Methods for the Detection of the Epoch of Reionization by Interferometers Measuring the 21 cm Signal

By

Carina Cheng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Astrophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Aaron Parsons, Chair
Professor Stuart Bale
Professor Mariska Kriek

Spring 2019

Methods for the Detection of the Epoch of Reionization by Interferometers Measuring the 21 cm Signal

Abstract

Methods for the Detection of the Epoch of Reionization by Interferometers Measuring the 21 cm Signal

by

Carina Cheng

Doctor of Philosophy in Astrophysics

University of California, Berkeley

Professor Aaron Parsons, Chair

The Epoch of Reionization is one of the last unexplored eras of our Universe's history. Beginning about a billion years after the Big Bang, this epoch is characterized by the births of the first stars and galaxies, whose light subsequently altered the nature of the gas surrounding them. There are several experiments aiming to detect the phase transition of this gas as it changes from neutral hydrogen to ionized hydrogen. Such a detection would open a wealth of information about our early Universe, revealing details about the nature of the first luminous sources and the evolution of structure formation.

Interferometers such as the Precision Array for Probing the Epoch of Reionization (PAPER) and the Hydrogen Epoch of Reionization Array (HERA) seek to measure the 21 cm signal from neutral hydrogen and map its evolution over spatial and temporal scales. A detection of reionization, however, is a difficult measurement. Though the 21 cm signal is a powerful 3D probe of the intergalactic medium, it is easily buried underneath bright foreground signals and instrumental systematics. A clean detection of reionization is ambitious and requires analysis methods that maximize data sensitivity and increase confidence in results.

The work presented in this thesis addresses many of the key challenges that face the current field of 21 cm cosmology. This includes algorithms to locate contaminated data, methods to ensure accurate power spectrum measurements, and techniques for removing unwanted systematics while preserving the reionization signal. These developments serve as the foundation of the latest 21 cm measurements from the PAPER-64 and PAPER-128 arrays, whose results lie near the forefront of the field. These methods are also fundamental to HERA and future experiments, as they provide a strong foundation for the continued exploration of our cosmic dawn.

to my parents

Contents

List of Figures

List of Tables

Acknowledgments

coming soon!

Chapter 1

Introduction

1.1 The Epoch of Reionization

Our Universe has a complex, rich history, and enormous progress has been made in the past few decades to unravel its story. Much has been learned about the very beginnings of the Universe, from the Big Bang's large explosion of energy to the relatively smooth and simple cosmic background radiation that was leftover. Additionally, observational feats have revealed characteristics of the present-day Universe and the intricate *cosmic web*, or large scale structure, of galaxies today.

The Epoch of Reionization (EoR) ties these two bookends together, taking place about a billion years after the Big Bang when young generations of stars and galaxies formed. How did the tiny density fluctuations from the cosmic microwave background develop into the structure we see today? How did the first luminous structures form, and how did they evolve and influence the gas around them? Exploring the reionization era opens up a new chapter of our Universe's story - a chapter that promises to connect the dots between our past and present.

1.1.1 Cosmic History

As the Universe expanded and cooled after the Big Bang, electrons and protons eventually combined to form neutral hydrogen atoms. At the young age of $\sim 380,000$ years, the Universe's baryonic content was almost entirely neutral hydrogen, while most of its total matter was dark matter (Loeb & Furlanetto 2013). Then, for the next several hundred million years, the *Dark Ages* proceeded, with concentrations of dark matter setting the foundations for the formation of the first luminous structures and black holes. More specifically, the tiny primordial density fluctuations that were established at the release of the CMB grew with inflation and the expansion of the Universe. The densest regions then collapsed to form dark matter halos, inside of which hydrogen gas could cool, condense, and fragment into stars (Dodelson 2003).

The first luminous structures are thought to have formed at an age of $\sim$ 200 million

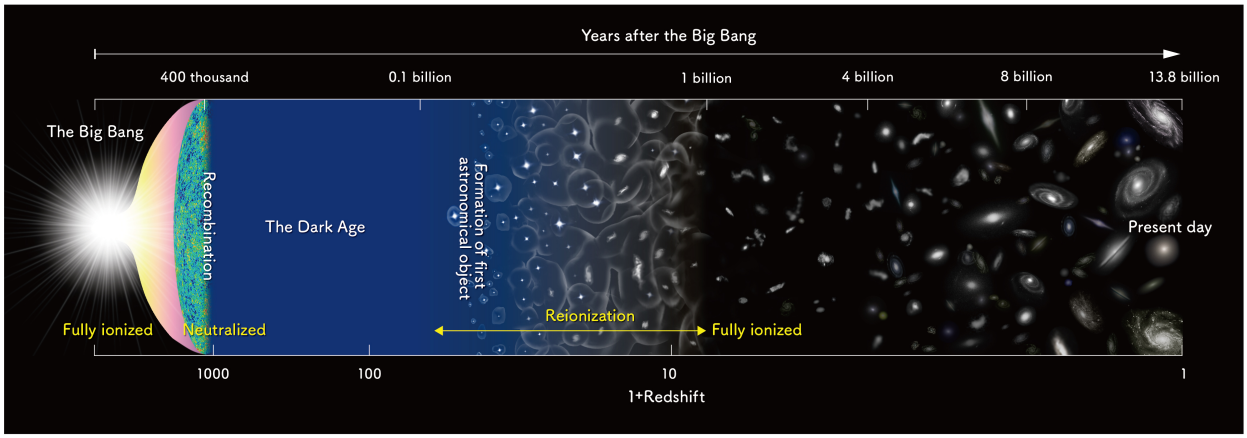

Figure 1.1: Timeline of the history of the Universe. The Epoch of Reionization marks the era when the first stars and galaxies formed and ionized the neutral hydrogen in the Universe. Image credit: NAOJ.

years ($z \sim 20$) and are predicted to have been massive stars with high luminosities and large ionizing powers (Loeb & Furlanetto 2013). Photons from these first stars are believed to have escaped their dark matter halo hosts, carving out pockets of ionized gas in the voids between halos ("outside-in" reionization; Miralda-Escude et al. 2000). As the number of sources grew, an increased number of ionized bubbles emerged and overlapped. Eventually, the first generations of stars and galaxies succeeded in ionizing most all of the neutral hydrogen in the Universe, with reionization predicted to be complete by about one billion years after the Big Bang ($z \sim 6$) (Furlanetto et al. 2006).

The exact timescale and details of the reionization process, which are shown within the context of the history of the Universe in Figure 1.1, are current research questions in the field of cosmology. The physics of reionization depends on several factors, including the nature of the first galaxies (i.e., the sources responsible for producing ionizing photons) and the surrounding gas (i.e., how efficiently the photons escaped their host galaxies to ionize the gas). Turning the argument around, deep investigations of the reionization era would lead to new understandings about the properties of the first galaxies and the intergalactic medium (IGM). There are several ways to approach the study of this epoch, with CMB measurements working to constrain the duration of reionization, galaxy measurements unveiling the end of reionization, and direct hydrogen measurements attempting to map out the changing density of the gas over time. All of these probes serve to illuminate this watershed era between a Universe dominated by darkness and a Universe defined by light.

1.1.2 CMB and Galaxy Measurements

There are several observational probes of the reionization epoch, and we highlight two broad categories in this section. The first is the study of CMB anisotropies, which carry

with them an imprint of the early Universe from the time of its release. But that's not the only imprint it has - CMB photons can scatter off of free electrons after reionization, and these scatterings leave behind polarization and temperature imprints (Haiman & Knox 1999). For example, the amplitude of the CMB is sensitive to scatterings, as an increased number of scatterings is akin to mixing different directions of the CMB together as photons are scattered in all directions. In other words, this scattering washes out anisotropies in the CMB and lowers its overall amplitude.

A useful parameter to quantify the amount of electron scattering that occurs is the optical depth, τ_{es}, defined as:

$$\tau_{es} = \int n_e \sigma_T dl \tag{1.1}$$

where n_e is the number density of free electrons, σ_T is the Thompson cross-section, and the integral is taken over a proper length dl. Once reionization begins, the number of free electrons increases, contributing to increasingly higher values of τ_{es}. Hence, an earlier start for reionization would yield higher optical depths than a late reionization scenario. For completeness, a more exact version of the 21 cm optical depth is:

$$\tau_{es} \sim 0.0092(1+\delta_b)(1+z)^{3/2} \frac{x_{HI}}{T_{\rm spin}} \Big(\frac{H(z)/(1+z)}{{\rm d}v_\parallel/{\rm d}r_\parallel} \Big), \tag{1.2}$$

where z is the redshift, δ_b is the fractional over-density of matter, x_{HI} is the fraction of neutral hydrogen (1 if all neutral, 0 if all ionized), $H(z)$ is the Hubble parameter, ${\rm d}v_\parallel/{\rm d}r_\parallel$ is the gradient of the proper velocity along the line of sight, and $T_{\rm spin}$ is the spin temperature, which is defined and explained in more detail in Chapter 1.1.3.

Observations of the CMB by WMAP and Planck have placed constraints on the optical depth parameter (Hinshaw et al. 2013; Planck Collaboration et al. 2016), with the more recent Planck result suggesting a value of $\tau_{es} \sim 0.066 \pm 0.016$. This value suggests that reionization ends at a redshift of $z \sim 6$, with instantaneous reionization ("mean" reionization) at $z \sim 8.8$ (Planck Collaboration et al. 2016).

Currently, the results from CMB measurements are in agreement with a second powerful probe of EoR — broadly speaking, that of galaxy observations. This probe comes in many flavors. For example, the spectra of distant quasars at high redshifts can illuminate the end of reionization. Quasars, being extraordinarily bright and energetic objects, are detectable at very far distances and their spectra reveal the amount of absorption their light has undergone due to the presence of neutral hydrogen. While nearby quasar spectra exhibit sharp absorption lines, distant ones show the Gunn-Peterson trough, implying that the quasar light was entirely suppressed by hydrogen absorption (i.e., neutral hydrogen existed). Studying the absorption features of quasars at different redshifts implies that reionization has indeed ended by $z \sim 6$ (Becker et al. 2001).

In addition to quasar observations, high-redshift galaxy observations can also reveal important characteristics about the state of the IGM. Namely, distant star-forming galaxies

can be detected using a variety of techniques, such as narrow-band imaging to find Lyman-α emitters (radiation that is produced by recombination near young stars) or broad-band observations to find Lyman-break galaxies (spectral breaks associated with absorption by neutral hydrogen). High-redshift galaxy observations can then be used to construct luminosity functions (number of stars per luminosity interval) and star formation histories, which in turn impact the evolution of the IGM.

More specifically, if star-forming galaxies dominated the reionization process, then the ionization rate can be related to the following star-formation parameters:

$$\dot{n}_{\rm ion} = f_{\rm esc}\xi_{\rm ion}\rho_{\rm SFR}, \tag{1.3}$$

where $\dot{n}_{\rm ion}$ is the cosmic ionization rate, $f_{\rm esc}$ is the escape fraction of photons into the IGM, $\xi_{\rm ion}$ is the rate of production of ionizing photons for a stellar population, and $\rho_{\rm SFR}$ is the star formation rate density. All three parameters influence the rate at which the IGM is ionized, and the star formation rate density is able to be constructed from galaxy luminosity functions. For example, Robertson et al. (2015) used data from the Hubble Space Telescope to construct a star formation rate history out to high redshifts, backing out an optical depth parameter that is consistent with that of Planck.

While galaxy measurements can be used to constrain the EoR, they are ultimately doing so by unveiling the properties of old, distant stars and galaxies. A similar, new technique that also aims to reconstruct the histories of the first luminous structures is observing nearby, metal-poor Local Group galaxies. Called "galactic archaeology," observations of nearby star-forming ancestors can be used to constrain the faint-end slope of the luminosity function. Determining the shape of this function has important implications on the number of galaxies needed to drive reionization and the types of sources dominating this epoch (Weisz & Boylan-Kolchin 2017). Additionally, studies of nearby metal-poor stars and galaxies can provide insight into the contents of the first generation of stars and the dynamics of high-redshift star formation, as observations of ultrafaint dwarf galaxies around the Milky Way suggest they are relatively clean tracers of the first generations of stars (Loeb & Furlanetto 2013).

Galaxy observations for reionization studies have been primarily driven by observations taken by the Hubble Space Telescope. In the coming years, the James Webb Space Telescope (JWST), a 6.5 m infrared space telescope, will be optimally primed for the detection of faint galaxies, including galaxies whose roots extend as far back as the cosmic dawn and who may exhibit signatures of first generation Population III stars. In addition to JWST, several large infrared ground-based telescopes are also underway, including the European Extremely Large Telescope (EELT), the Giant Magellan Telescope (GMT), and the Thirty Meter Telescope (TMT).

Although both CMB measurements and galaxy observations have much to look forward to, they currently each have their limitations. For example, CMB measurements can only reveal the integrated quantity of τ_{es}, therefore unable to provide insight into the evolution of reionization as it progresses over time. Similarly, galaxy observations are currently limited by finite observations of galaxies whose local environments may not be representative of

environments during the reionization era. A different, but complimentary, probe is needed to unlock the entire window into the EoR.

1.1.3 Measurements of Neutral Hydrogen

A direct measurement of neutral hydrogen gas over time would provide a fundamental way to track the IGM over the reionization process. Such a measurement is made possible by the spin-flip transition of hydrogen, which serves as a powerful probe that allows the tracing of gas over time, and it is this technique that serves as the basis for the remainder of this thesis (Furlanetto et al. 2006; Barkana & Loeb 2008; Morales & Wyithe 2010; Pritchard & Loeb 2010; Pritchard & Loeb 2012).

The spin-flip transition of neutral hydrogen occurs when a hydrogen atom changes energy state between two hyperfine levels. Namely, if a hydrogen atom moves from an aligned energy state (proton and electron having parallel spins) to an anti-aligned state (proton and electron having antiparallel spins), the energy difference is released in the form of a photon with a wavelength of 21 cm.

Because this transition has a well-defined wavelength, the signal can be directly mapped to a distance, or redshift, by measuring its wavelength upon detection. For example, a 21 cm photon that was initially emitted at a redshift of $z = 6$ would have expanded by a factor of $(1 + z)$ due to the expansion of the Universe and be 1.5 m long when it arrives at our telescopes. Hence, observing longer wavelengths of the hydrogen signal means that it has traveled for a greater distance (and has stretched out more) and thus comes from farther away at a higher redshift. This means that the 21 cm signal is a powerful tracer of neutral hydrogen at any distance (i.e., as a function of time), as long as it exists. This technique is especially compelling because it allows the direct exploration of the EoR as reionization occurs, opening up a window into a largely unexplored era sandwiched between CMB and galaxy observations (Figure 1.2).

