averages of the true power spectrum. We note, however, that ultimately our choice for $\mathbf{M}$ is irrelevant for the purposes of fitting theoretical models, provided it is invertible. With an invertible matrix, multiplying by the matrix results in no change in information content, and the constraining power a data set is unchanged.

8.3.2 What is being measured

In Equations 8.2, 8.3, and 8.7, two different covariances ($\mathbf{C}$ and $\mathbf{N}$) enter. The matrix $\mathbf{C}$ is a covariance matrix of the data vector $\mathbf{x}$, which includes all sources of signal in our data, whether from the cosmological signal, foregrounds, or instrumental noise. On the other hand, $\mathbf{N}$ contains *only* the covariance of contaminants, *i.e.* everything but the signal we are interested in.

Precisely which portions of the total covariance are to be considered contaminants depends on what one is attempting to measure. In Chapter 6 we envisioned an estimation of the *cosmological* 21 cm signal, which meant that our contaminant covariance $\mathbf{N}$ contained both foregrounds and instrumental noise. This is of course the ultimate goal of an instrument like the MWA. However, to perform such a measurement requires that one has an accurate foreground model, for otherwise it would be difficult to definitively say that $\mathbf{N}$ contains only contaminants. To avoid such complications, we instead choose to simply estimate the power spectrum of the total sky signal. The contaminant covariance then just consists of instrumental noise. This allows us to confidently establish an upper limit on the cosmological power spectrum without the need for an exquisite foreground model, since the foreground contributions are necessarily positive and would therefore only make an upper limit more conservative.

8.3.3 Cross-power spectra

Having elected to estimate the total sky power spectrum, our expression for the bias term reduces to

$$b_\alpha = \mathrm{tr}\left[\mathbf{N}_{inst}\mathbf{E}^\alpha\right], \tag{8.12}$$

where $\mathbf{N}_{inst}$ is the instrumental noise covariance. In principle, this can computed and subtracted in our estimator for $\widehat{\mathbf{p}}_\alpha$ (as the prescription in Equation 8.2 requires).

However, if this instrumental noise bias is non-negligible, one may end up in the numerically perilous situation of having to subtract two large numbers from each other.

Fortunately, the noise bias term can be eliminated entirely by measuring a *cross*-power spectrum instead of an auto-power spectrum. The idea is that instrumental noise fluctuations are uncorrelated with time, and thus a cross-correlation between data samples from two different times should have no average noise contribution, eliminating the noise bias term in our power spectrum estimator while preserving the sky signal, which is constant in time. This technique was successfully implemented by the WMAP team in their measurement of the CMB power spectrum (Hinshaw & et. al. [WMAP collaboration], 2007), and in our pipeline we find the cross-power spectrum of the two time-interleaved maps discussed in Section 8.2. The quadratic estimator formalism remains much the same, and our new estimator for the bandpowers becomes

$$\widehat{p}_\alpha = \mathbf{x}_1^t \mathbf{E}^\alpha \mathbf{x}_2, \tag{8.13}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the two maps formed from interleaved time-samples.

Note that while the noise bias has vanished from our estimator, our error bars will continue to have a noise contribution to them. In other words, the copy of the total covariance $\mathbf{C}$ in $\mathbf{E}^\alpha$ continues to represent the total sky covariance plus instrumental noise covariance. Our final measurement covariance (given by Equation 8.6) remains the same with the cross-correlation.

8.3.4 Modeling the total covariance

Given the considerations of the previous two subsections, we only need to model the total covariance $\mathbf{C}$. We do so by looking at the data itself, since it is unclear as to whether the models used in previous chapters are sufficiently accurate. (Indeed, one of the goals of the first generation of instruments is to provide better foreground models at low frequencies).

We consider the total covariance to be comprised of a term dominated by foregrounds ($\mathbf{C}_{fg}$) and a term dominated by instrumental noise ($\mathbf{N}_{inst}$):

$$\mathbf{C} = \mathbf{C}_{fg} + \mathbf{N}_{inst}. \tag{8.14}$$

From Chapter 6, we expect the main effects of the foreground covariance to be in the frequency direction. We therefore make the approximation that $\mathbf{C}_{fg}$ is block diagonal,

with correlations only between different frequencies of the same pixel, and not between different pixels on the sky. To estimate this covariance, we use the data and apply a procedure similar to that used in Chapter 3 and estimate a frequency-frequency covariance by averaging over spatial pixels:

$$\langle x(\nu)x(\nu')\rangle = \frac{1}{\Omega}\int T(\hat{\mathbf{r}},\nu)T(\hat{\mathbf{r}},\nu')d\Omega, \tag{8.15}$$

where Ω is the field of view of the observational field. The frequency-frequency covariance then forms the (identical) blocks of our block diagonal foreground covariance $\mathbf{C}_{fg}$.

To estimate the instrumental noise covariance $\mathbf{N}_{fg}$, we difference two high-resolution versions of our time-interleaved maps. This has the effect of removing any signals on the sky, leaving what should in principle be a map that contains only instrumental noise. Averaging over small cells of this map allows us to estimate the mean-square noise σ^2 in each spatial pixel of the map, providing us with the diagonal entries of $\mathbf{N}_{fg}$. The off-diagonal entries encode the spatial correlations between two cells, and are obtained by multiplying together σ^2 and two copies of the synthesized beam, one centered at each of the cells:

$$\mathbf{N}^{ij}_{inst} = \sum_m \sigma^2_m B_i(\widehat{\mathbf{r}}_m)B_j(\widehat{\mathbf{r}}_m), \tag{8.16}$$

where $B_i(\hat{\mathbf{r}})$ is the synthesized beam pattern in direction $\widehat{\mathbf{r}}$ when centered on the i^{th} pixel.

With $\mathbf{C}_{fg}$ and $\mathbf{N}_{inst}$, we can form the full covariance $\mathbf{C}$. This matrix is an $n_{vox} \times n_{vox}$ matrix, where n_{vox} is the number of voxels in the data cube. It is therefore quite unwieldy to examine on an element-by-element basis. However, we can gain intuition for the behavior of the system by averaging the covariance $\mathbf{C}$ over frequencies or over spatial pixels. The resulting covariances are reduced spatial-spatial and frequency-frequency covariances respectively. Performing an eigenvalue decomposition of the frequency-frequency covariance, we see in Figure 8-1 that the theoretical expectations of Chapter 3 are borne out by our data-derived covariance—the eigenvalues fall off exponentially and then hit a noise floor, suggesting that the foregrounds can be sequestered to a small number of modes. The spectral eigenmodes, shown in Figure 8-2, also bear a strong resemblance to their theoretical counterparts. The first three eigenmodes are seen to be roughly of the constant, linear, and quadratic forms of Chapter 3, albeit with some extra noise. The fourth eigenmode appears to be pure

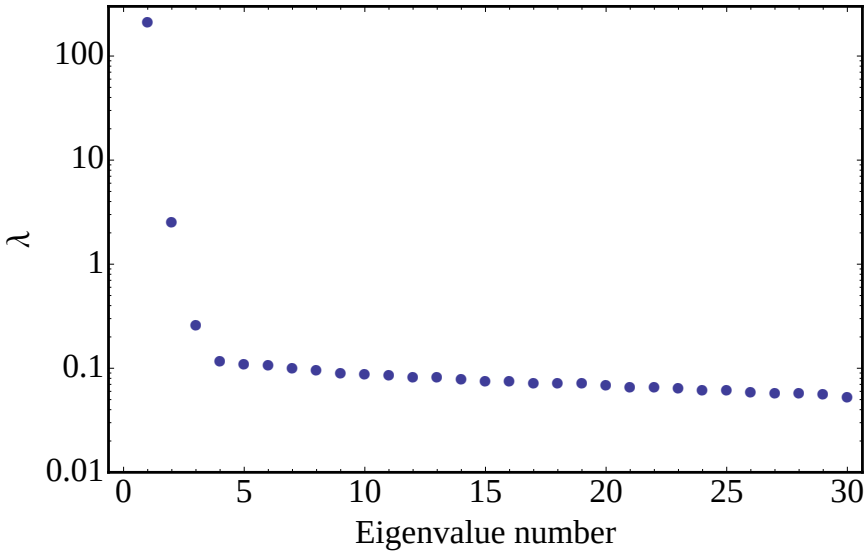

Figure 8-1: Eigenvalues of the reduced frequency-frequency covariance matrix obtained by spatially averaging over the full modeled covariance matrix.

noise, which is not surprising since Figure 8-1 suggests that we have hit the noise floor by then. Note that it is not surprising that the noise floor is reached sooner here than in the theoretical studies, for recall that in Chapter 3 we assumed much more integration time (1000 hrs), which would give a lower noise floor.