In practice, the 21 cm signal is encapsulated by the quantity $T_{\rm spin}$ (spin temperature), which measures the relative number of hydrogen atoms in the excited (aligned) versus ground (anti-aligned) spin-flip state. A high spin temperature means that the hydrogen gas is more likely to emit 21 cm photons, whereas a low $T_{\rm spin}$ implies that the gas is more likely to absorb 21 cm photons.

The spin temperature is always measured with respect to the temperature of the CMB ($T_{\rm CMB}$), which serves as a backlight for our measurement. During different stages of our cosmic history, $T_{\rm CMB}$ and $T_{\rm spin}$ take turns in the spotlight, with the *differential brightness temperature* δT_b describing their evolution:

$$\delta T_b \sim 9(1+z)^{1/2}(1+\delta_b)x_{HI}\Big(1-\frac{T_{\rm CMB}}{T_{\rm spin}}\Big)\Big(\frac{H(z)/(1+z)}{{\rm d}v_\parallel/{\rm d}r_\parallel}\Big). \tag{1.4}$$

Equation (1.4) captures the EoR signal that 21 cm experiments seek to measure, where z is the redshift, δ_b is the fractional over-density of matter, x_{HI} is the fraction of neutral hydrogen (1 if all neutral, 0 if all ionized), $H(z)$ is the Hubble parameter, and ${\rm d}v_\parallel/{\rm d}r_\parallel$ is

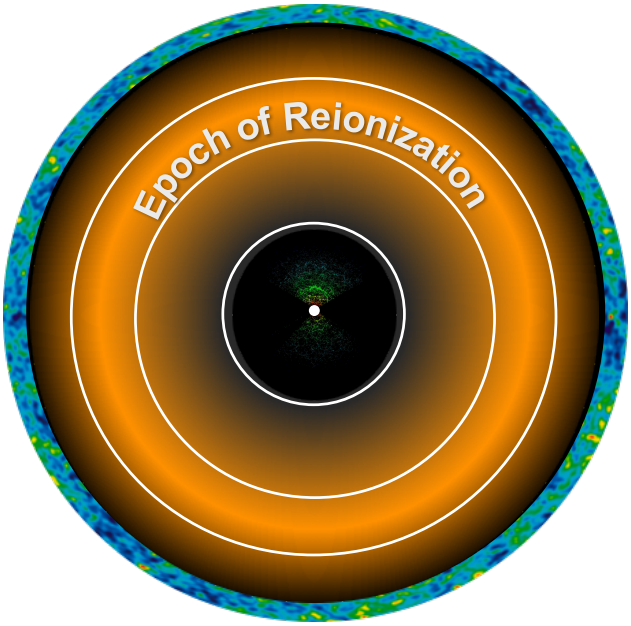

Figure 1.2: A cartoon diagram of the observable Universe, centered on us. Close-by, galaxy observations have mapped out cosmic web structure in our nearby Universe (image credit: SDSS). Far-away, the cosmic microwave background is observed at a redshift of $z \sim 1100$ (image credit: WMAP). The Epoch of Reionization represents a largely unexplored era between the two, and can be probed by measuring redshifted 21 cm radiation from neutral hydrogen.

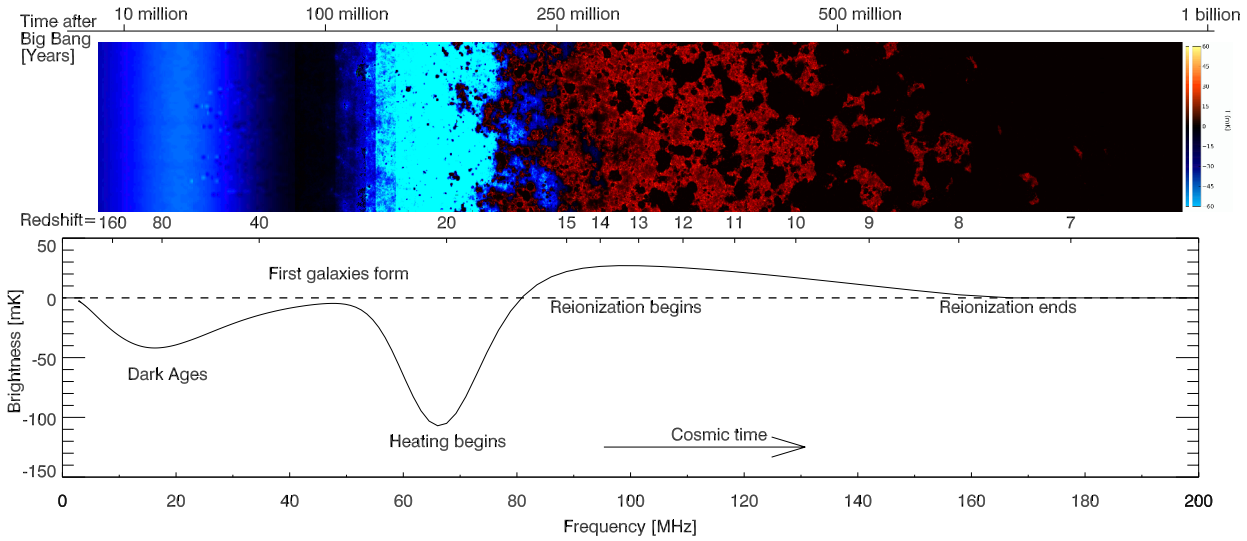

Figure 1.3: The evolution of the global 21 cm signal, starting with the Dark Ages, through galaxy formation and reionization (image credit: Pritchard & Loeb (2012)). The work in this thesis mainly focuses on a redshift range of $6 < z < 12$ when reionization is expected to progress and complete.

the gradient of the proper velocity along the line of sight (Furlanetto et al. 2006). The differential brightness temperature can be measured in multiple ways — in Chapter 1.2 we explain how interferometry (multiple telescopes) can be used to measure the correlations of δT_b on various spatial scales on the sky. Here we describe the evolution of the sky-averaged δT_b, called the *global signal*, in order to summarize how it is expected to behave during our cosmic dawn and through reionization.

A theoretical prediction for the evolution of δT_b is shown in Figure 1.3. At the very far left, a cooling, neutral IGM remains after recombination and the release of the CMB. Residual electrons collide off of both CMB photons and the hydrogen gas, driving couplings between $T_{\rm gas}$ and $T_{\rm CMB}$, and $T_{\rm gas}$ and $T_{\rm spin}$, respectively. Hence, we expect to see no signal ($\delta T_b = 0$) at this time.

During the Dark Ages, collisions still couple $T_{\rm gas}$ and $T_{\rm spin}$, but Compton scattering becomes rarer as the CMB dilutes with the expansion of the Universe. While the CMB dilutes as $T_{\rm CMB} \propto (1+z)$, the gas now follows an adiabatic expansion ($T_{\rm gas} \propto (1+z)^2$). Thus, the gas cools quicker than the CMB, and since it is still coupled to the spin temperature, $T_{\rm spin} < T_{\rm CMB}$ and the signal is expected to be seen in absorption. By the time the first galaxies begin forming, however, the gas is expected to be so dilute that it is no longer coupled to the spin temperature. The spin temperature therefore couples once again to the CMB, and no signal is produced.

As the first stars in the first galaxies begin emitting Lyman-α photons, they are resonantly scattered off of hydrogen via the Wouthuysen-Field effect (the absorption and emission of Lyman-α photons redistributes the spin-flip states), coupling $T_{\rm spin}$ and $T_{\rm gas}$ (Pritchard &

Loeb 2010). The gas, still cool from adiabatic expansion, implies that $T_{\rm spin} < T_{\rm CMB}$ and the signal is seen in absorption. Eventually, due to their long mean free paths, x-rays from the first sources are thought to be the primary drivers behind heating of the cooled, low-density gas (Furlanetto et al. 2006). This drives both the gas and spin temperatures above that of the CMB, where the signal is expected to be seen in emission for the first time.

Finally, even though the timing and details of reionization are unknown, UV photons from the first luminous structures are believed to eventually ionize all the neutral hydrogen, leaving no signal to be detected by a redshift of $z \sim 6$.

The shape of the global signal holds important science implications about our early Universe. For example, the timing of the heating trough reveals the types of sources responsible for heating (i.e., late heating implies harder x-ray spectras for x-ray binaries, as shown in Fialkov et al. (2014)). It also contains information regarding the sizes of the dark matter halos hosting those first sources and the cooling mechanisms responsible for star formation (Fialkov et al. 2014). The shape of the absorption feature is also dependent on a number of factors, such as x-ray and Lyman-α emissivities, which in turn are dependent on the nature of the first sources and properties of star formation. For high Lyman-α production rates, a deep trough would be present due to strong couplings between $T_{\rm spin}$ and $T_{\rm gas}$ as the gas cools. An even more pronounced absorption signature, such as the first tentative detection from the Experiment to Detect the Global Epoch of Reionization Signature (EDGES), requires additional physical explanations beyond known physics and commonly accepted scenarios (Bowman et al. 2018).

If the global signal's primary absorption feature unlocks clues about the first sources, the reionization peak and its subsequent decay hold the key for understanding the evolution of the neutral fraction x_{HI}. Namely, a direct measurement of δT_b during this time would shed light about the duration and rate of the reionization process, which in turn can be translated into an evolution for x_{HI}. A long reionization duration, for example, would yield a slowly varying neutral fraction evolution, while a more instantaneous reionization would produce a sharp drop-off feature (Pritchard & Loeb 2010). One thing is for certain though — as the community continues to investigate our cosmic dawn and the EoR through HI measurements (both the global signal and statistical fluctuations), we can expect to learn much about the constituents that make up the Universe and their complex interactions during this era.

1.2 Interferometry

Multiple radio telescopes (i.e., an interferometer) can be used in combination to probe 21 cm fluctuations. Rather than a single element, or aperture, many antennas can be used to increase the effective aperture size of a telescope.

As a simplistic example, two antennas may observe the same sky but each receives the sky signal at slightly different times, with a time delay determined by the antenna spacing, or baseline orientation and length, with respect to the sky. The two voltage streams from the antennas can then be correlated to form an output response with an amplitude dependent on

the sky's intensity and a phase dependent on the time delay between the two elements and the frequency of the light. The power received by this baseline, as we will see, represents one sample in the large "synthesized" aperture of the interferometer. Knowledge of the entire sky can be built up by having a large number of antennas and many different types, and copies, of baselines.

1.2.1 The Visibility Equation

The output measurement from correlating signals between two antennas is called a *visibility*. A visibility measurement is computed over the entire angular sky $d\Omega$ as:

$$V_{ij}(\nu) = \int A(\nu, \hat{s}) I(\nu, \hat{s}) e^{-2\pi i \frac{\vec{b}_{ij} \cdot \hat{s}}{\lambda}} d\Omega. \tag{1.5}$$

Focusing first on the exponential part, i and j represent a pair of antennas, $\vec{b}_{ij}$ is the baseline vector, $\hat{s}$ is a unit-vector in the direction of a source in the sky, and λ is the wavelength of the signal. The fractional term in the exponential reflects the changing number of wavelengths between the two antennas as a signal goes in and out of phase as the source passes overhead. The entire exponential term represents the phase of the visibility, which can also be described as the fringe pattern, or diffraction, or interference pattern, between two antenna elements.

In Equation (1.5), the amplitude of the sky is broken up into a primary beam component $A(\nu, \hat{s})$ and sky intensity component $I(\nu, \hat{s})$. The primary beam describes the power pattern of an antenna element and determines its field of view, and combined with the sky intensity can be thought of as the "perceived intensity". The total power received by an antenna is a combination of the intensity distribution on the sky and how receptive the antenna is, or more specifically, the convolution between the two terms (Thompson et al. 2001).

The visibility equation can be re-interpreted as the 2-dimensional Fourier-transform of the sky, or a sample of the uv-plane, where u and v are spatial frequencies (more specifically, they are East/West and North/South line-of-sight projections of a baseline towards a phase center). In other words, every baseline measures a different Fourier-mode of the sky. To form an image, the Fourier-transform of a visibility would produce a *dirty image* of the sky, from which the true sky can be reconstructed by de-convolving out information from the antenna beam. In this thesis, however, we focus on the 3D Fourier-transform of the sky, or the power spectrum (Chapter 1.2.2), instead of making images. Hence, we work directly with visibilities as a starting point, which has already taken two Fourier-transforms for us.

1.2.2 The 21 cm Power Spectrum

In this thesis, we focus on cross-correlations, or power spectral measurements, of visibilities. Recalling that we seek to measure the differential brightness temperature on various spatial scales of the sky, we can form the quantity:

$$\langle \delta\tilde{T}_b(\vec{k})^* \delta\tilde{T}_b(\vec{k}) \rangle = (2\pi)^3 \delta^D(\vec{k} - \vec{k}') P_{21}(\vec{k}), \tag{1.6}$$

where $\delta\tilde{T}_b(\vec{k})$ is the Fourier-transform of the differential sky brightness as a function of cosmological wavenumber $\vec{k}$ (i.e., our visibility measurement, up to scaling factors), δ^D is the Dirac-delta function, and P_{21} is the 21 cm power spectrum quantity we are interested in eventually forming.

Simply speaking, because our visibility measurements have already taken two spatial Fourier-transforms out of the three needed for a 3D power spectrum, we need only to take one last Fourier-transform (along frequency), and then multiply and average the visibilities together for a given baseline in order to compute a power spectrum measurement. Having repeated baseline copies then increases the sensitivity to a given Fourier-mode on the sky, while having different types of baselines makes it possible to measure multiple Fourier-modes and build up an image of the sky. Since the EoR signal is expected to be present everywhere on the sky, in this work we focus on the former technique in order to maximize our sensitivity to the cosmological signal.

We note that the wavenumber $\vec{k}$ can be broken up into a perpendicular component $\vec{k}_\perp$ and a parallel component $k_\parallel$, where $\vec{k}_\perp$ is proportional to the (x,y) spatial coordinates on the sky and $k_\parallel$ is proportional to the line-of-sight direction on the sky (i.e., frequency). Every unique baseline probes a single $\vec{k}_\perp$, and it's worth noting that, because we focus on redundant baselines in the analysis to follow, most of our power spectrum sensitivity comes from the frequency-direction. Accounting for cosmological distance, a 1D wavenumber has units of Mpc^{-1}, so that the 3D power spectrum has units of mK^2Mpc^3. Visibility measurements typically have units of Janskys.

Just as the shape of the global signal provides insight about the early Universe, the shape of the cross-power spectrum, as defined by Equation (1.6), also delivers a wealth of information. Figure 1.4 shows the 21 cm "dimensionless" power spectrum $\Delta^2(k)$ (units of mK), defined as:

$$\Delta^2(k) = \frac{k^3}{2\pi^2} P_{21}(k), \tag{1.7}$$

as a function of the magnitude of k. This figure shows the expected evolution of the power spectrum, where the overall signal moves to small scales (large k) as more hydrogen becomes ionized (the large regions of neutral hydrogen turn into smaller and smaller pockets). This effect can be seen by both the steepening of the spectrum as the neutral fraction decreases, and the time-evolution of the spectrum at a specific (large) k.

The 21 cm power spectrum therefore encodes important information about the spatial and temporal evolution of reionization, and the shape of the spectrum can be directly mapped to sizes of the ionized bubbles as they grow. Additionally, the power spectrum, which is a function of the differential brightness temperature, can be used to constrain both T_{spin} (and T_{gas}, since they're coupled during this era) and x_{HI} via Equation (1.4). It is thus a powerful tool that will enable constraints to be made on the properties of the IGM.

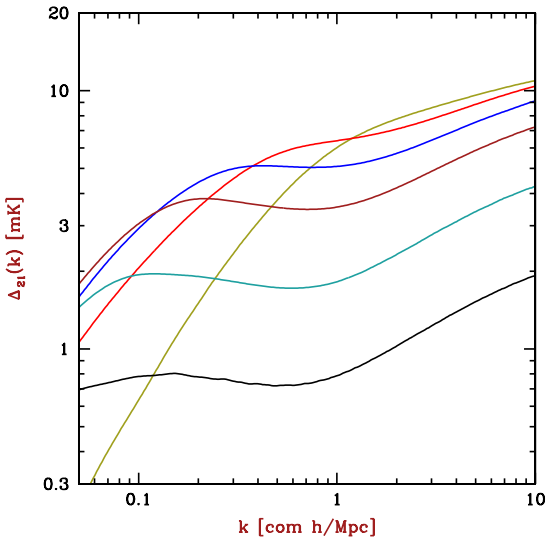

Figure 1.4: The theoretical evolution of the cross-21 cm power spectrum for a specific model (image credit: Barkana (2009)), where the neutral fraction x_{HI} = 10%, 30%, 50%, 70%, 90%, and 98% from top to bottom at large k. This figure shows the expected evolution of the power spectrum which interferometers seek to measure.

1.2.3 Calibration

We now transition to a brief overview of key data processing steps that are often standard routines when going from visibilities to power spectra. We speak broadly about these steps in this chapter, and go into detail about them for specific experiments in later chapters.

As discussed previously, an interferometer measures a visibility for every baseline pair and every time integration. Repeated baseline types and multi-day observations can be stacked to gain sensitivity to Fourier-modes on the sky. However, ensuring clean measurements at EoR sensitivities requires many other crucial processing steps, including calibration, which we give an overview of in this section.

We've seen that the visibility measurement is dependent on the sky, baseline, and antenna beam, but it is also affected by instrumental systematics. For example, certain components in the signal chain of an instrument can contribute variable amounts of noise, additional interfering signals can arise from reflections, and blockage and scattering may occur. These are all effects that must be considered in order to accurately extract the EoR signal.

Many of these effects can be mitigated through precise calibration. There are two main types of calibration used by 21 cm interferometers: *redundant* calibration and *absolute* calibration. We will briefly describe each here.

Redundant calibration is a type of calibration based on the redundant nature of baselines. For experiments like PAPER and HERA, which have many repeated copies of baselines, redundant calibration can be used to bring all identical baselines into agreement (i.e., it calibrates out deviations between baselines of the same type). This calibration is a powerful

technique because it does not use any knowledge of the sky, yet can correct for instrumental-induced gain and phase effects brought on by differences in the signal chain attributable to antennas, cables, and receivers (Liu et al. 2010).

Mathematically, the visibility v_{ij} of every baseline can be written as:

$$v_{ij} = g_i^* g_j y_{ij} + n_{ij}, \tag{1.8}$$

where g_i and g_j are the complex gains of each antenna, y_{ij} is the "true" model visibility for that particular baseline type, and n_{ij} is noise. The goal of redundant calibration is to solve for the gain of each antenna and the "true" visibility of each baseline type. This can be accomplished by setting up a system of linear equations containing every visibility measurement (the method used by PAPER and HERA is detailed in Chapter 3.1). If there are more measurements than the sum of the number of unique baselines and antennae, then it is a solvable system.

The complex gains can be further broken down to be written as $g_i = e^{\eta_i + i\phi_i}$ (Liu et al. 2010). In other words, by solving for the gains, we are solving for both an amplitude component and a phase offset for each antenna. The gains can then be divided out of every visibility measurement, producing redundantly-calibrated measurements across the whole array.

While redundant calibration is a clever technique for the internal calibration of an interferometer, the calibrated visibility measurements are still on an arbitrary gain scale that has not been matched up to the sky. Hence, absolute calibration refers to using sources in the sky of known brightness (or sky models) in order to solve for the four remaining internal degrees of freedom: an overall gain, an overall phase, and the tip and tilt of the array. Interferometers typically use a standard self-calibration routine to accomplish this, where y_{ij} is known for specific sky models.

Ultimately, calibration is a crucial step in preparing interferometric data for a power spectrum analysis. Precise calibration helps to smooth measurements and minimize the interaction between our chromatic instrument and spectrally smooth foregrounds. Our goal, after all, is an accurately-calibrated instrument that will result in cleaner data from which the EoR signal can be accessed.

1.2.4 Foreground Filtering

Arguably the largest challenge of processing 21 cm data is in removing bright foregrounds. There are several techniques to do this, which fall into two main categories: foreground subtraction and foreground avoidance. The former consists of modeling and subtracting out foreground sources, while the latter involves making EoR measurements in a domain where foregrounds are minimal.

For interferometers with imaging capabilities, foreground removal techniques include modeling approaches to spatially localize and remove contaminants (e.g., Santos et al. 2005; Wang et al. 2006; Jelić et al. 2008; Liu et al. 2009; Bowman et al. 2009; Harker et al. 2009; Chapman et al. 2016). This can be done by fitting polynomials to data or by using

non-parametric methods, which make fewer assumptions about the form of the foregrounds. While foreground subtraction would be ideal if done accurately, modeling is difficult and subtraction poses the risk of cosmological signal loss.

The other method commonly used, foreground avoidance, is a strategy employed by both PAPER and HERA. Foreground avoidance was originally suggested as an alternate method to the subtraction method, which has stringent requirements in order to yield uncontaminated results. In order to understand foreground avoidance, we must first define the "EoR Window".

A 3D power spectrum can be split into two directions along $k_\perp$ and $k_\parallel$, which correspond to modes perpendicular to the line-of-sight and along the line-of-sight, respectively. In this two-dimensional space, there are two main regions — one relatively free of foregrounds (the "EoR Window") and one contaminated by foregrounds (the "wedge"), as shown in Figure 1.5 (e.g., Datta et al. 2010; Vedantham et al. 2012; Pober et al. 2013). We can see this by thinking of the $k_\parallel$ direction to be akin to the physical time delay associated with light hitting two antennas (a good approximation, especially for short baselines). As $k_\perp$, which is proportional to baseline length, increases, the maximum time delay also increases because it is set by the length of the baseline (i.e., a maximum delay occurs when a source is at the horizon; therefore, the time delay is simply the time it takes for the light to travel the distance of the baseline). Hence, the "wedge" is formed, representing a region where smooth-spectrum foregrounds are expected to be contained. Said differently, foregrounds are expected to be bound by the light-crossing time between two antennas, and therefore there is a maximum limit for $k_\parallel$ (time delay) given a $k_\perp$ (baseline).

Delay-filtering is the process by which foregrounds within the wedge are filtered out, leaving a relatively clean window behind from which the cosmological signal can be extracted (Parsons et al. 2012). This approach can suffer from some foreground leakage as explained in Chapter 2.5, but its advantages include its simplicity and conservativeness (i.e., it leaves all the cosmological signal within the window unaltered).

Whether it's avoidance or removal, there have been many approaches at tackling the challenge of foregrounds (summarized in Chapman et al. (2016)). But, regardless of the method used, removing Galactic and extragalactic foregrounds from 21 cm data is absolutely a critical step for analyses seeking the EoR signal.

1.2.5 Fringe-rate Filtering

The final analysis technique to introduce in this section is fringe-rate filtering, a filtering scheme carried out in a domain which is the Fourier-dual to time. This type of filtering aims to optimize the process of combining time-ordered data and has been investigated in Roshi & Perley (2003), Parsons & Backer (2009), Offringa et al. (2012), and Parsons et al. (2016).

A "fringe-rate" is the rate at which the sky moves relative to the fringe pattern of an interferometer. As sources pass overhead, they walk in and out of the interference pattern of two antennas, and the rate at which this movement happens is dependent on the source's declination and hour angle. For example, a source located near a celestial pole has a zero

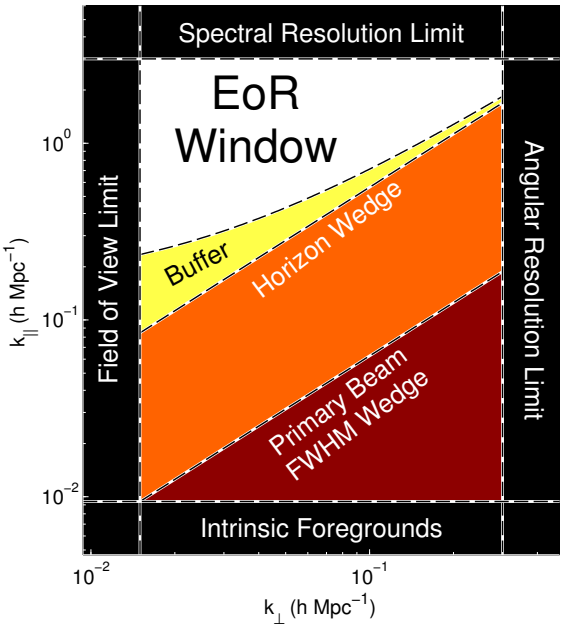

Figure 1.5: A cartoon diagram of the "EoR Window" and "wedge" of foreground contamination in Fourier space (image credit: Dillon et al. (2015b)). A foreground avoidance approach makes power spectrum measurements in the window, while a foreground subtraction approach subtracts out foregrounds so that measurements can be made in the wedge. The overall power spectrum measurement space is limited by an interferometer's field of view and angular resolution along the horizontal axis, and spectral resolution and intrinsic foregrounds along the vertical axis.

fringe-rate, as it does not appear to move in the sky as the Earth rotates. However, a source located on the celestial equator will have a maximum fringe-rate set by the rate of Earth's rotation. The sky can therefore be decomposed into fringe-rate bins, each of which forms a concentric circle of constant fringe-rate around the celestial sphere.

One initial advantage of filtering in fringe-rate space is that it allows the filtering of noise that is not associated with movement locked on the sky (i.e., filtering out fringe-rates greater than the maximum allowed by the Earth's rotation). This excess noise can come from the instrument itself or from signals with an origin not on the sky. Additionally, one can up-weight and down-weight certain portions of the sky by choosing different linear combinations of fringe-rates (Parsons et al. 2016). This allows what is effectively a beam-sculpting operation, where the most sensitive parts of one's beam can be up-weighted compared to others.

By weighting data in the fringe-rate domain, time-integration of measurements can be achieved. Depending on the shape of the filter in the fringe-rate domain, the effect in the time domain can be an averaging operation along time. This is advantageous because it allows an optimal way of combining measurements (by weighting fringe-rates differently based on signal-to-noise ratios in each fringe-rate bin, for example) compared to a more traditional boxcar average, which does not use information from individual fringe-rate bins.

Broadly, fringe-rate filtering can be thought of as a tailored filtering step that can increase the sensitivity of a measurement by differentiating between noise- and signal-like modes in data. When carefully chosen, a fringe-rate filter can enhance modes containing emission from the celestial sphere, where the 21 cm signal lies.

1.3 Instruments

The recent exploration of our cosmic dawn has led to the development of multiple experiments that are aiming to detect the 21 cm signal from neutral hydrogen during reionization. In this section we will highlight the two main radio interferometers focused on in this thesis, and then discuss other similar experiments along with the current status of the field.

1.3.1 The Precision Array for Probing the Epoch of Reionization

The Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER) is a first generation EoR experiment. Its history dates back to 2007, when an initial four dipole antennas observed the sky from Western Australia. A year later, the array increased to eight stations and moved to Green Bank, West Virginia. These first two deployments are summarized in Parsons et al. (2010) and were used to characterize important aspects of the instrument, including system performance, beam models, instrumental temperatures, and sensitivity to radio frequency interference (RFI).

PAPER then moved to the Karoo Desert in South Africa, near the Square Kilometre Array South Africa (SKA-SA). The PAPER array doubled in size each year, starting with

Figure 1.6: A PAPER antenna in the Karoo Desert in South Africa. A dual-polarization dipole sits at the center, surrounded by wire mesh panels that measure 2 m on each side.

32 antennas in 2011 and ending with 128 a few years later. PAPER's observing seasons using these three arrays (PAPER-32, PAPER-64, and PAPER-128) have primarily been used to develop analysis techniques, understand instrumental design, and begin to place limits on the EoR and connect them to science implications.