Similar computations of the eigenvalues and eigenmodes of the reduced spatial-spatial covariance yield essentially a white-noise eigenvalue spectrum and eigenmodes that are mostly noise maps. The results are more illuminating if we plot a single row of the spatial covariance (which, remember, has dimensions of a map). This can be seen in Figure 8-3, and what is recovered is a map of the square of the synthesized beam of the instrument.

8.3.5 Deconvolution

Implicit in the formalism of Chapter 6 was the assumption that the input sky maps had already been deconvolved. In other words, to estimate the true power spectrum of the sky, it is necessary to first undo the blurring effects of the synthesized beam. Without this step, an estimation of the power spectrum can still be made, but the result formally becomes the power spectrum of the sky as seen by the instrument, not the power spectrum of the true sky.

With a radio interferometer, it is formally impossible to fully deconvolve the synthesized beam from the data, since there are many parts of the uv-plane that are not sampled. However, with some approximations it is possible to remove the effects of

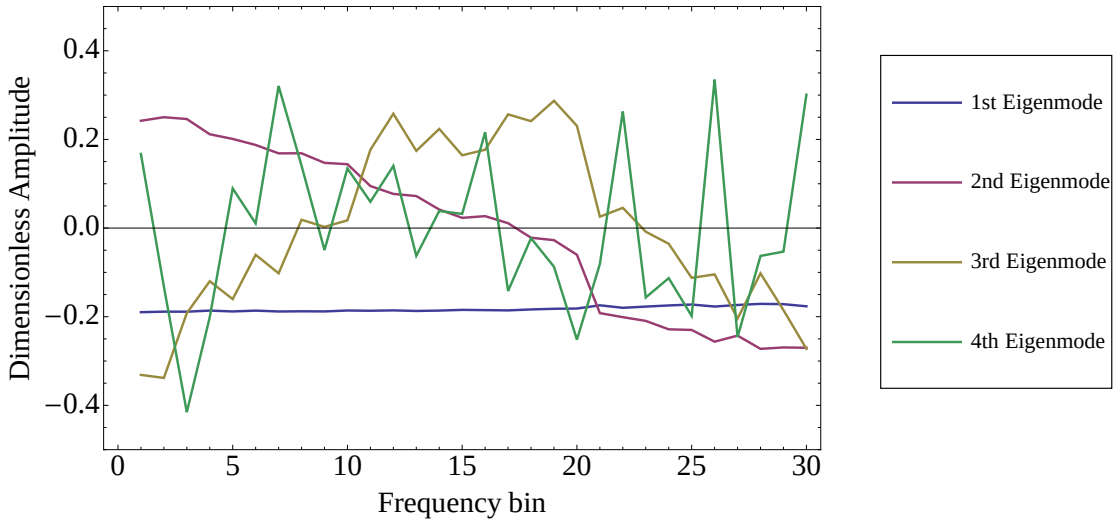

Figure 8-2: First four eigenvectors of the reduced frequency-frequency covariance matrix obtained by spatially averaging over the full modeled covariance matrix.

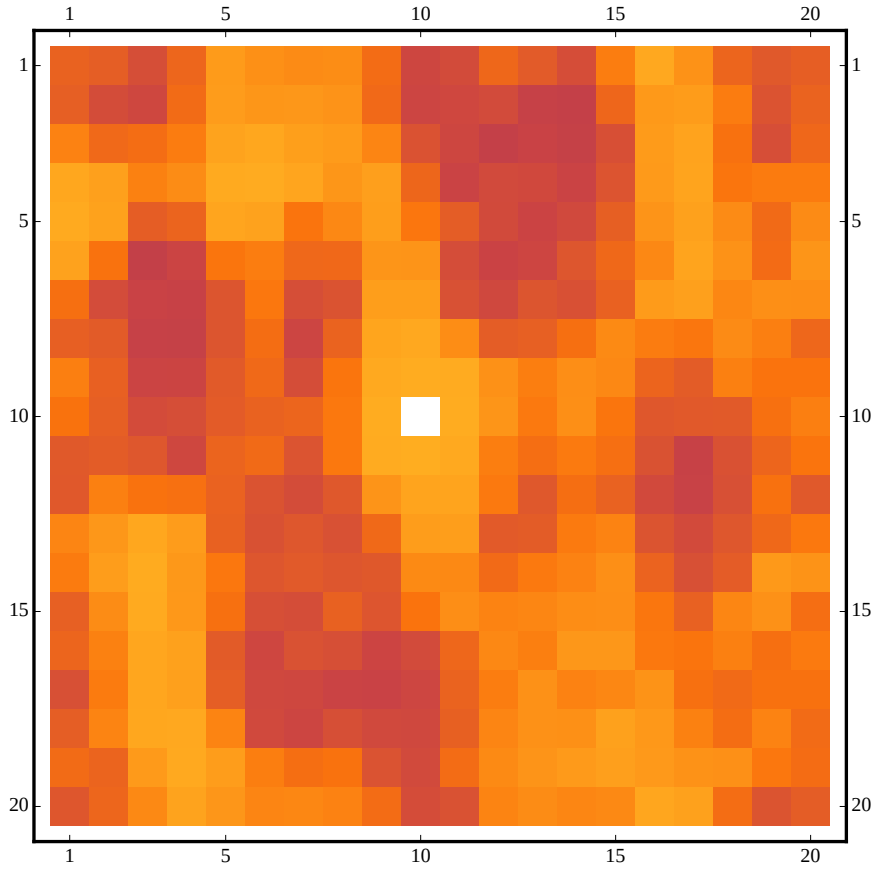

Figure 8-3: A single row of the reduced spatial-spatial covariance matrix obtained by averaging over the frequencies of the full modeled covariance matrix.

the beam from the final power spectrum. To see this, consider our power spectrum estimator for a dirty map (*i.e.* one whose synthesized beam has not been removed):

$$\widehat{\mathbf{p}}_\alpha^{dirty} = \mathbf{x}_1^t \mathbf{B}^t \mathbf{E}^\alpha \mathbf{B} \mathbf{x}_2, \tag{8.17}$$

where $\mathbf{B}$ is a matrix describing the convoluting action of a synthesized beam, and we have simply made the substitution $\mathbf{x} \to \mathbf{B}\mathbf{x}$. Now, $\mathbf{E}^\alpha$ is a matrix that is diagonal in Fourier space, so if $\mathbf{B}$ were also diagonal there, the factors of $\mathbf{B}$ would factor out of our expression for $\widehat{\mathbf{p}}_\alpha^{dirty}$. Deconvolving the power spectrum would then just amount to taking the Fourier transform of the synthesized beam, binning it appropriately, and dividing the dirty map power spectrum by its square.

Unfortunately, the synthesized beam matrix is *not* diagonal in a full, three-dimensional Fourier space. Instead, it is diagonal in Fourier modes perpendicular to the line-of-sight (*i.e.* in $k_\perp$) and in image-space parallel to the line-of-sight, since the synthesized beam smooths the maps frequency-by-frequency. In principle, then, one must correct for the beam prior to Fourier transforming in the frequency direction. For the purposes of this thesis, however, we notice that the fractional change in the synthesized beam over our bandwidth is sub-percent level, which means we can approximate the beam as being achromatic over our band (but see the Section 8.4 for some chromatic effects that appear in our data). With this, our correction can be made to the cylindrical power spectrum on the $k_\perp$-$k_\parallel$ plane, since it would only require dividing by a function of $k_\perp$, and so the Fourier transform in the line-of-sight direction can be performed either before or after.

8.3.6 Cylindrical and spherical power spectra

As we discussed in Chapter 6, the presence of redshift space distortions and systematics such as foregrounds mean that it is often prudent to form a cylindrically binned power spectrum $P(k_\perp, k_\parallel)$ as an initial step. Doing so means using the cylindrically binned expression for $\mathbf{C}_{,\alpha}$ discussed in Chapter 6:

$$\mathbf{C}_{,\alpha}^{ij} = \frac{(2\pi^2)^{-1}}{(r_{ij}^\perp)^2 r_{ij}^\parallel} \left(\sin k_b^\parallel r_{ij}^\parallel - \sin k_{b-1}^\parallel r_{ij}^\parallel \right) \left[k_a^\perp r_{ij}^\perp J_1(k_a^\perp r_{ij}^\perp) - k_{a-1}^\perp r_{ij}^\perp J_1(k_{a-1}^\perp r_{ij}^\perp) \right], \tag{8.18}$$

where $k_a^\perp$ and $k_b^\perp$ delimit the $k_\perp$ extent of the α^{th} bin on the $k_\perp$-$k_\parallel$ plane, $k_a^\parallel$ and $k_b^\parallel$ delimit the $k_\parallel$ extent, J_1 denotes the first cylindrical Bessel function, $r_{ij}^\perp$ denotes the distance between i^{th} and j^{th} voxels projected perpendicular to the line-of-sight,

and $r_{ij}^{\parallel}$ denotes the distance between i^{th} and j^{th} voxels along the line-of-sight. In the following section, we will use the cylindrical power spectrum to understand the systematics and error properties of our measurement.