A brief overview of the PAPER instrument and digital backend follows. The PAPER dipole itself (Figure 1.6) is made from two rods of copper sandwiched between two aluminum disks. The sizes of each are fine-tuned to produce an antenna frequency response between 100-200 MHz. Each PAPER antenna is sensitive to two orthogonal polarizations, those being the East/West and North/South directions (XX and YY linear polarizations) given the antenna's orientation on the ground. A grounding structure, made of wire-mesh and held in place by PVC pipes, is both underneath the dipole and surrounds it as four angled panels. The design of the antenna's framework was driven by the desire to produce spectrally and spatially smooth beam responses as discussed in Parsons et al. (2010) and Pober et al. (2012). Altogether, an entire PAPER dipole measures about 2 m on each side and sits still while the Earth's rotation moves the sky above ("drift-scan" mode). When photons hit the dipoles' copper rods, electrons are excited and their movement turns the electric field into a voltage that can be measured.

In short, PAPER's analog system consists of a balun attached to each dipole element (which measures the voltage and amplifies the signal) and coaxial cables which transport the antenna signals. The signals then travel to dual-channel receiver boards which are cooled inside thermal enclosures in order to prevent the introduction of high gains from temperature fluctuations. The receivers both amplify and filter (in frequency) the signals before sending them to the digital system.

PAPER's digital system is housed inside a refrigerated container on the observing site.

The bulk of the processing is carried out by a series of real-time digital FX correlators which cross-correlate pairs of antenna signals to form visibility measurements. More specifically, the FX correlators comprise of "F-engines" which digitize, down-convert, and channelize antenna inputs, a switch that divides the data into frequency subsets and routes the packets, GPUs which cross-multiply signals from all antenna pairs and integrate in time ("X-engines"), and finally another switch and computer which collect the data, write it to disk, and send the final products over an ethernet connection as "raw" visibility data products that are ready to be analyzed.

Much of PAPER's digital signal processing (DSP) system has been made possible due to the development of hardware by the Center for Astronomy Signal Processing and Electronics Research (CASPER; Parsons et al. 2006; Hickish et al. 2016). We rely on their Field-programmable Gate Array (FPGA) processors which can quickly perform the fast Fourier-transforms required to produce visibilities. The development of FPGA's into PAPER's DSP system has been critical in allowing for the scalable expansion of the array.

PAPER's first power spectrum limits on the EoR came from its 32-element array (Parsons et al. 2014). PAPER-32 observed in a redundantly-configured layout from 2011-2012, producing a published 2σ upper limit on the 21 cm power spectrum of $(41\,\mathrm{mK})^2$ for $k = 0.27\,h\,\mathrm{Mpc}^{-1}$ at $z = 7.7$ (Parsons et al. 2014). This result, while orders of magnitude above predicted EoR signals, was used to generate constraints on the brightness temperature of 21 cm emission for various reionization models and rule out cold reionization scenarios (i.e., some heating of the IGM is necessary by $z = 7.7$ to be consistent with PAPER-32's results).

PAPER expanded to 64 elements in 2012, keeping its redundant layout in order to maximize power spectrum sensitivity. The analysis and initial results for PAPER-64's observing season is outlined in Ali et al. 2015, where a 2σ upper limit on the EoR is published as $(22\,\mathrm{mK})^2$ for $0.15 < k < 0.5\,h\,\mathrm{Mpc}^{-1}$ at $z = 8.4$. A result at this sensitivity can begin to place more interesting limits on IGM heating models and on the temperature of the IGM during this time (Pober et al. 2015).

From 2013-2015, PAPER-128 marked the last era for the PAPER experiment. The data collected with this array has not been published publicly. While most of this thesis focuses on PAPER-64, specifically on analysis methods developed to revise the initial (incorrect) PAPER-64 results (Chapters 2 and 3) and the presentation of revised limits (Chapter 4), we also present a first-look at PAPER-128 and discuss how PAPER's final observing season has influenced analysis metrics for next generation experiment HERA (Chapter 5).

The PAPER experiment as a whole has been absolutely fundamental to the growth of the field of 21 cm cosmology. This first generation experiment set a standard for other similar experiments and has provided countless lessons on all aspects of the signal chain. The array may be retired, but its influence will not be forgotten.

1.3.2 The Hydrogen Epoch of Reionization Array

The development of the Hydrogen Epoch of Reionization Array (HERA) was largely driven by the need for increased sensitivity, as even PAPER-128 lacked the collecting area

Figure 1.7: A HERA dish in the Karoo Desert in South Africa. Wire-mesh, PVC pipes, and wooden structures serve as the foundation for the 14 m diameter parabola. A PAPER dipole is suspended upside-down with a wire pulley-system and surrounded by a prototype wire-mesh skirt structure. HERA-350 will use an updated design for its feed; however, HERA's initial data releases use the old PAPER infrastructure as depicted here.

for a significant detection of the EoR. HERA is a second generation EoR experiment currently being built in the Karoo. It features a staged build-out of parabolic dishes with 14 m diameters, with construction beginning in 2015 and an eventual 350 dishes planned to be completed by the end of 2019.

A HERA dish (Figure 1.7) is made up of wire-mesh, PVC pipes, and wooden support structures. The size, shape, and total number of the dishes were chosen in order to optimize sensitivity (i.e., minimize chromatic effects that would leak power into the EoR window), minimize costs, and be easily scalable and robust for a five year lifetime (DeBoer et al. 2017). While work on a new feed design (with a wider bandwidth) is ongoing, the first observations from HERA use recycled PAPER dipoles. Suspended upside-down with a wire pulley-system, the PAPER dipoles are surrounded by a wire-mesh backplane structure that minimizes cross-coupling between antennas while optimizing beam efficiency, frequency response, and polarization match (DeBoer et al. 2017). Similarly, the first stages of the HERA array are using the existing PAPER signal chain and hardware, while work is progressing towards an underground node-based architecture that will house the DSP system and minimize cable reflections by allowing for shorter cable paths.

The configuration of HERA, like PAPER, is highly redundant and optimized for a robust foreground avoidance approach. Because this power spectrum approach requires short baselines (which minimize the wedge), the HERA antennas are densely-packed next to each other into a main core. This core is segmented into three displaced sections, whose sectioning is

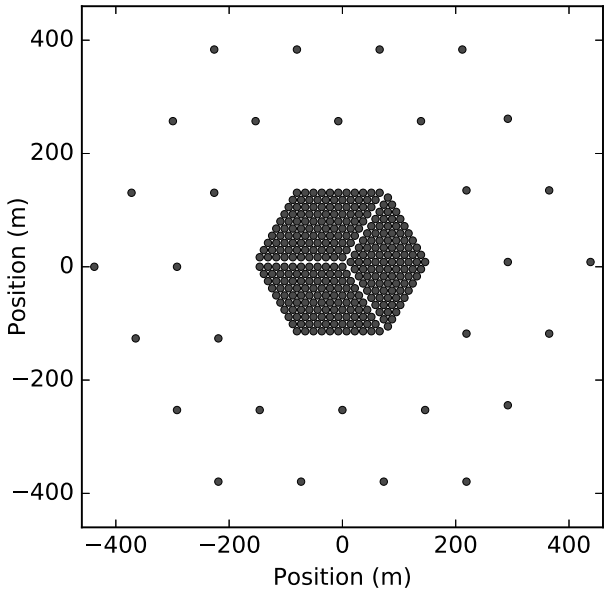

Figure 1.8: The full HERA-350 array (image credit: DeBoer et al. (2017)). The array is comprised of a segmented densely-packed core (to optimize redundancy for a foreground avoidance approach) and surrounding outrigger elements (for imaging capabilities).

designed to improve HERA's imaging ability (Dillon & Parsons 2016). Additionally, there will be 30 outrigger elements joining the full array, allowing for a more complete uv-plane coverage and imaging capabilities that can be leveraged for a foreground removal approach. The full HERA array is depicted in Figure 1.8.

The primary science goal of HERA is to make a high-significance detection of the cosmological signal in order to constrain the timing and morphology of reionization. With precision constraints on the EoR, we can begin to understand the role of the first stars and galaxies in driving reionization, their complex interactions with their environments, and the evolution of cosmic structure. With the full array and an upgraded feed sensitive from 50-250 MHz, secondary science goals of the instrument include precision cosmology, imaging of the reionization epoch, and an investigation of the pre-reionization heating epoch (DeBoer et al. 2017). HERA, with a collecting area of $\sim 0.1\,\mathrm{km}^2$, is well-poised for these challenges (Pober et al. 2014) and has already undergone numerous tests and simulations to ensure that its design meets specification (Neben et al. 2016; Ewall-Wice et al. 2017; Patra et al. 2018). As the fully realized HERA soon approaches, there is much to look forward to in the coming years from this array.

1.3.3 Other Experiments and Status of Field

Although this thesis focuses mostly on PAPER data and analysis methods that will be used by HERA, these two arrays are not alone in their quest for the EoR. Other radio

interferometers which seek to measure statistical power spectra include the Giant Metre-wave Radio Telescope located in India (GMRT; Paciga et al. 2013), the LOw Frequency ARray in Europe (LOFAR; van Haarlem et al. 2013), the Murchison Widefield Array in Australia (MWA; Tingay et al. 2013), the 21 Centimeter Array in China (21CMA; Pen et al. 2004; Wu 2009), and the Square Kilometre Array in South Africa (SKA; Koopmans et al. 2015).

Several of these experiments have succeeded in placing upper limits on the EoR, including results from the 32-tile MWA (Dillon et al. 2014), 128-tile MWA (Dillon et al. 2015a; Beardsley et al. 2016), GMRT (Paciga et al. 2013), and LOFAR (Patil et al. 2017). PAPER has also previously published results using 32 antennas (Parsons et al. 2014; Jacobs et al. 2015) and 64 antennas (Ali et al. 2015), though we highlight the errors found in PAPER's analysis pipeline throughout this thesis and thus refer the reader to Chapter 4 for updated results from PAPER.

The work in the 21 cm community that has led to these power spectrum limits (Figure 1.9) has largely revolved around the key challenge of controlling foregrounds and systematics. To accomplish this, significant progress has been made in all aspects of the experimental process, ranging from carefully designed interferometers (Lonsdale et al. 2009; Parsons et al. 2012; Dillon & Parsons 2016), to novel methods for understanding and dealing with foregrounds (e.g., Morales et al. 2006; Datta et al. 2010; Sullivan et al. 2012; Moore et al. 2013; Hazelton et al. 2013; Pober et al. 2013; Liu et al. 2014a; Liu et al. 2014b; Thyagarajan et al. 2015b), to statistical analysis techniques for precise calibration and power spectrum estimation (e.g., Liu et al. 2010; Trott et al. 2012; Liu et al. 2014b; Zheng et al. 2014; Dillon et al. 2014; Jacobs et al. 2016). PAPER's foreground avoidance strategies have, in particular, led to detailed understandings of redundant calibration and the effects of filtering on the EoR window, while MWA's foreground subtraction techniques have provided complementary improvements in imaging capabilities. While the experiments with published results currently lack the sensitivities needed for an EoR detection, both the delay-filtering and map-making methods, along with hybrid approaches (Trott et al. 2016), have set a strong foundation for controlling foregrounds by future, more sensitive interferometers.

Related to the fundamental goal of simultaneously maximizing sensitivity and minimizing contaminants, some other challenges that face current 21 cm experiments include polarization leakage from Faraday-rotated emission (Moore et al. 2013; Kohn et al. 2016; Nunhokee et al. 2017), direction-dependent beam effects, and other low level sources of chromaticity induced by the instrument or calibration. These effects will require thorough investigations as experiments approach EoR sensitivities.

In addition to power spectrum experiments, there are several complementary experiments that aim to measure the sky-averaged global 21 cm signal (i.e., the mean brightness temperature of the EoR relative to the CMB). These include the Experiment to Detect the Global EoR Signature (EDGES; Bowman & Rogers 2010), the Large Aperture Experiment to Detect the Dark Ages (LEDA; Bernardi et al. 2016), the Dark Ages Radio Explorer (DARE; Burns et al. 2012), the Sonda Cosmológica de las Islas para la Detección de Hidrógeno NeutroSciHi (SCI-HI; Voytek et al. 2014), the Broadband Instrument for Global HydrOgen ReioNisation Signal (BIGHORNS; Sokolowski et al. 2015), and the Shaped Antenna measurement of the

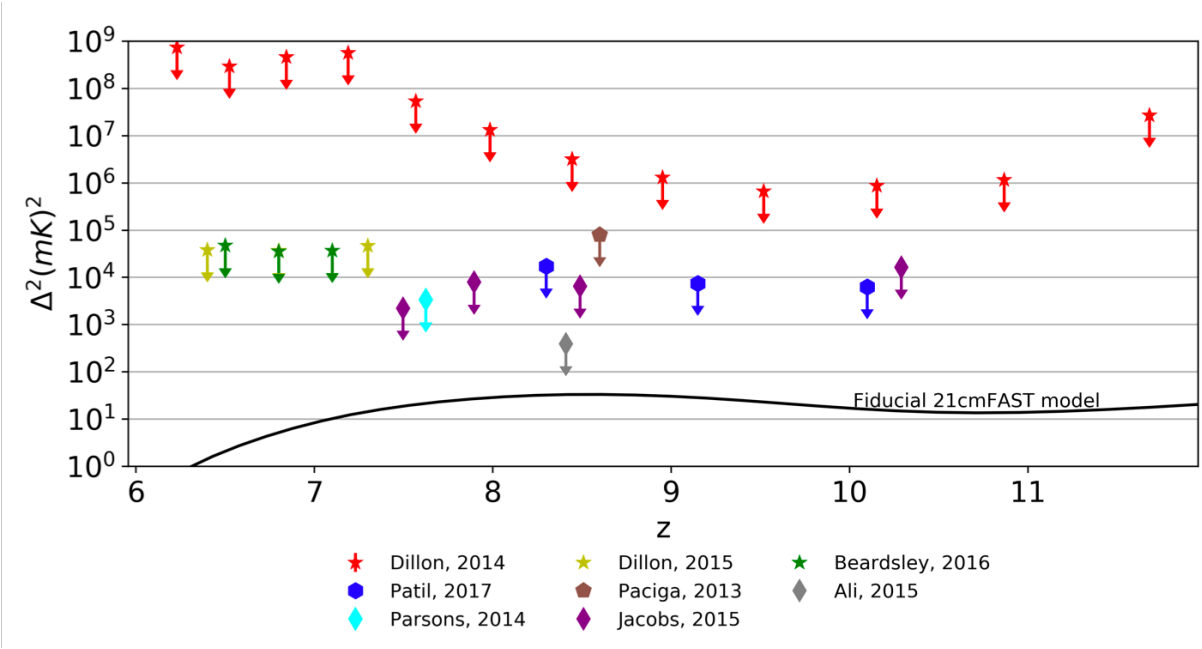

Figure 1.9: Published upper limits on the EoR placed by different 21 cm experiments, prior to the work in this thesis. All PAPER results shown (PAPER-32 is in cyan and magenta, and PAPER-64 is in gray) are suspect to the errors discussed throughout this work and are superseded by the ones presented in Chapter 4.

background RAdio Spectrum (SARAS; Patra et al. 2015).

Like the power spectrum experiments, global signal experiments also face challenges of bright foregrounds and instrumental systematics. In particular, they require extremely precise calibration in order to avoid overfitting when subtracting foregrounds, and they also face additional challenges when observing the lowest frequencies, such as ionospheric fluctuations and brighter foregrounds (Vedantham et al. 2014; Datta et al. 2014).

However, most global signal telescopes consist of single elements and can therefore be more easily constructed. Additionally, EoR sensitivities can be reached with relatively short observations. Thus, global signal experiments are actively being worked on as a complementary view into the evolution of the 21 cm signal. Accurate detections of the features of the global signal will delineate the different epochs in the early Universe, and their shapes will carry important implications about the nature of the IGM and the first sources.

For example, the first potential detection of the 21 cm signal has been recently made by the global signal experiment EDGES (Bowman et al. 2018). This exciting result suggests the presence of an absorption feature in the sky-averaged spectrum at 78 MHz, thought to be the result of the absorption of CMB photons by HI gas. Because the detected feature is best-fit by an amplitude much larger than what is consistent with expectations for the 21 cm signal during this epoch, alternate scientific explanations have been offered, such as the influence of dark matter on baryons and the effect of their interaction on the temperature of the gas (Barkana 2018; Slatyer & Wu 2018). Measurements from other experiments in the future will certainly be informative in shaping the community's confidence in this detection, and this result marks just the beginning of many more in the field to come.

1.4 This Thesis

Although 21 cm observations promises an uninterrupted window into the EoR, from which we can learn much about galaxy formation and the properties of the IGM, there are many challenges facing this field of cosmology. In general, the 21 cm signal is extremely faint, with bright foregrounds (mostly synchrotron radiation from our own Galaxy) and radio interference easily overshadowing the target signal. As a consequence, instruments need to be extremely well-understood, precisely calibrated, and sensitive enough for a successful detection. In addition, analysis techniques must be innovative and rigorously construed so as to be able to extract clean and accurate measurements.

Broadly speaking, in this thesis I present work associated with data from radio interferometers seeking to measure 21 cm fluctuations during the EoR. While a confirmed detection by an interferometer remains elusive at this time, this work serves as a huge leap forward in working with large data sets and extracting measurements of the cosmological signal with confidence.

More specifically, a major theme throughout this thesis is the unexpected discovery of errors in previous PAPER analyses. Subtle mathematical shortcuts, unanticipated interactions between techniques, and unfortunate oversights combine together to motivate new methods,

deeper understandings, and careful revisions. In addition to the practical, technical insights offered in this thesis, we also make a point throughout to advertise that scientific progress comes in many forms, even those motivated by mistakes and errors.

The majority of the rest of this thesis is focused on the analysis of data from PAPER. Chapter 2 develops intuition behind our power spectrum analysis methods and choices. Chapter 3 applies lessons learned to a subset of PAPER-64 data, and Chapter 4 builds off of both to place revised limits on the 21 cm power spectrum from PAPER-64. Finally, we conclude with an overview of PAPER's final observing season in Chapter 5 and a preliminary look into ongoing work related to HERA in Chapter 6. Taken as a whole, this thesis represents deep analysis investigations that have resulted in nuanced understandings and a raised bar for a growing field.

Chapter 2

Power Spectrum Methods

2.1 Power Spectrum Themes and Techniques

The major challenge that faces all 21 cm experiments is isolating a small signal that is buried underneath foregrounds and instrumental systematics that are, when combined, four to five orders of magnitude brighter (e.g., Santos et al. 2005; Ali et al. 2008; de Oliveira-Costa et al. 2008; Jelić et al. 2008; Bernardi et al. 2009, 2010; Ghosh et al. 2011; Pober et al. 2013; Bernardi et al. 2013; Dillon et al. 2014; Kohn et al. 2016). A clean measurement therefore requires an intimate understanding of the instrument and a rigorous study of data analysis choices. With continual progress being made in the field and HERA on the horizon, it is becoming increasingly important to understand how the methods we choose interact with each other to affect power spectrum results. More specifically, it is imperative to develop techniques and tests that ensure the accuracy and reliability of a potential EoR detection. In this chapter and the next, we discuss three topics (signal loss, error estimation, and bias) that are essential to investigate for a robust 21 cm power spectrum analysis. We also highlight four power spectrum techniques (fringe-rate filtering, weighting, bootstrapping, jackknife testing) and their trade-offs, potential pitfalls, and connections to the themes. We first approach the themes from a broad perspective (Chapter 2), introducing the themes of our focus and using toy models to develop intuition into each one. We then perform a detailed case study (Chapter 3) using data from the 64-element configuration of PAPER, highlighting key changes from the methods used in the previously published result in Ali et al. (2015), henceforth known as A15, which have led to a revised PAPER-64 power spectrum result. In these two chapters we use a subset of PAPER-64 data to illustrate our revised analysis methods, while Chapter 4 builds off of the methods to finally present revised PAPER-64 results for multiple redshifts and baseline types.

We note that this thesis as a whole adds to the growing foundations of lessons which have been documented, for example, in Paciga et al. (2013), Patil et al. (2016), and Jacobs et al. (2016), by the GMRT, LOFAR, and MWA projects respectively. These lessons are imperative as the community moves towards higher sensitivities and potential EoR detections.

There are many choices a 21 cm data analyst must consider. How can time-ordered

measurements be combined? How can the variance of the data be estimated? In what way(s) can the data be weighted to suppress contaminated modes while not destroying an EoR signal? How can a statistically significant detection of a signal be properly identified? Many common techniques, such as averaging data, weighting, bootstrapping, and jackknife testing, address these issues but harbor additional trade-offs. For example, an aggressive filtering method may succeed in eliminating interfering systematics but comes at the cost of losing some EoR signal. A chosen weighting scheme may theoretically maximize sensitivity but fail to suppress foregrounds in practice.

Though there are many data analysis choices, measuring the statistical 21 cm power spectrum ultimately requires robust methods for determining accurate confidence intervals and rigorous techniques to identify and control systematics. In this thesis, we focus on three 21 cm power spectrum themes that encapsulate this goal and discuss four techniques that interplay with each other and impact the themes. We will give brief definitions now, and build intuition for each theme in the sections to follow.

- **Signal Loss** (Chapters 2.2 and 2.3): Signal loss refers to attenuation of the target *cosmological* signal in a power spectrum estimate. Certain analysis techniques can cause this loss, and if the amount of loss is not quantified accurately, it could lead to false non-detections and overly aggressive upper limits. Determining whether an analysis pipeline is lossy, and estimating the amount of loss if so, has subtle challenges but is necessary to ensure the accuracy of any result.

- **Error Estimation** (Chapter 2.4): Confidence intervals on a 21 cm power spectrum result determine the difference between a detection and a null result, which have two very different implications. Additionally, accurate error estimation is crucial for the comparison of results to theoretical models. Errors can be estimated in a variety of ways, and we will discuss a few of them.

- **Bias** (Chapter 2.5): There are several possible sources of power offset in a visibility measurement that can show up as a detection in a power spectrum, such as bias from noise and foregrounds. In particular, a successful EoR detection would also imitate a bias. Proving that a bias is an EoR detection may be the most difficult challenge for future 21 cm analyses, as it is crucial to be able to distinguish a detection of foreground leakage, for example, from that of EoR. In this chapter, we will highlight some sources of bias, discuss ways to mitigate their effects, and describe example tests that a true EoR detection must pass.

The following techniques each have advantages when it comes to maximizing sensitivity and understanding systematics in data. However, some have limitations, and we will discuss circumstances in which there are trade-offs. We choose to focus on these four techniques because they represent major steps in PAPER's power spectrum pipeline, with several of them also being standard steps in general 21 cm analyses.

- **Fringe-rate filtering:** Fringe-rate filtering is an averaging scheme for time-ordered data (Parsons et al. 2016). Broadly, a fringe-rate filter, as described in more detail in Chapter 1.2.5, averages visibilities in time to produce a smaller number of more sensitive independent samples. However, such a filter also affects the presence of foregrounds and systematics. We explain the trade-offs of filtering in more detail in Chapter 2.2.3.

- **Weighting:** A data set can be weighted to emphasize certain features and minimize others. One particular flavor of weighting employed by previous PAPER analyses is inverse covariance weighting in frequency, which is a generalized version of inverse variance weighting that also takes into account frequency correlations (Liu & Tegmark 2011; Dillon et al. 2013; Liu et al. 2014a; Liu et al. 2014b; Dillon et al. 2014; Dillon et al. 2015a). Using such a technique enables the down-weighting of contaminant modes that obey a different covariance structure from that of cosmological modes. However, a challenge of inverse covariance weighting is in estimating a covariance matrix that is closest to the true covariance of the data; the discrepancy between the two can have large impacts on signal loss. We investigate the impact of different types of weighting on signal loss in Chapter 2.2.

- **Bootstrapping:** In addition to using theoretical models for covariance matrices and theoretical error estimation methods, bootstrapping is one way to estimate errors. Namely, bootstrapping is a useful method for estimating errors of a data set from itself (Andrae 2010). By randomly drawing many subsamples of the data, we obtain a sense of its inherent variance, though there are subtleties to consider such as the independence of values in a data set. We explore this potential pitfall of bootstrapping in Chapter 2.4.

- **Jackknife testing:** A resampling technique useful for estimating bias, jackknives can be taken along different dimensions of a data set to cross-validate results. In particular, null tests can be used to verify whether results are free of systematics, as done with CO power spectra (Keating et al. 2016) and CMB measurements (see e.g., Ade et al. 2008; Chiang et al. 2010; Bischoff et al. 2011; Das et al. 2011a; Araujo et al. 2012; Crites et al. 2015; BICEP2 Collaboration et al. 2016; Ade et al. 2017; Sherwin et al. 2017). An EoR detection must pass both jackknife and null tests, which we highlight in Chapter 2.5.2.

In this chapter, we study each theme in depth, focusing on how power spectrum technique trade-offs affect each. We use toy data models to develop intuition into why certain analysis choices may be appealing and discuss ways in which they are limited. We highlight problems that can arise regarding each theme and offer suggestions to mitigate the issues. Ultimately, we show that rigorous investigations into signal loss, error estimation, and bias must be performed for robust 21 cm results.

2.2 Signal Loss Toy Model

Signal loss can arise in a variety of ways in an analysis pipeline, such as by fitting a polynomial during spectral calibration, applying a delay domain filter, or deriving weights from data and applying them to itself. Here we focus on signal loss associated with the use of an empirically estimated covariance matrix with the "optimal quadratic estimator" formalism. This loss was significantly underestimated in the A15 analysis.

2.2.1 The Quadratic Estimator Method

We begin with an overview of the quadratic estimator (QE) formalism used for power spectrum estimation. The goal of power spectrum analysis is to produce an unbiased estimator of the EoR power spectrum in the presence of both noise and foreground emission. Prior to power spectrum estimation, the data will often have been prepared to have minimal foregrounds by some method of subtraction, so this foreground emission may appear either directly (because it was not subtracted) or as a residual of some subtraction process not in the power spectrum domain. If an accurate estimate of the total covariance of the data is known, including both the desired signal and any contaminants, then the "optimal quadratic estimator" formalism provides a method of producing a minimum variance, unbiased estimator of the desired signal, as shown in Liu & Tegmark (2011), Dillon et al. (2013), Liu et al. (2014a), Liu et al. (2014b), Trott et al. (2012), Dillon et al. (2014), Dillon et al. (2015a), Switzer et al. (2015), and Trott et al. (2016).

Suppose that the measured visibilities for a single baseline in Jy are arranged as a data vector, $\mathbf{x}$. It has length $N_t N_f$, where N_t is the number of time integrations and N_f is the number of frequency channels. The covariance of the data is given by

$$\mathbf{C} \equiv \langle \mathbf{x}\mathbf{x}^\dagger \rangle = \mathbf{S} + \mathbf{U} \tag{2.1}$$

where the average over an ensemble of data realizations produces the true covariance, and we further assume it may be written as the sum of the desired cosmological signal $\mathbf{S}$ and other terms $\mathbf{U}$.

We are interested in estimating the three-dimensional power spectrum of the EoR. Visibilities are measurements of the Fourier transform of the sky along two spatial dimensions (using the flat-sky approximation), and since we are interested in three-dimensional Fourier modes we only need to take one Fourier transform of our visibilities along the line-of-sight dimension. We consider band powers P^α of the power spectrum of $\mathbf{x}$ over some range in cosmological $\mathbf{k}$, where α indexes a waveband in $k_\parallel$ (a cosmological wavenumber $k_\parallel$ is the Fourier dual to frequency under the delay approximation (Parsons et al. 2012), which is a good approximation for the short baselines that PAPER analyzes). The fundamental dependence of the covariance on the power spectrum band powers P^α is encoded as

$$\mathbf{S} = \sum_\alpha P^\alpha \frac{\partial \mathbf{C}}{\partial P^\alpha} \equiv \sum_\alpha P^\alpha \mathbf{Q}^\alpha \tag{2.2}$$

where we define $\frac{\partial \mathbf{C}}{\partial P^\alpha} \equiv \mathbf{Q}^\alpha$. In other words, $\mathbf{Q}$ describes the response of the covariance to a change in the power spectrum, relating a quadratic statistic of the data (the covariance) to a quadratic statistic in Fourier-space (the power spectrum).