Ultimately, however, it is the spherically-binned power spectrum $P(k)$ that is of most interest for comparisons with theoretical models. If the beam corrections discussed in Section 8.3.5 were not necessary, computing a spherically-binned power spectrum would be straightforward. One would simply use a different expression for $\mathbf{C}_{,\alpha}$ in our formalism:

$$\mathbf{C}_{,\alpha}^{ij} = \frac{1}{2\pi^2 r^3} \left(\sin k_b r - \sin k_a r - k_b r \cos kr + k_a r \cos k_a r\right). \tag{8.19}$$

As mentioned above, however, one cannot go directly to the spherical power spectrum because the beam corrections need to be performed on the $k_\perp$-$k_\parallel$ plane. We therefore first compute a cylindrical power spectrum, and then re-bin the data along contours of constant $k = \sqrt{k_\perp^2 + k_\parallel^2}$ to estimate a spherical power spectrum. For numerical stability, we perform this binning in a uniformly-weighted fashion, except we exclude regions on the $k_\perp$-$k_\parallel$ plane that are clearly foreground contaminated.

8.4 Results and Discussion

We now present the results of our power spectrum estimation. Instead of plotting the cylindrical power spectrum directly, we plot the quantity

$$\Delta(k_\perp, k_\parallel) \equiv \sqrt{\frac{k_\perp^2 k_\parallel}{2\pi^2} P(k_\perp, k_\parallel)}, \tag{8.20}$$

which, as the contribution to the temperature field variance per logarithmic interval in k, has units of temperature, making the results more intuitively interpretable. In Figure 8-4 we show our minimum-variance quadratic estimator measurement of the dirty map cylindrical cross-power spectrum. The discreteness of the plot is due to the coarseness of the Fourier-space bins used in the computation (10 bins in the $k_\perp$ direction, spanning $k_\perp = 0.005\,\mathrm{Mpc}^{-1}$ to $0.1\,\mathrm{Mpc}^{-1}$; 18 bins in the $k_\parallel = 0.05\,\mathrm{Mpc}^{-1}$ to $2.0\,\mathrm{Mpc}^{-1}$). Going to a finer set of bins is of course possible, but computationally difficult. It is for similar computational reasons that only $12,000$ voxels were used in this analysis (400 pixels $\times$ 30 frequencies as described in Section 8.2). The white regions on the figure are regions that would be accessible if the full data set were used.

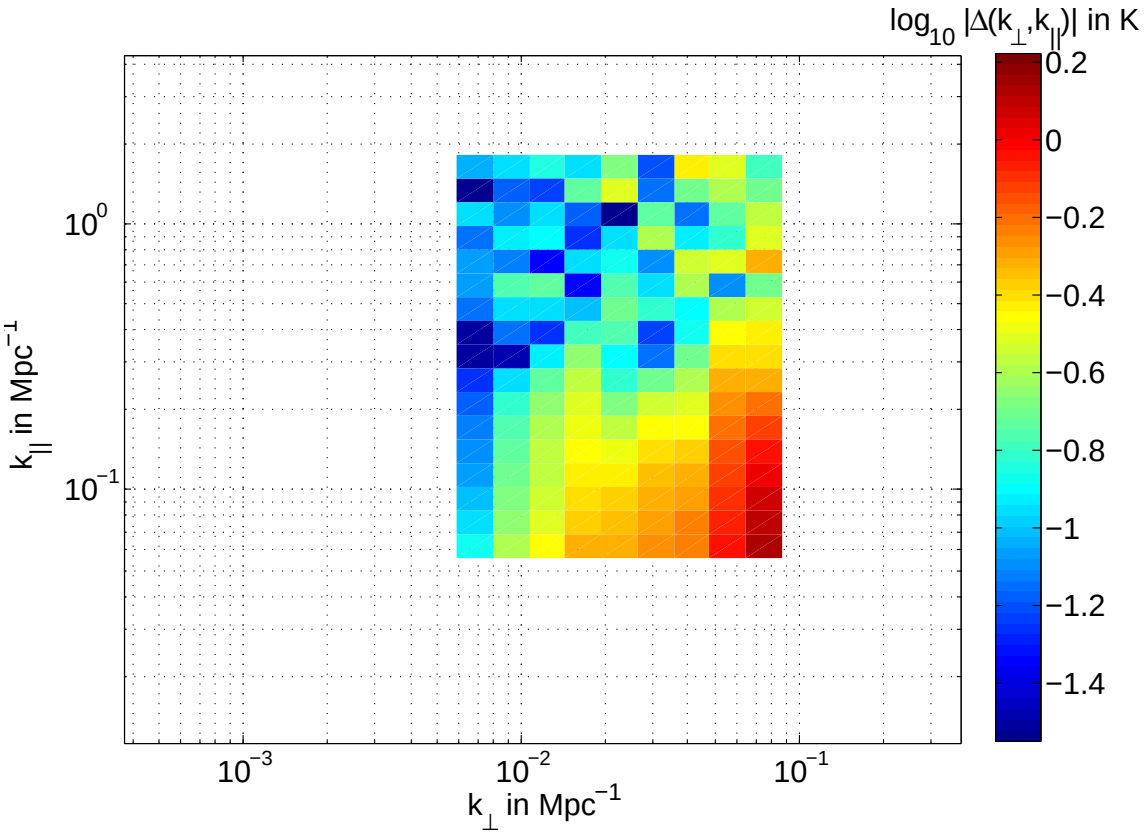

Figure 8-4: A minimum-variance quadratic estimator measurement of the dirty map cylindrical cross-power spectrum. The absolute value of the cross-power spectrum has been taken before forming $\Delta(k_\perp, k_\parallel)$ in this and the following figures, since cross-correlations can be negative in low signal-to-noise regions.

As a comparison, one can form a similar power spectrum by performing Fast Fourier Transforms (FFT) and binning the results in annuli. The result is shown in Figure 8-5 spanning the same region in k-space. The advantage of using FFTs is that they are computationally quick, allowing the full data sample to be used, increasing both the accessible range of Fourier modes and the raw sensitivity. The disadvantage of using FFTs is that assumptions must be made about the physical geometry of the data. In particular, the data are assumed to be rectilinear and contiguous, and neither assumption is strictly correct. The rectilinearity assumption is violated because lines of sight diverge in physical space, and because equally spaced frequency channels do not map to equally spaced co-moving line-of-sight cells. As for contiguity, the need to discard flagged frequency channels (as discussed in Section 8.2) means that there are gaps in the data cube. To produce Figure 8-5, it was therefore necessary to interpolate between missing channels. With the quadratic estimators, a power spectrum can be produced exactly, regardless of the geometry of the input data.

A further difference between the FFT plot of Figure 8-5 and the quadratic estimator plot of Figure 8-4 is in the weighting of the data, which is non-optimal in the FFT method. The form of $\mathbf{E}^\alpha$ detailed above involves a $\mathbf{C}^{-1}$ inverse variance weighting of the data. As discussed in Chapter 6, this has the effect of downweighting foreground and noise contaminated modes in an optimal way. Figure 8-5 performs no such downweighting, and thus one sees what is almost certainly strong foreground contamination at low $k_\parallel$, consistent with theoretical expectations. The drop-off to the left of the red, heavily contaminated region is due to the fact that our array has a minimum baseline length of about $\sim 10\,\mathrm{m}$, which maps to a minimum $k_\perp$. The drop-off to the right, similarly, is due to the maximum baseline of about $\sim 340\,\mathrm{m}$.

Thus, to properly compare the FFT plot to the quadratic estimator plot, it is necessary to perform some downweighting of contaminated modes before performing the FFT. This was done to produce Figure 8-6, where foreground eigenmodes were identified using the methods of Chapter 3 and Figure 8-2. In other words, eigenvalues for the spectral covariance were plotted (for the full data set used to generate the FFT plots), and a noise floor was seen to be reached after 50 eigenvalues. Correspondingly, the first 50 eigenmodes were projected out of the data. Comparing Figures 8-5 and 8-6, one sees that such a projection affects mainly the low $k_\parallel$ modes, but there is still a non-negligible effect at high $k_\parallel$. This is consistent with the theoretical work in Chapter 3, where we saw that with our phenomenological foreground model, the foreground eigenmodes were qualitatively similar but not identical to Fourier modes. Taking out the first 50 modes thus mainly erases power from the lower parts of the

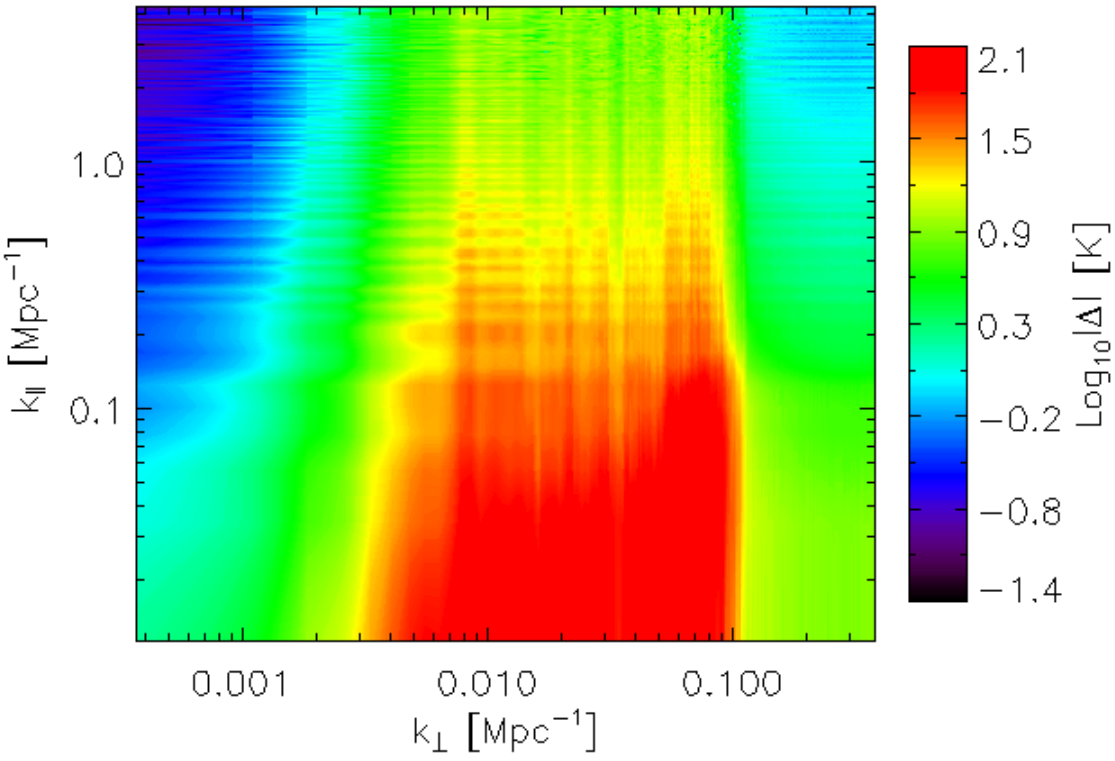

Figure 8-5: An FFT-based measurement of the dirty map cylindrical cross-power spectrum.

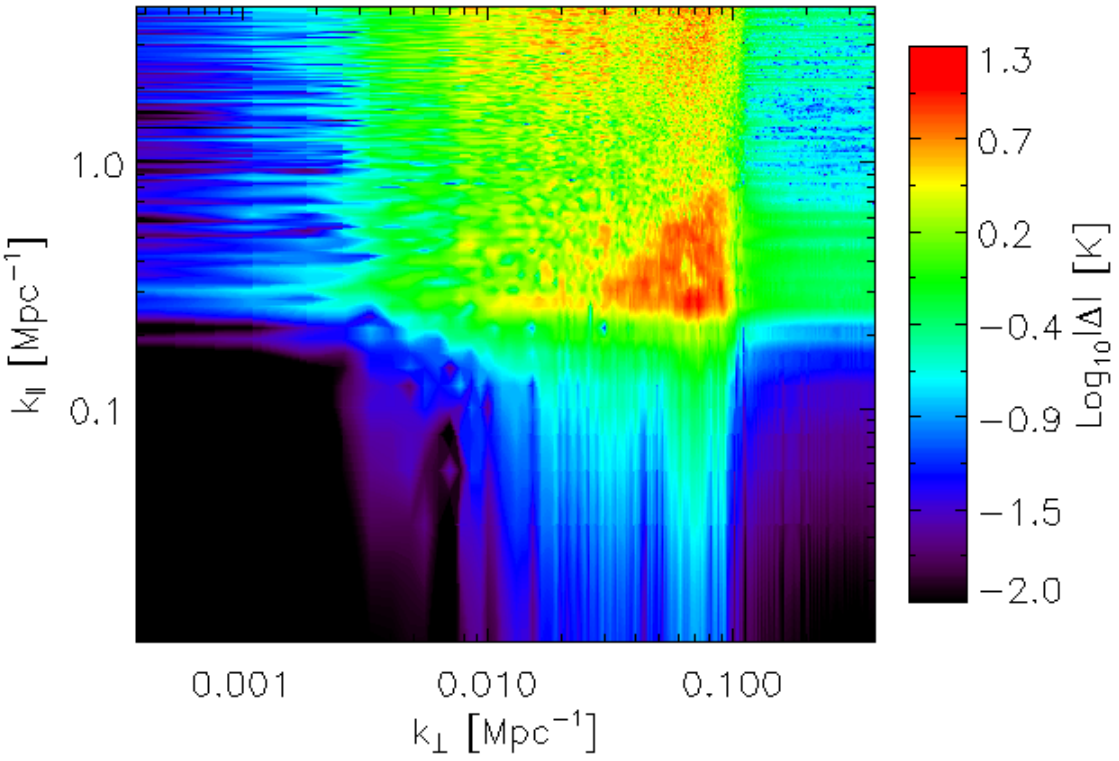

Figure 8-6: An FFT-based measurement of the dirty map cylindrical cross-power spectrum, with the first 50 frequency eigenmodes projected out.

Fourier plane, but the effects are not exclusive to that region.

Figures 8-4 and 8-6 are our best measurements of the dirty map cylindrical power spectrum, and it is reassuring that they appear to be qualitatively similar in the region where they overlap. One should not expect precise qualitative agreement for several reasons. First are the geometric discrepancies described above. In addition, the different amounts of data (albeit from the same master data set) went into producing the two plots, so the sensitivities are formally different. Finally, the quadratic estimator plot employs inverse variance weighting, which differs from the eigenmode truncation of the FFT plots.

Focusing on the region of overlap, the predominant structure in both plots appears to be a triangular "wedge" with higher temperature and a cooler (and therefore ultimately cleaner) region above the wedge. The wedge has been seen in previous measurement simulations (Datta et al., 2010; Vedantham et al., 2012; Morales et al., 2012; Trott & Wayth, 2012), and is attributed to the response of the instruments' chromatic beam to bright point source foreground contaminations[4]. Essentially, a

[4]There were no signs of this wedge in the results of Chapter 6 because those computations did