The optimal quadratic estimator prescription is then to compute

$$\widehat{P}^\alpha = \sum_\beta (\mathbf{F}^{-1})^{\alpha\beta}(\widehat{q}^\beta - \widehat{b}^\beta) \tag{2.3}$$

where $\mathbf{F}$ is the Fisher matrix (which determines errors on the power spectrum estimate)

$$F^{\alpha\beta} \equiv \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\beta\right), \tag{2.4}$$

$\widehat{q}$ is the un-normalized power spectrum estimate

$$\widehat{q}^\alpha = \frac{1}{2}\mathbf{x}^\dagger\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{x}, \tag{2.5}$$

and $\widehat{b}$ is the additive bias

$$\widehat{b}^\alpha = \frac{1}{2}\,\mathrm{tr}\left(\mathbf{U}\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\right). \tag{2.6}$$

The power spectrum estimator in Equation (2.3) is the minimum variance (smallest error bar) estimate of the power spectrum subject to the constraint that it is also unbiased; that is, the ensemble average of the estimator is equal to its true value

$$\langle\widehat{P}^\alpha\rangle = P^\alpha \tag{2.7}$$

(Tegmark 1997; Bond et al. 1998).

Intuitively, the estimator must be capable of "suppressing" or "removing" the effects of contaminants in order to obtain an unbiased estimate of the power spectrum. By construction, the subtraction of the residual foreground and noise bias accomplishes this, removing any additive bias. However, the $\mathbf{C}^{-1}$ piece of Equation (2.5) also has the effect of suppressing residual foregrounds and noise, in both the additive bias and any contributions the residuals may have to the variance. The way in which the optimal estimator accomplishes this is illustrated with a toy model in Chapter 2.3.1. There, we show that the effect of the weighting in Equation 2.5 is to project out the modes of $\mathbf{U}$ with a different covariance structure than $\mathbf{S}$ in the power spectrum estimate, and the effect of Equation 2.6 is to subtract out the remaining bias. Similar effects for a realistic model of the EoR and foregrounds are shown in Liu & Tegmark (2011).

The toy model in Chapter 2.3.1 also illustrates that if the covariance structure of the contaminants is sufficiently different from the desired power spectrum, then the linear bias term may be expected to be quite small, and it is only necessary to know $\mathbf{C}$ and $\mathbf{Q}^\alpha$, but not $\mathbf{U}$. Since the foregrounds are expected to be strongly correlated between frequencies whereas the EoR is not, we expect different covariance structures and therefore a small linear bias. Moreover, because the linear bias is always positive and there is no multiplicative bias, the

quadratic-only term will always produce an estimate which is *high* relative to the true value, and which can conservatively be interpreted as an upper limit. These considerations, and the difficulty of obtaining an estimate for $\mathbf{U}$, motivate the neglect of the linear bias in the rest of this analysis.

Motivated by the desire to retain the advantageous behavior of suppressing contributions of $\mathbf{U}$ to estimates of the EoR power spectrum, we note that is possible to define a modified version of the quadratic estimator where Equation (2.5) is replaced by

$$\widehat{q}^{\alpha} = \frac{1}{2}\mathbf{x}^{\dagger}\mathbf{R}\mathbf{Q}^{\alpha}\mathbf{R}\mathbf{x} \tag{2.8}$$

where $\mathbf{R}$ is a weighting matrix chosen by the data analyst. For example, inverse covariance weighting (the optimal form of QE) would set $\mathbf{R} \equiv \mathbf{C}^{-1}$ and a uniform-weighted case would use $\mathbf{R} \equiv \mathbf{I}$, the identity matrix. Again, the matrix $\mathbf{Q}^{\alpha}$ encodes the dependence of the covariance on the power spectrum but in practice also also does other things, including implementing a transform of the frequency domain visibilities to $\mathbf{k}$-space, taking into account cosmological scalings, and converting the visibilities from Jansky to Kelvin.

With an appropriate normalization matrix $\mathbf{M}$, the quantity

$$\widehat{\mathbf{P}} = \mathbf{M}\widehat{\mathbf{q}} \tag{2.9}$$

is a sensible *estimate* of the *true* power spectrum $\mathbf{P}$.

To ensure that $\mathbf{M}$ correctly normalizes our power spectrum, one may take the expectation value of Equation (2.9) to obtain

$$\begin{aligned} \langle \widehat{P}^{\alpha} \rangle &= \frac{1}{2}\sum_{\beta\gamma} M^{\alpha\gamma}\mathrm{tr}\left(\mathbf{R}\mathbf{Q}^{\gamma}\mathbf{R}\mathbf{Q}^{\beta}\right)P^{\beta} + \frac{1}{2}\sum_{\gamma}\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{Q}^{\gamma}\mathbf{R}\right) \\ &\equiv \sum_{\beta} W^{\alpha\beta}P^{\beta} + \frac{1}{2}\sum_{\gamma}\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{Q}^{\gamma}\mathbf{R}\right), \end{aligned} \tag{2.10}$$

where $W^{\alpha\beta}$ are elements of a window function matrix. Considering the first term of this expression (again, we are assuming that the linear bias term is significantly suppressed; and if this is not the case, we are simply assuming that we are setting a conservative upper limit), if $\mathbf{W}$ ends up being the identity matrix for our choices of $\mathbf{R}$ and $\mathbf{M}$, then we recover Equation (2.7) for the first term, and we have an estimator that has no multiplicative matrix bias. However, Equation (2.7) is a rather restrictive condition, and it is possible to violate it and still have a sensible (and correctly normalized) power spectrum estimate. In particular, as long as the rows of $\mathbf{W}$ sum to unity, our power spectrum will be correctly normalized. Beyond this, the data analyst has a choice for $\mathbf{M}$, and for simplicity throughout this thesis we choose $\mathbf{M}$ to be diagonal. In a preview of what is to come, we also stress that the derivation that leads to Equation (2.10) assumes that $\mathbf{R}$ and $\mathbf{x}$ are not correlated. If this assumption is violated, a simple application of the (now incorrect) formulae in this section can result in an improperly normalized power spectrum estimator that does not conserve power, i.e., one that has signal loss.

Given the advantages of inverse covariance weighting, a question arises of how one goes about estimating $\mathbf{C}$. One method is to empirically derive it from the data $\mathbf{x}$ itself. Similar types of weightings that are based on variance information in data are done in Chang et al. (2010) and Switzer et al. (2015). In previous PAPER analyses, one time-averages the data to obtain:

$$\widehat{\mathbf{C}} \equiv \langle \mathbf{x}\mathbf{x}^\dagger \rangle_t \approx \langle \mathbf{x}\mathbf{x}^\dagger \rangle, \tag{2.11}$$

assuming $\langle \mathbf{x} \rangle_t = 0$ (a reasonable assumption since fringes average to zero over a sufficient amount of time), where $\langle \rangle_t$ denotes a finite average over time. The weighting matrix for our empirically estimated inverse covariance weighting is then $\mathbf{R} \equiv \widehat{\mathbf{C}}^{-1}$, where we use a hat symbol to distinguish the empirical covariance from the true covariance $\mathbf{C}$.

In the next three sections, we use toy models to investigate the effects of weighting matrices on signal loss by experimenting with different matrices $\mathbf{R}$ and examining their impact on the resulting power spectrum estimates $\widehat{\mathbf{P}}$. Our goal in experimenting with weighting is to suppress foregrounds and investigate EoR losses associated with it. We note that we purposely take a thorough and pedagogical approach to describing the toy model examples given in the next few sections. The specifics of how signal loss appears in PAPER's analysis is later described in Chapter 3.

As a brief preview, we summarize our findings in the following sections here:

- If the covariance matrix is estimated from the data, a strong correlation between the estimated modes and the data will in general produce an estimate of the signal power spectrum which is strongly biased *low* relative to the true value. In this context, this is what we call "signal loss" (Chapter 2.2.2).
- The effect of the bias is worsened when the number of independent samples used to estimate the covariance matrix is reduced (Chapter 2.2.3).
- The rate at which empirical eigenvectors converge to their true forms depends on the sample variance in the empirical estimate and the shape of the empirical eigenspectrum. In general, larger sample variances lead to more loss (Chapter 2.2.3).
- Knowing these things, there are some simple ways of altering the empirical covariance matrix to decouple it from the data and produce unbiased power spectrum estimates (Chapter 2.2.4).

2.2.2 Empirical Inverse Covariance Weighting

Using a toy model, we will now build intuition into how weighting by the inverse of the empirically estimated covariance, $\widehat{\mathbf{C}}^{-1}$, can give rise to signal loss. We construct a simple data set that contains visibility data with 100 time integrations and 20 frequency channels. This model represents realistic dimensions of about an hour of PAPER data which might be used for a power spectrum analysis. For PAPER-64 (both the A15 analysis and our new

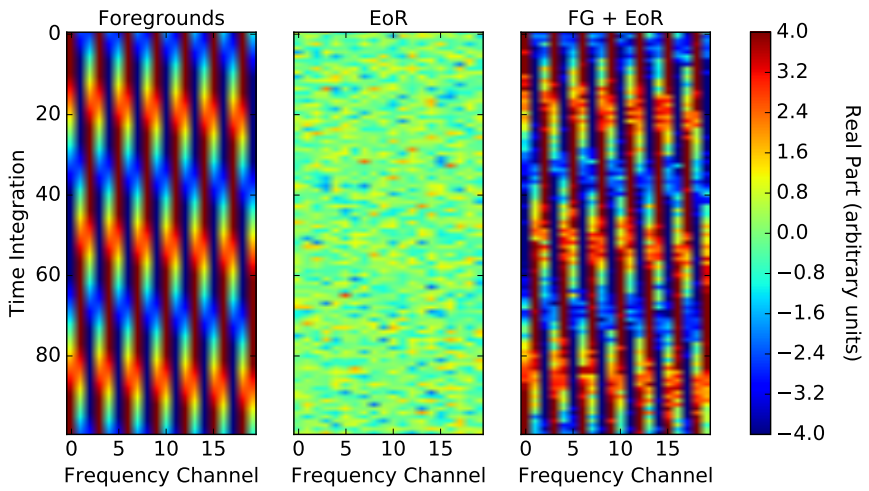

Figure 2.1: Our toy model data set to which we apply different weighting schemes to in order to investigate signal loss. We model a mock foreground-only visibility with a sinusoid signal that varies smoothly in time and frequency. We model a mock visibility of an EoR signal as a random Gaussian signal. We add the two together to form $\mathbf{x} = \mathbf{x}_{\rm FG} + \mathbf{x}_{\rm EoR}$. Real parts are shown here.

analysis) we use ~ 8 hours of data (with channel widths of 0.5 MHz and integration times of 43 seconds), but here we scale it down with no loss of generality.

We create mock visibilities, $\mathbf{x}$, and assume a non-tracking, drift-scan observation. Hence, flat spectrum sources (away from zenith) lead to measured visibilities which oscillate in time and frequency. We therefore form a mock visibility measurement of a bright foreground signal, $\mathbf{x}_{\rm FG}$, as a complex sinusoid that varies smoothly in time and frequency, a simplistic but realistic representation of a single bright source. We also create a mock visibility measurement of an EoR signal $\mathbf{x}_{\rm EoR}$ as a complex, Gaussian random signal. A more realistic EoR signal would have a sloped power spectrum in $p(k)$ (instead of flat, as in the case of white noise), which could be simulated by introducing frequency correlations into the mock EoR signal. However, here we treat all k's separately, so a simplistic white noise approximation can be used. Our combined data vector is then $\mathbf{x} = \mathbf{x}_{\rm FG} + \mathbf{x}_{\rm EoR}$, to which we apply different weighting schemes throughout Chapter 2.2. The three data components are shown in Figure 2.1.

We compute the power spectrum of our toy model data set $\mathbf{x}$ using Equations (2.8) and (2.9), with $\mathbf{R} \equiv \widehat{\mathbf{C}}^{-1}$. Figure 2.2 shows the estimated covariances of our toy model data sets along with the $\widehat{\mathbf{C}}^{-1}$ weighted data. The foreground sinusoid is clearly visible in $\widehat{\mathbf{C}}_{\rm FG}$. The power spectrum result is shown in green in the left plot of Figure 2.3. Also plotted in the figure are the uniform-weighted ($\mathbf{R} \equiv \mathbf{I}$) power spectrum of the individual components $\mathbf{x}_{\rm FG}$ (blue) and $\mathbf{x}_{\rm EoR}$ (red). As shown, our $\widehat{\mathbf{C}}^{-1}$ weighted result successfully suppresses foregrounds, demonstrated in Figure 2.3 by the missing foreground peak in the

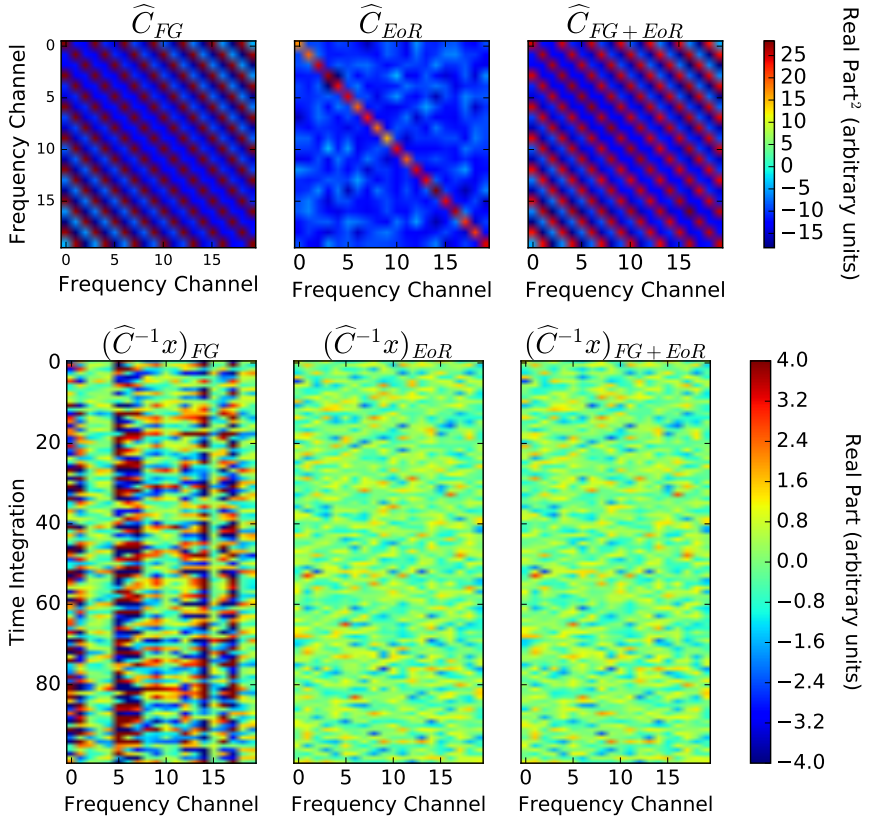

Figure 2.2: The estimated covariance matrices (top row) and inverse covariance-weighted data (bottom row) for FG only (left), EoR only (middle), and FG + EoR (right). Real parts are shown here.

weighted power spectrum estimate (green). It is also evident that our result fails to recover the EoR signal — it exhibits the correct shape, but the amplitude level is slightly low. It is this behavior which we describe as signal loss.