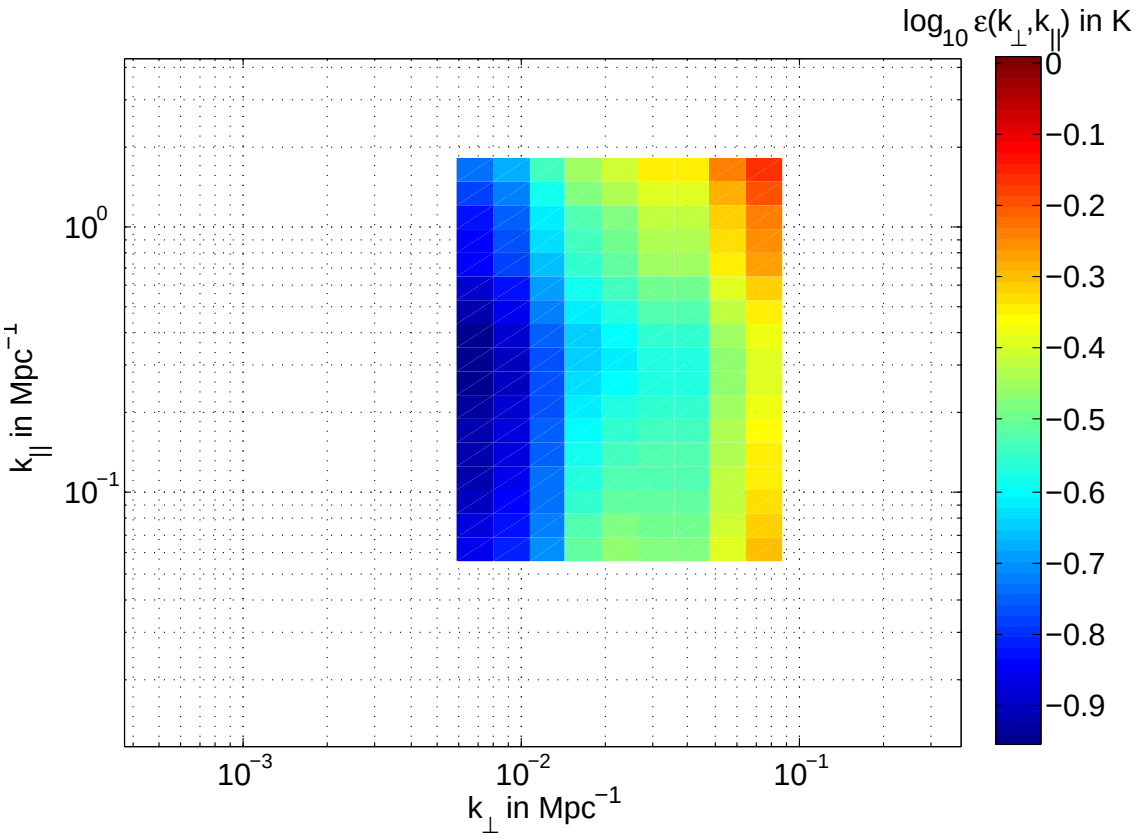

Figure 8-7: Error bars on the quadratic estimator measurement of the dirty map cylindrical cross-power spectrum.

chromatic beam imprints extra frequency structure into a point source, producing power higher up in the $k_\perp$-$k_\parallel$ plane than would otherwise be expected from a smooth spectral source (which are confined to the low $k_\parallel$ regions). The effect is more pronounced at higher $k_\perp$ because higher $k_\perp$ values are sampled by longer baselines, whose phases dilate more quickly with frequency. The result is the characteristic wedge shape, seen clearly in observational data here.

In addition to the bandpower estimates, one can also compute error bars in the quadratic estimator formalism. The results are shown in Figure 8-7. The rather vertical orientation of the plot suggests that there is a strong component instrumental noise components in the errors. However, the structures on the plot do have some $k_\parallel$ dependence to them, suggesting that there is some residual foreground contamination.

Taking the ratio of Figures 8-4 and 8-7, we arrive at Figure 8-8, which is a rough signal-to-noise ratio[5]. The most prominent feature is the bottom right corner, which

not include bright point sources

[5]Note that here, the "noise" includes all sources of error, such as the cosmic variance of the signal.

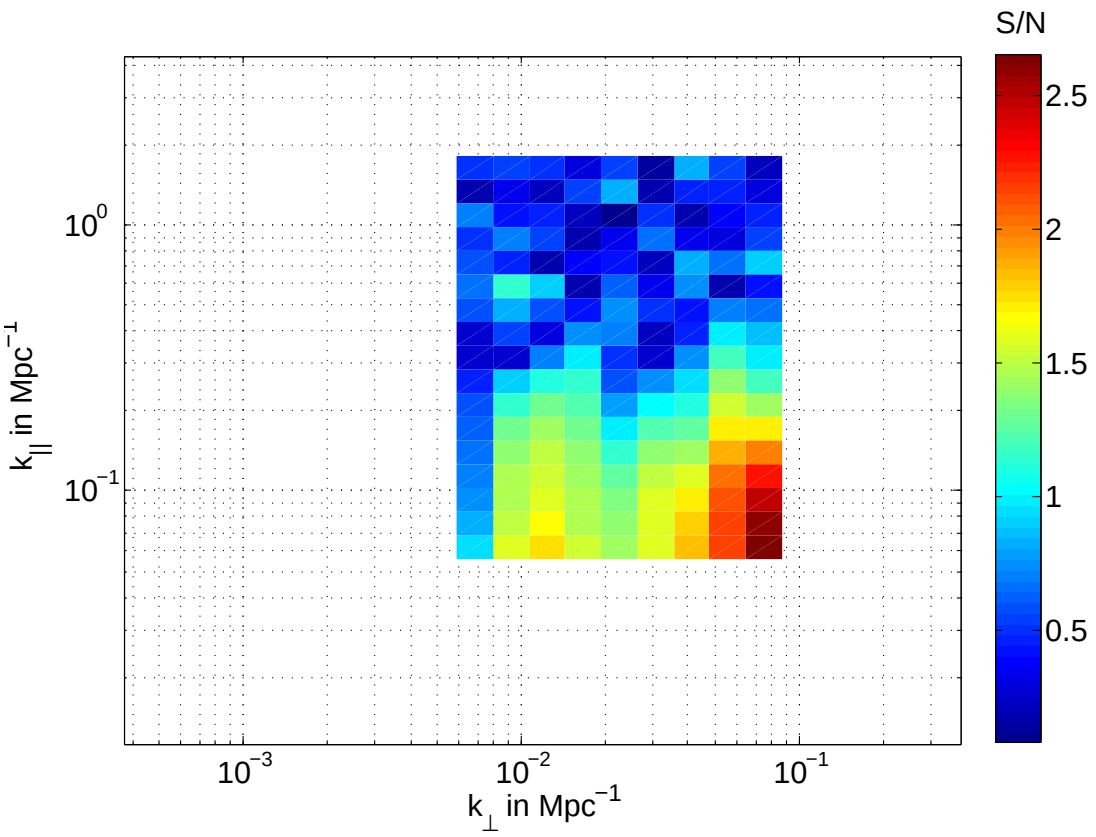

Figure 8-8: The rough signal-to-noise ratio, computed by taking the ratio of Figures 8-4 and 8-7.

has a high signal-to-noise, confirming that have indeed observed the theoretically predicted wedge[6]. The lower signal-to-noise at higher $k_{||}$ is not a cause for concern. Since our current power spectrum estimates are estimates of the total emission on the sky, a low power in those regions means that those parts of the $k_\perp$-$k_{||}$ plane are relatively free from foreground contamination, as is hoped. In addition, we have already argued that the error bars are likely to decrease with further integration, making the region a promising part of Fourier space for a future detection of the EoR cosmological signal.

Having understood many of the characteristics of our measurement using dirty map cylindrical power spectra, we now turn to producing a spherical power spectrum $P(k)$. Since the spherical power spectrum is a quantity of theoretical interest, we take

[6]Note that our error bars will not include a term corresponding to the power of the wedge itself (*i.e.* a "cosmic variance" term if the wedge were considered a signal). This is because the wedge is expected to be due to bright point source contamination as view through a chromatic beam, and by construction our covariance modeling did not include such effects. The high signal-to-noise ratio in the region of the wedge is thus truly a detection of the wedge over other foreground contaminants and instrumental noise.

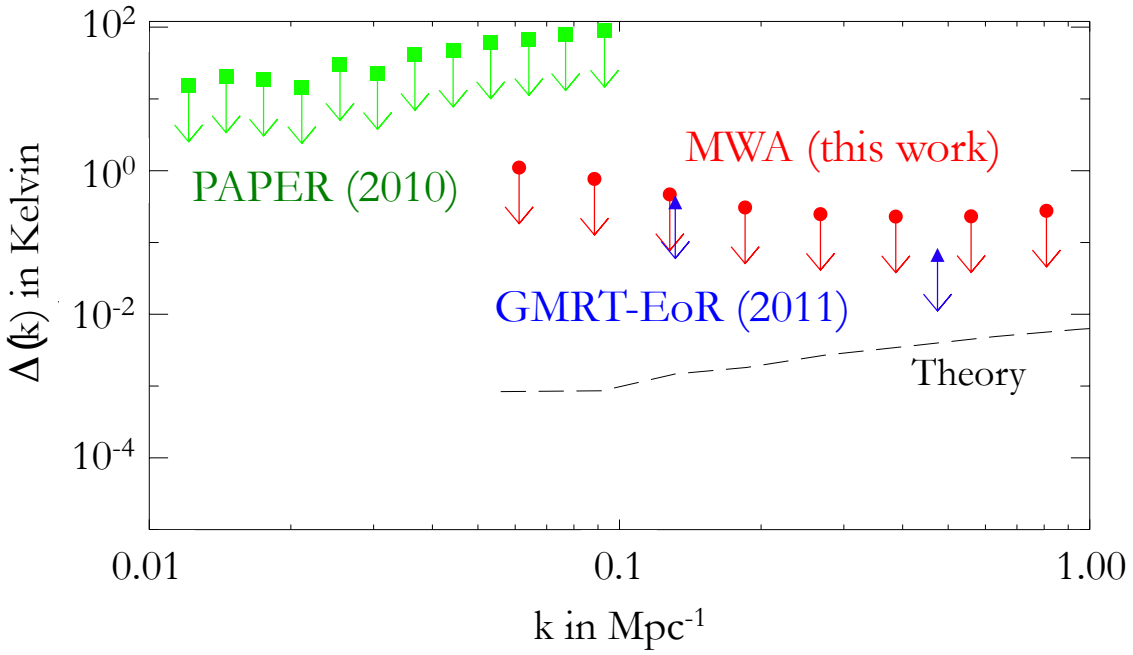

Figure 8-9: An upper limit on the spherically binned, deconvolved power spectrum, plotted as $\Delta(k) \equiv \sqrt{\frac{k^3}{2\pi^2}P(k)}$. This serves as an upper limit for the EoR power spectrum at $z = 9.1$. This work is denoted by the red circles, the PAPER limit by the green squares (Parsons et al., 2010), and the GMRT-EoR limit by the blue triangles (Paciga et al., 2010).

the extra step adjusting for the effective beam of the instrument in the final power spectrum estimate (as discussed in Section 8.3.5). The result is shown in Figure 8-9, where we have once again elected to plot the variance per logarithmic k:

$$\Delta(k) \equiv \sqrt{\frac{k^3}{2\pi^2}P(k)} \tag{8.21}$$

This serves as our upper limit on the EoR cosmological power spectrum, which ranges from 0.45 K at $k = 0.06\,\mathrm{Mpc}^{-1}$ to 3.8 K at $k = 0.81\,\mathrm{Mpc}^{-1}$. While this limit is several orders of magnitude above the theoretically predicted signal, it is lower than the one published by the PAPER experiment (Parsons et al., 2010), and competitive with the GMRT-EoR limit (Paciga et al., 2010), which used 10 times more integration time and vastly more collecting area. Our limit is thus an encouraging validation of the MWA and our data analysis methods.

8.5 Conclusion

In this chapter, we have established an observational upper limit on the EoR power spectrum at $\sim 140\,\mathrm{MHz}$, corresponding to $z \sim 9.1$. The upper limit was at the sub-Kelvin level. Using two independent pipelines, we also produced cylindrical power spectra of the dirty maps. A theoretically expected "wedge" from bright point source foregrounds and instrumental chromatic effects was observed, and a clear window on the $k_\perp$-$k_\parallel$ plane emerged as a promising location to look for the EoR in future measurements. Our current measurements are likely dominated by instrumental noise at high $k_\parallel$. This suggests that by simply integrating for longer and using more frequency slices from the data, one should be able to improve on the limits set by this work, which bodes well for the field of hydrogen cosmology as a whole.

Chapter 9

Conclusion

In this thesis, we have tackled a number of the challenges that stand between the theoretical promise and experimental reality of hydrogen cosmology. As we argued in Chapter 1, these challenges are quite considerable, given the exquisite sensitivity, calibration, and foreground subtraction requirements that are necessary to detect an extremely faint ($\sim 10\,\mathrm{mK}$) signal from beneath a set of bright foregrounds ($\sim$ 100's to 1000's of K). Following foreground subtraction, it is important to analyze the data in a way that is lossless and unbiased to give the best possible constraints on theoretical models.

In Chapter 2 we developed a method for calibrating radio interferometers with a large number of redundant baselines. We also quantified the calibration errors of the technique, and generalized the method to include near-redundant arrays. Given that redundant (or close to redundant) baselines will be necessary to meet the sensitivity requirements of most 21 cm interferometer arrays anyway, redundant calibration should be further explored as a practical technique. Indeed, this is what the LOFAR group is in the process of doing, and early results are encouraging (Noorishad et al., 2012). If redundant calibration does ultimately turn out to work sufficiently well for the purposes of hydrogen cosmology, it would be wise to design next-generation arrays with explicit considerations of baseline redundancy, as this can reduce the cost of correlations from $O(n^2)$ to $O(n \log n)$, where n is the number of receiving elements.

Foreground modeling was considered in Chapter 3, where a phenomenological foreground model was used to elucidate the fact that foreground spectra can typically be described using a small set of smooth basis functions. Over small frequency ranges, these functions appeared to look very much like polynomials, and in Chapters 4 and 5 we used this approximation to show that foregrounds could indeed be subtracted to a reasonable residual levels by fitting out the smooth components of measured spectra.

Chapter 6 generalized the foreground subtraction process to a matrix-based inverse variance subtraction, which can be proved to be optimal in the limit of Gaussian noise. Our framework was adapted from the quadratic estimator formalism commonly used in CMB studies and galaxy surveys, and allowed us to consider foreground subtraction and power spectrum estimation in a single unified step. Doing so provided a power spectrum estimator that was optimal and unbiased by residual noise and foregrounds.

Chapter 7 turned to the optimization of global 21 cm experiments. Again developing a rigorous statistical framework for data analysis, we derived an optimal way to extract a global signal spectrum from a noisy and foreground-contaminated experiment. We found that the error bars on final measurement could be reduced by incorporating angular information. This suggests that future global signal experiments should be designed to have at least a coarse level of angular sensitivity.

The techniques explored in this thesis culminated in an application to real data in Chapter 8. Data was taken for 5 hrs using the MWA in its 32-tile prototype configuration, sky maps were produced, and a series of power spectra were computed. Using the cylindrical power spectrum $P(k_\perp, k_\parallel)$ of the dirty maps as a diagnostic, we detected a theoretically expected "wedge" due to the chromatic convolution of bright point source foregrounds by our instrument. We also confirmed the existence of promising regions in the $k_\perp$-$k_\parallel$ plane for future EoR signal detection, where contaminants such as foregrounds are expected to contribute relatively little to the measurement. Finally, we placed an upper limit on the spherically-binned power spectrum $P(k)$, finding it to be competitive with limits from other instruments.

In conclusion, we have derived a number of data analysis methods for hydrogen cosmology. These include calibration, foreground subtraction, and signal extraction techniques. Tests on real data have demonstrated their potential, yielding interesting limits. This bodes well for hydrogen cosmology, as it suggests that even though there is more work to be done, many of the proposed analysis methods appear to be robustly applicable, providing the necessary tools for a successful detection of the EoR signal in the near future and exquisite constraints on our Universe beyond that.

Bibliography

Ali, S. S., Bharadwaj, S., & Pandey, B. 2005, MNRAS, 363, 251

Barkana, R. 2006, MNRAS, 372, 259

Barkana, R. & Loeb, A. 2005a, ApJ Lett., 624, L65

—. 2005b, ApJ, 626, 1

Bernardi, G., de Bruyn, A. G., Harker, G., Brentjens, M. A., Ciardi, B., Jelić, V., Koopmans, L. V. E., Labropoulos, P., Offringa, A., Pandey, V. N., Schaye, J., Thomas, R. M., Yatawatta, S., & Zaroubi, S. 2010, AA, 522, A67+

Bernardi, G., Mitchell, D. A., Ord, S. M., Greenhill, L. J., Pindor, B., Wayth, R. B., & Wyithe, J. S. B. 2011, MNRAS, 413, 411

Bhatnagar, S., Cornwell, T. J., Golap, K., & Uson, J. M. 2008, A&A, 487, 419

Bittner, J. M. & Loeb, A. 2011, JCAP, 4, 38

Bond, J. R., Jaffe, A. H., & Knox, L. 1998, Phys. Rev. D, 57, 2117

Bowman, J. D. 2007, PhD thesis

Bowman, J. D., Morales, M. F., & Hewitt, J. N. 2006, ApJ, 638, 20

—. 2009, ApJ, 695, 183

Bowman, J. D. & Rogers, A. E. E. 2010, Nature, 468, 796

Bowman, J. D., Rogers, A. E. E., & Hewitt, J. N. 2008, ApJ, 676, 1

Burns, J. O., Lazio, J., Bale, S., Bowman, J., Bradley, R., Carilli, C., Furlanetto, S., Harker, G., Loeb, A., & Pritchard, J. 2012, Advances in Space Research, 49, 433

Carilli, C. L. & Rawlings, S. 2004, Science with the Square Kilometer Array (Amsterdam: Elsevier)

Chang, T., Pen, U., Peterson, J. B., & McDonald, P. 2008, Phys. Rev. Lett., 100, 091303

Clark, B. G. 1980, A&A, 89, 377

Cooray, A., Li, C., & Melchiorri, A. 2008, PRD, 77, 103506

Cornwell, T. & Fomalont, E. B. 1999, in Astronomical Society of the Pacific Conference Series, Vol. 180, Synthesis Imaging in Radio Astronomy II, ed. G. B. Taylor, C. L. Carilli, & R. A. Perley, 187–+