As discussed in Chapter 2.2.1, this behavior is *not* expected in the case that we were to use a true $\mathbf{C}^{-1}$ weighting. Rather, we would obtain the behavior shown in the toy model in Chapter 2.3.1[1], with suppression of the foreground mode resulting in a nearly unbiased estimate of the power spectrum. The key difference is that since $\widehat{\mathbf{C}}$ is estimated from the data, its eigenvectors and eigenvalues are strongly coupled to the particular data realization that was used to compute it, and this coupling leads to loss.

For the case of an eigenmode which can be safely assumed to be predominantly a foreground, its presence in the true covariance matrix will result in the desired suppression via a kind of projection; whether or not it is strongly correlated with the the actual data vector is irrelevant. However, in the case of an empirically estimated covariance matrix, the eigen-

[1]Note that there the true covariance matrix is also the sum of a diagonal portion describing the signal and a single mode describing the contaminant (similar to Figure 2.2).

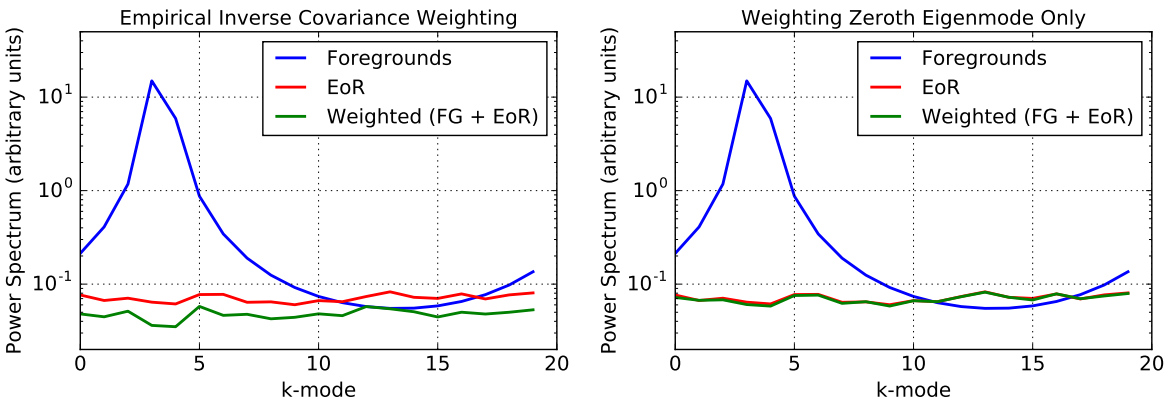

Figure 2.3: Resulting power spectrum estimates for the toy model simulation described in Chapter 2.2.2 — foregrounds only (blue), EoR only (red), and the weighted FG + EoR data set (green). The power spectrum of the foregrounds peaks at a k-mode based on the frequency of the sinusoid used to create the mock FG signal. In the two panels, we compare using empirically estimated inverse covariance weighting where $\mathbf{C}$ is derived from the data (left), and projecting out the zeroth eigenmode only (right). In the former case, signal loss arises from the coupling of the eigenmodes of $\widehat{\mathbf{C}}$ to the data. There is negligible signal loss when all eigenmodes besides the foreground one are no longer correlated with the data.

modes of $\widehat{\mathbf{C}}_{\mathrm{EoR}}$ will both be incorrect and can be correlated with the data. If these incorrect eigenmodes are not correlated with the data, it will lead to non-minimum variance estimates but will not produce the suppression of the power spectrum amplitude as seen in the left plot of Figure 2.3. As shown mathematically in Chapter 2.3.2, however, if $\widehat{\mathbf{C}}_{\mathrm{EoR}}$ *is* correlated with the data vector $\mathbf{x}$, there is a kind of projection of power in the *non*-foreground modes from the resulting power spectrum estimate, thus producing an estimate that is biased low. In short, *if the covariance is computed from the data itself, it carries the risk of overfitting information in the data and introducing a multiplicative bias (per k) to estimates of the signal.*

The danger of an empirically estimated covariance matrix comes mostly from not being able to describe the EoR-dominated eigenmodes of $\mathbf{C}$ accurately, for which the EoR signal is brighter than foregrounds. In such a case, the coupling between these modes to the data realization leads to the overfitting and subtraction of the EoR signal. More specifically, the coupling between the estimated covariance and the data is anti-correlated in nature (which is explained in more detail in Chapters 2.3.2 and 3.2.1), which leads to loss. Mis-estimating $\mathbf{C}$ for EoR-dominated eigenmodes is therefore more harmful than for FG-dominated modes, and since the lowest-valued eigenmodes of an eigenspectrum are typically EoR-dominated, using this part of the spectrum for weighting is most dangerous.

Armed with this information, we can tweak the covariance in a simple way to suppress foregrounds and yield minimal signal loss. Recall that our toy model foreground can be perfectly described by a single eigenmode. Using the full data set's (foreground plus EoR

signal) empirical covariance, we can project out the zeroth eigenmode and then take the remaining covariance to be the identity matrix. This decouples the covariance from the data for the EoR modes. The resulting power spectrum estimate for this case is shown in the right plot of Figure 2.3. In this case we recover the EoR signal, demonstrating that if we can disentangle the foreground-dominated modes and EoR-dominated modes, we can suppress foregrounds with negligible signal loss.

Altering $\widehat{\mathbf{C}}$ as such is one specific example of a regularization method for this toy model, in which we are changing $\widehat{\mathbf{C}}$ in a way that reduces its coupling to the data realization. There are several other simple ways to regularize $\widehat{\mathbf{C}}$, and we will discuss some in Chapter 2.2.4.

2.2.3 Effect of Fringe-Rate Filtering

We have shown how signal loss can arise due to the coupling of EoR-dominated eigenmodes to the data. We will next show how this effect is exacerbated by reducing the total number of independent samples in a data set.

A fringe-rate filter is an analysis technique designed to maximize sensitivity by integrating in time (Parsons et al. 2016). Rather than a traditional box-car average in time, a time domain filter can be designed to up-weight temporal modes consistent with the sidereal motion on the sky, while down-weighting modes that are noise-like.

Because fringe-rate filtering is analogous to averaging in time, it comes at the cost of reducing the total number of independent samples in the data. With fewer independent modes, it becomes more difficult for the empirical covariance to estimate the true covariance matrix of the fringe-rate filtered data. We can quantify this effect by evaluating a convergence metric $\varepsilon(\widehat{\mathbf{C}})$ for the empirical covariance, which we define as

$$\varepsilon(\widehat{\mathbf{C}}) \equiv \sqrt{\frac{\sum_{ij}(\widehat{C}_{ij} - C_{ij})^2}{\sum_{ij} C_{ij}^2}}, \tag{2.12}$$

where $\mathbf{C}$ is the true covariance matrix. To compute this metric, we draw different numbers of realizations (different draws of Gaussian noise) of our toy model EoR measurement, $\mathbf{x}_{\mathrm{EoR}}$, and take their ensemble average. We then compare this to the "true" covariance, which in our simulation is set to be the empirical covariance after a large number (500) of realizations. As shown in Figure 2.4, we perform this computation for a range of total independent ensemble realizations (horizontal axis) and number of independent samples in the data following time-averaging, or "fringe-rate filtering" (different colors). With more independent time samples (i.e., more realizations) in the data, one converges to the true fringe-rate filtered covariance more quickly.

The situation here with using a finite number of time samples to estimate our covariance is analogous to a problem faced in galaxy surveys, where the non-linear covariance of the matter power spectrum is estimated using a large — but finite — number of expensive simulations. There, the limited number of independent simulations results in inaccuracies in estimated covariance matrices (Dodelson & Schneider 2013; Taylor & Joachimi 2014),

which in turn result in biases in the final parameter constraints (Hartlap et al. 2007). In our case, the empirically estimated covariances are used for estimating the power spectrum, and as we discussed in the previous section (and will argue more thoroughly in Chapters 2.3.2 and 3.2.1), couplings between these covariances and the data can lead to power spectrum estimates that are biased *low*—which is precisely signal loss. In future work, it will be fruitful to investigate whether advanced techniques from the galaxy survey literature for estimating accurate covariance matrices can be successfully adapted for 21 cm cosmology. These techniques include the imposition of sparsity priors (Padmanabhan et al. 2016), the fitting of theoretically motivated parametric forms (Pearson & Samushia 2016), covariance tapering (Paz & Sánchez 2015), marginalization over the true covariance (Sellentin & Heavens 2016), and shrinkage methods (Pope & Szapudi 2008; Joachimi 2017).

The overall convergence of the covariance is important, but also noteworthy is the fact that different eigenvectors converge to their true forms at different rates. This is illustrated by Figure 2.5, which shows the convergence of eigenvectors in an empirical estimate of a covariance matrix. For this particular toy model, we construct a covariance whose true form combines the same mock foreground from the previous toy models with an EoR component that is modeled as a diagonal matrix with eigenvalues spanning one order of magnitude (more specifically, we construct the EoR covariance as a diagonal matrix in the Fourier domain, where the signal is expected to be uncorrelated; its Fourier transform is then the true covariance of the EoR in the frequency domain, or $\mathbf{C}_{\rm EoR}$). For different numbers of realizations, we draw random EoR signals that are consistent with $\mathbf{C}_{\rm EoR}$, add them to the mock foreground data, and compute the combined empirical covariance by averaging over the realizations. The eigenvectors of this empirical covariance are then compared to the true eigenvectors $\mathbf{v}$, where we use as a convergence metric $\varepsilon(\widehat{\mathbf{v}})$, defined as:

$$\varepsilon(\widehat{\mathbf{v}}) \equiv \sqrt{\sum_{i}^{N_f} |\mathbf{v} - \widehat{\mathbf{v}}|_i^2}, \tag{2.13}$$

where N_f is the number of frequencies (20) in the mock data. The eigenmode convergence curves in Figure 2.5 are ranked ordered by eigenvalue, such that "Eigenmode #0" illustrates the convergence of the eigenvector with the largest eigenvalue, "Eigenmode #1" for the second largest eigenvalue, and so on. We see that the zeroth eigenmode — the mode describing the foreground signal — is quickest to converge.

Our numerical test reveals that the convergence rates of empirical eigenvectors is related to the sample variance in our empirical estimate. In general, computing an empirical covariance from a finite ensemble average means that the empirical eigenmodes have sample variances. Consider first a limiting case where all eigenvalues are equal. In such a scenario, any linear combination of eigenvectors is also an eigenvector, and thus there is no sensible way to define the convergence of eigenvectors. In our current test, aside from the zeroth mode, the eigenvalues have similar values but are not precisely equal. Hence, there is a well-defined set of eigenvectors to converge to. However, due to the sample variance of our empirical covariance estimate, there may be accidental degeneracies between modes, where

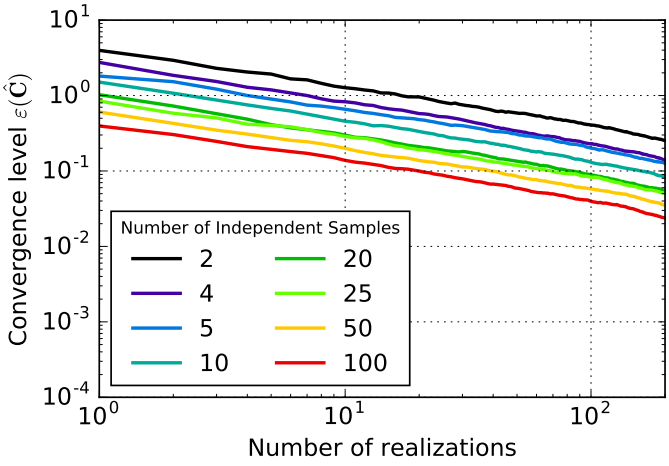

Figure 2.4: The convergence level, as defined by Equation (2.12), of empirically estimated covariances of mock EoR signals with different numbers of independent samples. In red, the mock EoR signal is comprised entirely of independent samples (100 of them). Subsequent colors show time-averaged signals. As the number of realizations increases, we see that the empirical covariances approach the true covariances. With more independent samples, the quicker an empirical covariance converges (i.e., the quicker it decouples from the data), and the less signal loss we would expect to result.

some modes are mixing and swapping with others. Therefore, the steeper an eigenspectrum, the easier it is for the eigenmodes to decouple from each other and approach their true forms. A particularly drastic example of this can be seen in the behavior of mode 0 (the foreground mode), whose eigenvalue differs enough from the others that it is able to converge reasonably quickly despite substantial sample variance in our empirical covariance estimate. To break degeneracies in the remaining modes, however, requires many more realizations.

While the connection between the rate of convergence of an empirical eigenvector with the sample variance of an eigenspectrum is interesting, it is also important to note that regardless of convergence rate, any mode that is coupled to the data is susceptible to signal loss. The true eigenvectors are not correlated with the data realizations; thus, if our empirical eigenvectors are converged fully, there will not be any signal loss. However, an unconverged eigenvector estimate will retain some memory of the data realizations used in its generation, leading to signal loss.

In the toy models throughout Chapter 2.2, we exploit the fact that the strongest eigenmode (highest eigenvalue mode) is dominated by foregrounds in order to purposely incur signal loss for that mode. Even for the case of real PAPER data (Chapter 3), we make the assumption that the strongest eigenmodes are likely the most contaminated by foregrounds. However, in general, foregrounds need not be restricted to the strongest eigenmodes, and as we have seen, it is really the degeneracies between modes that determines how quickly they converge, and hence how much signal loss can result.