Cornwell, T. J., Golap, K., & Bhatnagar, S. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 86

Datta, A., Bhatnagar, S., & Carilli, C. L. 2009, Ap. J., 703, 1851

Datta, A., Bowman, J. D., & Carilli, C. L. 2010, ApJ, 724, 526

de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., Jonas, J., Landecker, T. L., & Reich, P. 2008, MNRAS, 388, 247

Di Matteo, T., Ciardi, B., & Miniati, F. 2004, MNRAS, 355, 1053

Di Matteo, T., Perna, R., Abel, T., & Rees, M. J. 2002, ApJ, 564, 576

Dillon, J. S., Liu, A., & Tegmark, M. 2012, in preparation

Fan, X., Strauss, M. A., Becker, R. H., White, R. L., Gunn, J. E., Knapp, G. R., Richards, G. T., Schneider, D. P., Brinkmann, J., & Fukugita, M. 2006, AJ, 132, 117

Fisher, R. A. 1935, J. Roy. Stat. Soc., 98, 39

Furlanetto, S. R., Lidz, A., Loeb, A., McQuinn, M., Pritchard, J. R., Alvarez, M. A., Backer, D. C., Bowman, J. D., Burns, J. O., Carilli, C. L., Cen, R., Cooray, A., Gnedin, N., Greenhill, L. J., Haiman, Z., Hewitt, J. N., & et al. 2009, in Astronomy, Vol. 2010, AGB Stars and Related Phenomenastro2010: The Astronomy and Astrophysics Decadal Survey, 83–+

Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, Physics Reports, 433, 181

Furlanetto, S. R., Sokasian, A., & Hernquist, L. 2004a, MNRAS, 347, 187

Furlanetto, S. R., Zaldarriaga, M., & Hernquist, L. 2004b, ApJ, 613, 16

Geil, P. M., Gaensler, B. M., & Wyithe, J. S. B. 2010, ArXiv e-prints: 1011.2321

Gleser, L., Nusser, A., & Benson, A. J. 2008, MNRAS, 391, 383

Greenhill, L. J. & Bernardi, G. 2012, ArXiv e-prints: 1201.1700

Hamilton, A. J. S. & Tegmark, M. 2000, MNRAS, 312, 285

Harker, G., Zaroubi, S., Bernardi, G., Brentjens, M. A., de Bruyn, A. G., Ciardi, B., Jelic, V., Koopmans, L. V. E., Labropoulos, P., Mellema, G., Offringa, A., Pandey, V. N., Pawlik, A. H., Schaye, J., Thomas, R. M., & Yatawatta, S. 2010, ArXiv e-prints: 1003.0965

Harker, G., Zaroubi, S., Bernardi, G., Brentjens, M. A., de Bruyn, A. G., Ciardi, B., Jelic, V., Koopmans, L. V. E., Labropoulos, P., Mellema, G., Offringa, A., Pandey, V. N., Schaye, J., Thomas, R. M., & Yatawatta, S. 2009, ArXiv e-prints: 0903.2760 (astro-ph)

Harker, G. J. A., Pritchard, J. R., Burns, J. O., & Bowman, J. D. 2012, MNRAS, 419, 1070

Hinshaw, G., Barnes, C., Bennett, C. L., Greason, M. R., Halpern, M., Hill, R. S., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Page, L., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2003, ApJS, 148, 63

Hinshaw, G. & et. al. [WMAP collaboration]. 2007, ApJS, 170, 288

Högbom, J. A. 1974, A&AS, 15, 417

Iliev, I. T., Shapiro, P. R., Ferrara, A., & Martel, H. 2002, ApJL, 572, L123

Ishiguro, M. 1974, Astronomy and Astrophysics Supplement, 15, 431

Jackson, C. 2005, PASA, 22, 36

Jelic, V., Zaroubi, S., Labropoulos, P., Thomas, R. M., Bernardi, G., Brentjens, M., de Bruyn, G., Ciardi, B., Harker, G., Koopmans, L. V. E., Pandey, V., Schaye, J., & Yatawatta, S. 2008, ArXiv e-prints: 0804.1130 (astro-ph)

Komatsu, E., Smith, K. M., Dunkley, J., Bennett, C. L., Gold, B., Hinshaw, G., Jarosik, N., Larson, D., Nolta, M. R., Page, L., Spergel, D. N., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2011, Ap. J. Supplement, 192, 18

Lewis, A. & Challinor, A. 2007, PRD, 76, 083005

Lidz, A., Zahn, O., McQuinn, M., Zaldarriaga, M., & Hernquist, L. 2008, ApJ, 680, 962

Liu, A. & Tegmark, M. 2011, Phys. Rev. D, 83, 103006

Liu, A. & Tegmark, M. 2012, MNRAS, 419, 3491

Liu, A., Tegmark, M., Bowman, J., Hewitt, J., & Zaldarriaga, M. 2009a, MNRAS, 398, 401

Liu, A., Tegmark, M., Morrison, S., Lutomirski, A., & Zaldarriaga, M. 2010, MNRAS, 408, 1029

Liu, A., Tegmark, M., & Zaldarriaga, M. 2009b, MNRAS, 394, 1575

Loeb, A. & Wyithe, J. S. B. 2008, Physical Review Letters, 100, 161301

Loeb, A. & Zaldarriaga, M. 2004, Physical Review Letters, 92, 211301

Lonsdale, C. J., Cappallo, R. J., Morales, M. F., Briggs, F. H., Benkevitch, L., Bowman, J. D., Bunton, J. D., Burns, S., Corey, B. E., Desouza, L., Doeleman, S. S., Derome, M., Deshpande, A., Gopala, M. R., Greenhill, L. J., Herne, D. E., Hewitt, J. N., Kamini, P. A., Kasper, J. C., Kincaid, B. B., Kocz, J., Kowald, E., Kratzenberg, E., Kumar, D., Lynch, M. J., Madhavi, S., Matejek, M., Mitchell, D. A., Morgan, E., Oberoi, D., Ord, S., Pathikulangara, J., Prabu, T., Rogers, A., Roshi, A., Salah, J. E., Sault, R. J., Shankar, N. U., Srivani, K. S., Stevens, J., Tingay, S., Vaccarella, A., Waterson, M., Wayth, R. B., Webster, R. L., Whitney, A. R., Williams, A., & Williams, C. 2009, IEEE Proceedings, 97, 1497

Madau, P., Meiksin, A., & Rees, M. J. 1997, ApJ, 475, 429

Mao, X.-C. 2012, ArXiv e-prints: 1204.1812

Mao, Y., Tegmark, M., McQuinn, M., Zaldarriaga, M., & Zahn, O. 2008, PRD, 78, 023529

Masui, K. W., McDonald, P., & Pen, U.-L. 2010, PRD, 81, 103527

McQuinn, M., Hernquist, L., Zaldarriaga, M., & Dutta, S. 2007, MNRAS, 381, 75

McQuinn, M., Lidz, A., Zahn, O., Dutta, S., Hernquist, L., & Zaldarriaga, M. 2006a, ArXiv e-prints: 0610094 (astro-ph)

McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. R. 2006b, ApJ, 653, 815

Mesinger, A., McQuinn, M., & Spergel, D. N. 2012, MNRAS, 2672

Morales, M. F. 2005, ApJ, 619, 678

Morales, M. F., Bowman, J. D., & Hewitt, J. N. 2006, ApJ, 648, 767

Morales, M. F., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ArXiv e-prints: 1202.3830

Morales, M. F. & Hewitt, J. 2004, ApJ, 615, 7

Morales, M. F. & Matejek, M. 2009, MNRAS, 400, 1814

Morales, M. F. & Wyithe, J. S. B. 2009, ArXiv e-prints: 0910.3010

Noordam, J. E. & de Bruyn, A. G. 1982, Nature, 299, 597

Noorishad, P., Wijnholds, S. J., van Ardenne, A., & van der Hulst, T. 2012, ArXiv e-prints: 1205.1413

Nusser, A. 2005, MNRAS, 364, 743

Oh, S. P. & Mack, K. J. 2003, MNRAS, 346, 871

Paciga, G., Chang, T., Gupta, Y., Nityanada, R., Odegova, J., Pen, U., Peterson, J., Roy, J., & Sigurdson, K. 2010, ArXiv Astrophysics e-prints: 1006.1351

Padmanabhan, N., Seljak, U., & Pen, U. L. 2003, New Astronomy, 8, 581

Parsons, A., McQuinn, M., Jacobs, D., Aguirre, J., & Pober, J. 2011, ArXiv e-prints: 1103.2135

Parsons, A. R., Backer, D. C., Foster, G. S., Wright, M. C. H., Bradley, R. F., Gugliucci, N. E., Parashare, C. R., Benoit, E. E., Aguirre, J. E., Jacobs, D. C., Carilli, C. L., Herne, D., Lynch, M. J., Manley, J. R., & Werthimer, D. J. 2010, The Astronomical Journal, 139, 1468

Pearson, T. J. & Readhead, A. C. S. 1984, Annu. Rev. of Astronomy and Astrophysics, 22, 97

Peebles, P. J. E. 2012, ArXiv e-prints: 1203.6334

Pen, U. 2003, MNRAS, 346, 619

Peterson, J. B., Bandura, K., & Pen, U. L. 2006, ArXiv Astrophysics e-prints: astro-ph/0606104

Petrovic, N. & Oh, S. P. 2011, MNRAS, 413, 2103

Pindor, B., Wyithe, J. S. B., Mitchell, D. A., Ord, S. M., Wayth, R. B., & Greenhill, L. J. 2010, ArXiv Astrophysics e-prints: 1007.2264

Pritchard, J. R. & Loeb, A. 2010, Phys. Rev. D, 82, 023006

Ramesh, R., Subramanian, K. R., & Sastry, C. V. 1999, Astronomy and Astrophysics Supplement, 139, 179

Rau, U., Bhatnagar, S., Voronkov, M. A., & Cornwell, T. J. 2009, IEEE Proceedings, 97, 1472

Reid, B. A., Percival, W. J., Eisenstein, D. J., Verde, L., Spergel, D. N., Skibba, R. A., Bahcall, N. A., Budavari, T., Frieman, J. A., Fukugita, M., Gott, J. R., Gunn, J. E., Ivezić, Ž., Knapp, G. R., Kron, R. G., Lupton, R. H., McKay, T. A., Meiksin, A., Nichol, R. C., Pope, A. C., Schlegel, D. J., Schneider, D. P., Stoughton, C., Strauss, M. A., Szalay, A. S., Tegmark, M., Vogeley, M. S., Weinberg, D. H., York, D. G., & Zehavi, I. 2010, MNRAS, 404, 60

Robertson, B. E., Ellis, R. S., Dunlop, J. S., McLure, R. J., & Stark, D. P. 2010, Nature, 468, 49

Santos, M. G., Cooray, A., & Knox, L. 2005, ApJ, 625, 575

Shaver, P. A., Windhorst, R. A., Madau, P., & de Bruyn, A. G. 1999, A & A, 345, 380

Tegmark, M. 1997a, PRD, 56, 4514

—. 1997b, Ap. J. Let., 480, L87+

—. 1997c, Phys. Rev. D, 55, 5895

Tegmark, M., Blanton, M. R., Strauss, M. A., & et al. 2004, ApJ, 606, 702

Tegmark, M. & Efstathiou, G. 1996, MNRAS, 281, 1297

Tegmark, M., Eisenstein, D. J., Hu, W., & de Oliveira-Costa, A. 2000, ApJ, 530, 133

Tegmark, M., Eisenstein, D. J., Strauss, M. A., Weinberg, D. H., Blanton, M. R., Frieman, J. A., Fukugita, M., Gunn, J. E., Hamilton, A. J. S., Knapp, G. R., Nichol, R. C., Ostriker, J. P., Padmanabhan, N., Percival, W. J., Schlegel, D. J., Schneider, D. P., Scoccimarro, R., Seljak, U., Seo, H.-J., Swanson, M., Szalay, A. S., Vogeley, M. S., Yoo, J., Zehavi, I., Abazajian, K., Anderson, S. F., Annis, J., Bahcall, N. A., Bassett, B., Berlind, A., Brinkmann, J., Budavari, T., Castander, F., Connolly, A., Csabai, I., Doi, M., Finkbeiner, D. P., Gillespie, B., Glazebrook, K., Hennessy, G. S., Hogg, D. W., Ivezić, Ž., Jain, B., Johnston, D., Kent, S., Lamb, D. Q., Lee, B. C., Lin, H., Loveday, J., Lupton, R. H., Munn, J. A., Pan, K., Park, C., Peoples, J., Pier, J. R., Pope, A., Richmond, M., Rockosi, C., Scranton, R., Sheth, R. K., Stebbins, A., Stoughton, C., Szapudi, I., Tucker, D. L., vanden Berk, D. E., Yanny, B., & York, D. G. 2006, PRD, 74, 123507

Tegmark, M., Hamilton, A. J. S., Strauss, M. A., Vogeley, M. S., & Szalay, A. S. 1998, ApJ, 499, 555

Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, Ap. J., 480, 22

Tegmark, M. & Villumsen, J. V. 1997, MNRAS, 289, 169

Tegmark, M. & Zaldarriaga, M. 2009, PRD, 79, 083530

—. 2010, PRD, 82, 103501

Toffolatti, L., Argueso Gomez, F., de Zotti, G., Mazzei, P., Franceschini, A., Danese, L., & Burigana, C. 1998, MNRAS, 297, 117

Tozzi, P., Madau, P., Meiksin, A., & Rees, M. J. 2000a, ApJ, 528, 597

—. 2000b, Nuclear Physics B Proceedings Supplements, 80, C509+

Trott, C. M. & Wayth, R. B. 2012, submitted

Vedantham, H., Udaya Shankar, N., & Subrahmanyan, R. 2012, ApJ, 745, 176

Visbal, E., Loeb, A., & Wyithe, S. 2009, JCAP, 10, 30

Wang, X., Tegmark, M., Santos, M. G., & Knox, L. 2006, ApJ, 650, 529

Wieringa, M. H. 1992, Experimental Astronomy, 2, 203

Williams, C. L., Hewitt, J. N., Levine, A. M., de Oliveira-Costa, A., Bowman, J. D., Briggs, F. H., Gaensler, B. M., Hernquist, L. L., Mitchell, D. A., Morales, M. F., Sethi, S. K., Subrahmanyan, R., Sadler, E. M., Arcus, W., Barnes, D. G., Bernardi, G., Bunton, J. D., Cappallo, R. C., Crosse, B. W., Corey, B. E., Deshpande, A., deSouza, L., Emrich, D., Goeke, R. F., Greenhill, L. J., Hazelton, B. J., Herne, D., Kaplan, D. L., Kasper, J. C., Kincaid, B. B., Koenig, R., Kratzenberg, E., Lonsdale, C. J., Lynch, M. J., McWhirter, S. R., Mitchell, . A., Morales, . F., Morgan, E. H., Oberoi, D., Ord, S. M., Pathikulangara, J., Prabu, T., Remillard, R. A., Rogers, A. E. E., Roshi, A. A., Salah, J. E., Sault, R. J., Udaya Shankar, N., Srivani, K. S., Stevens, J. B., Tingay, S. J., Wayth, R. B., Waterson, M., Webster, R. L., Whitney, A. R., Williams, A. J., & Wyithe, J. S. B. 2012, ArXiv e-prints: 1203.5790

Wouthuysen, S. A. 1952, AJ, 57, 31

Wyithe, J. S. B. & Loeb, A. 2009, MNRAS, 397, 1926

Wyithe, J. S. B., Loeb, A., & Geil, P. M. 2008, MNRAS, 383, 1195

Yang, Y. 1988, Astronomy and Astrophysics, 189, 361

Zahn, O., Reichardt, C. L., Shaw, L., Lidz, A., Aird, K. A., Benson, B. A., Bleem, L. E., Carlstrom, J. E., Chang, C. L., Cho, H. M., Crawford, T. M., Crites, A. T., de Haan, T., Dobbs, M. A., Dore, O., Dudley, J., George, E. M., Halverson, N. W., Holder, G. P., Holzapfel, W. L., Hoover, S., Hou, Z., Hrubes, J. D., Joy, M., Keisler, R., Knox, L., Lee, A. T., Leitch, E. M., Lueker, M., Luong-Van, D., McMahon, J. J., Mehl, J., Meyer, S. S., Millea, M., Mohr, J. J., Montroy, T. E., Natoli, T., Padin, S., Plagge, T., Pryke, C., Ruhl, J. E., Schaffer, K. K., Shirokoff, E., Spieler, H. G., Staniszewski, Z., Stark, A. A., Story, K., van Engelen, A., Vanderlinde, K., Vieira, J. D., & Williamson, R. 2011, ArXiv e-prints: 1111.6386

Zaldarriaga, M., Furlanetto, S. R., & Hernquist, L. 2004, ApJ, 608, 622