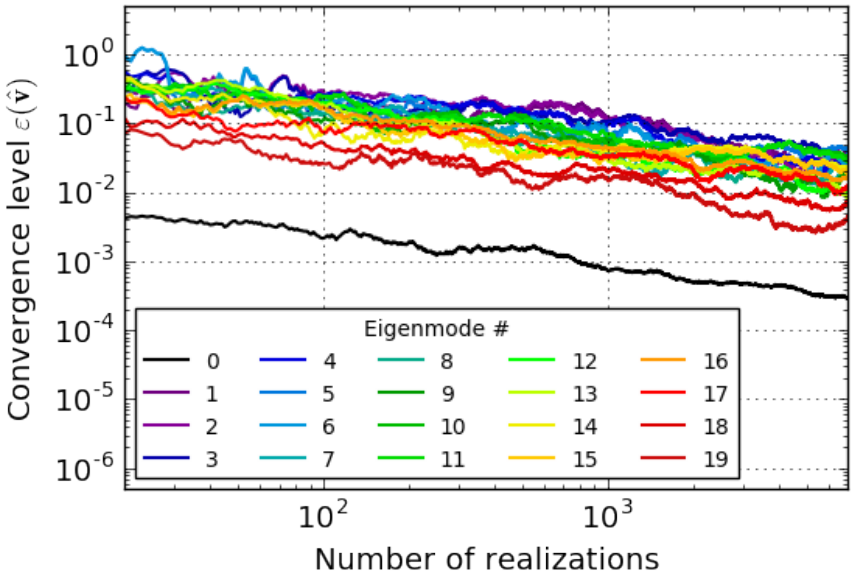

Figure 2.5: The convergence level, as defined by Equation (2.13), of empirically estimated eigenvectors for different numbers of mock data realizations. The colors span from the 0th eigenmode (has the highest eigenvalue) to the 19th eigenmode (has the lowest eigenvalue), where they are ordered by eigenvalue in descending order. This figure shows that the zeroth eigenmode converges the quickest, implying that eigenvectors with eigenvalues that are substantially different than the rest (the FG-dominated mode has a much higher eigenvalue than the EoR modes) are able to converge to the true eigenvectors the quickest. On the other hand, eigenmodes 1-19 have similar eigenvalues and are slower to converge because of degeneracies between them.

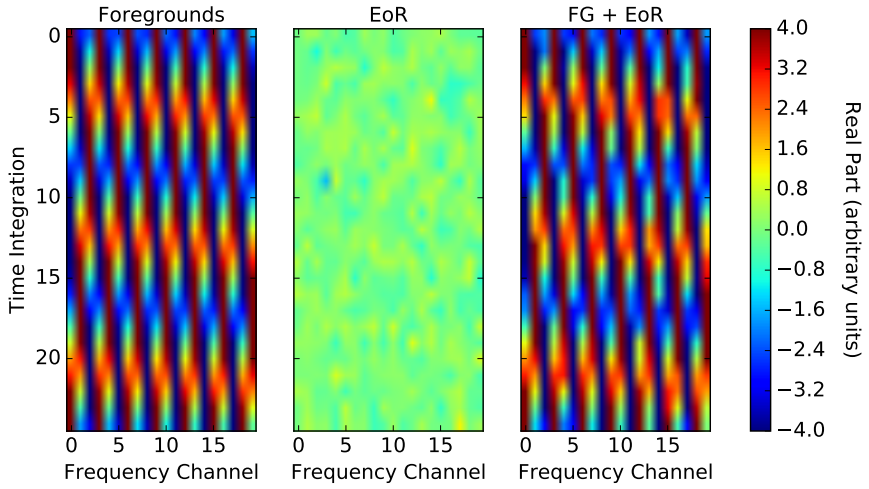

Figure 2.6: Our "fringe-rate filtered" (time-averaged) toy model data set. We average every four samples together, yielding 25 independent samples in time. Real parts are shown here.

With Figures 2.4 and 2.5 establishing the connection between convergence rates (of empirical covariances and eigenvectors) and number of realizations, we now turn back to our original toy model used in Chapter 2.2.2, which is comprised of a mock foreground and mock EoR signal. We mimic a fringe-rate filter by averaging every four time integrations of our toy model data set together, yielding 25 independent samples in time (Figure 2.6). We choose these numbers so that the total number of independent samples is similar to the number of frequency channels — hence our matrices will be full rank. We use this "fringe-rate filtered" mock data for the remainder of Chapter 2.2.

The power spectrum results for this model are shown in Figure 2.7, and as expected there is a much larger amount of signal loss for this time-averaged data set since we do a worse job estimating the true covariance. In addition, as a result of having fewer independent samples, we obtain an estimate with more scatter. This is evident by noticing that the green curve in Figure 2.7 fails to trace the shape of the uniform-weighted EoR power spectrum (red).

Using our toy model, we have seen that a sensitivity-driven analysis technique like fringe-rate filtering has trade-offs of signal loss and noisier estimates when using data-estimated covariance matrices. Longer integrations increase sensitivity but reduce the number of independent samples, resulting in eigenmodes correlated with the data that can overfit signal greatly. We note that a fringe-rate filter does have a range of benefits, many described in Parsons et al. (2016), so it can still be advantageous to use one despite the trade-offs.

2.2.4 Other Weighting Options

In Chapter 2.2.2 we showed one example of how altering $\widehat{\mathbf{C}}$ can make the difference between nearly zero and some signal loss. We will now use our toy model to describe several other ways to tailor $\widehat{\mathbf{C}}$ in order to minimize signal loss. We choose four independent regularization methods to highlight in this section, which have been chosen due to their simplicity in implementation and straightforward interpretations. We illustrate the resulting power spectra for the different cases in Figure 2.8. These examples are not meant to be taken as suggested analysis methods but rather as illustrative cases.

As a first test, we model the covariance matrix of EoR as a proof of concept that if perfect models are known, signal loss can be avoided. We know that our simulated EoR signal should have a covariance matrix that mimics the identity matrix, with its variance encoded along the diagonal. We model $\mathbf{C}_{\mathrm{EoR}}$ as such (i.e., the identity), instead of computing it based on $\mathbf{x}_{\mathrm{EoR}}$ itself. Next, we add $\mathbf{C}_{\mathrm{EoR}} + \widehat{\mathbf{C}}_{\mathrm{FG}}$ (where $\widehat{\mathbf{C}}_{\mathrm{FG}} = \langle \mathbf{x}_{\mathrm{FG}} \mathbf{x}_{\mathrm{FG}}^{\dagger} \rangle_t$) to obtain a final $\widehat{\mathbf{C}}_{\mathrm{reg}}$ (regularized empirical covariance matrix) to use in weighting. In Figure 2.8 (upper left), we see that there is negligible signal loss. This is because by modeling $\mathbf{C}_{\mathrm{EoR}}$, we avoid overfitting EoR fluctuations in the data that our model doesn't know about (but, an empirically derived $\widehat{\mathbf{C}}_{\mathrm{EoR}}$ would know about the fluctuations). In practice such a weighting option is not feasible, as it is difficult to model $\mathbf{C}_{\mathrm{EoR}}$, and $\widehat{\mathbf{C}}_{\mathrm{FG}}$ is unknown because we do not know how to separate out the foregrounds from the EoR in our data.

The second panel (top right) in Figure 2.8 uses a regularization method of setting $\widehat{\mathbf{C}}_{\mathrm{reg}} \equiv$

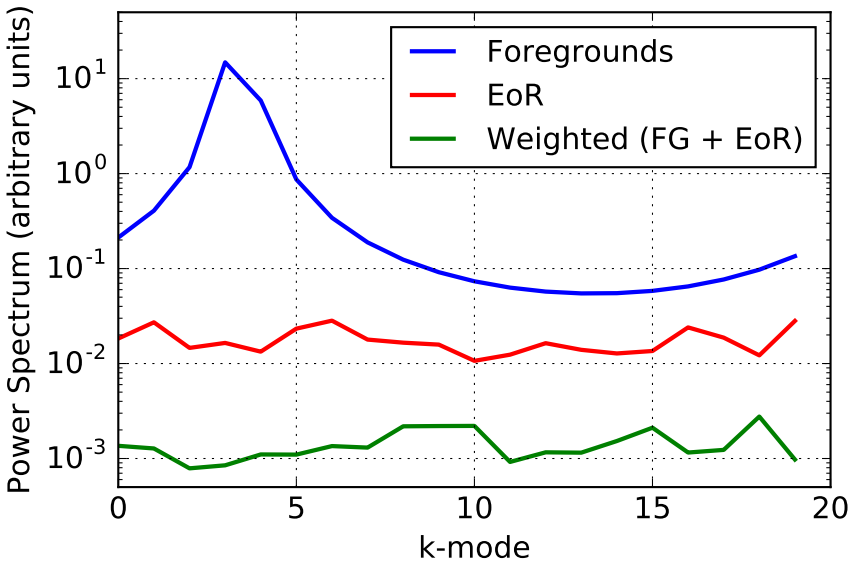

Figure 2.7: Resulting power spectrum estimate for the "fringe-rate filtered" (time-averaged) toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR data set (green). We use empirically estimated inverse covariance weighting where $\mathbf{C}$ is computed from the data. There is a larger amount of signal loss than for the non-averaged data, a consequence of weighting by eigenmodes that are more strongly coupled to the data due to there being fewer independent modes in the data.

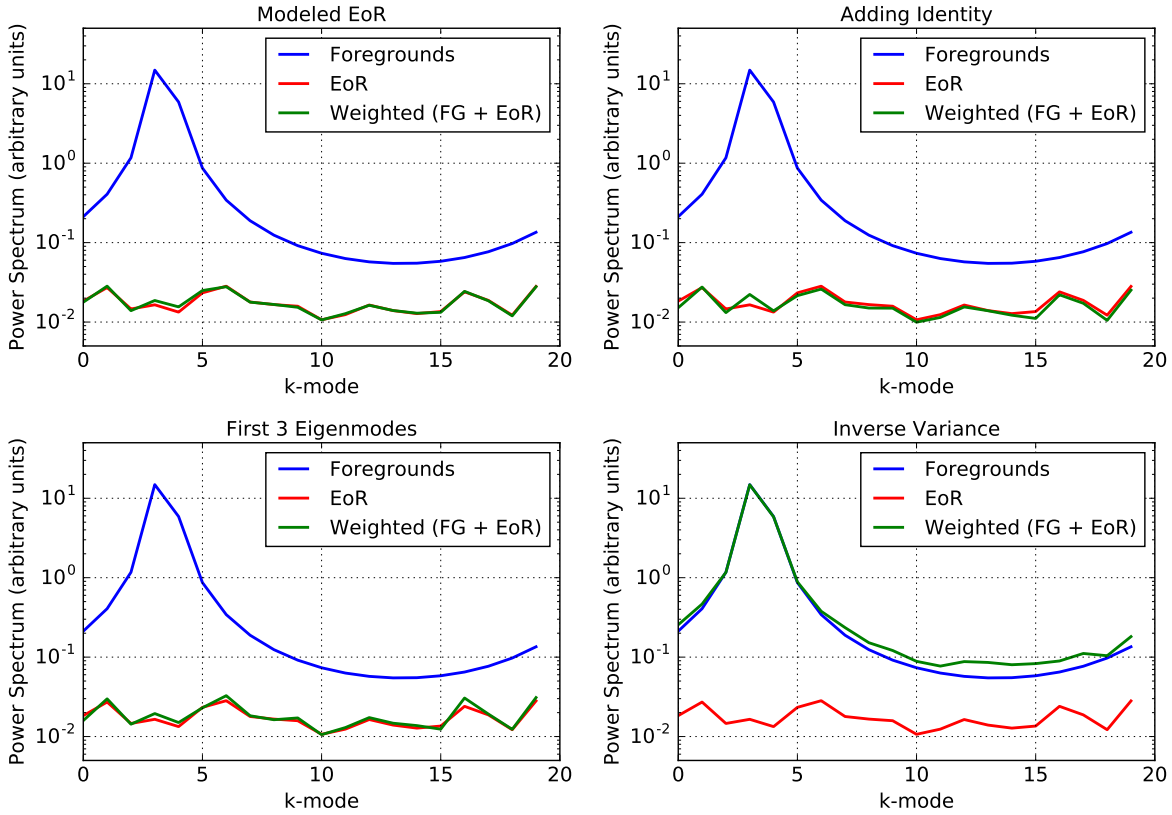

Figure 2.8: Resulting power spectra estimates for our "fringe-rate filtered" (time-averaged) toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR data set (green). We show four alternate weighting options that each minimize signal loss, including modeling the covariance matrix of EoR (upper left), regularizing $\widehat{\mathbf{C}}$ by adding an identity matrix to it (upper right), using only the first three eigenmodes of $\widehat{\mathbf{C}}$ (lower left), and keeping only the diagonal elements of $\widehat{\mathbf{C}}$ (lower right). The first case (upper left) is not feasible in practice since we do not know $\mathbf{C}_{\mathrm{FG}}$ and $\mathbf{C}_{\mathrm{EoR}}$ like we do in the toy model.

$\widehat{\mathbf{C}}+\gamma\mathbf{I}$, where $\gamma = 5$ (an arbitrary strength of $\mathbf{I}$ for the purpose of this toy model). By adding the identity matrix, element-wise, we are weighting the diagonal elements of the estimated covariance matrix more heavily than those off-diagonal. Since the identity component does not know anything about the data realization, it alters the covariance to be less coupled to the data and there is no loss.

The third panel (bottom left) in Figure 2.8 minimizes signal loss by only using the first three eigenmodes of the estimated covariance. Recalling that our toy model foregrounds can be described entirely by the zeroth eigenmode, this method intentionally projects out the highest-valued modes only by replacing all but the three highest weights in the eigenspectrum with 1's (equal weights). Again, avoiding the overfitting of EoR-dominated modes which are coupled to the data results in negligible signal loss. While this case is illuminating for the toy model, in practice it is not obvious which eigenmodes are foreground or EoR dominated (and they could be mixed as well), so determining which subset of modes to down-weight is not trivial. We experiment with this idea using PAPER data in Chapter 3.2.3.

The last regularization scheme we are highlighting here is setting $\widehat{\mathbf{C}}_{\text{reg}} \equiv \widehat{\mathbf{C}} \circ \mathbf{I}$ (element-wise multiplication), or inverse variance weighting (i.e., keeping only the diagonal elements of $\widehat{\mathbf{C}}$). In the bottom right panel of Figure 2.8, we see that this method does not down-weight the foregrounds at all — this regularization altered $\widehat{\mathbf{C}}$ in a way where it is no longer coupled to *any* of the empirically estimated eigenmodes, including the FG-dominated one. To understand this, we recall that our foregrounds are spread out in frequency and therefore have non-negligible frequency-frequency correlations. Multiplying by the identity matrix, element-wise, results in a diagonal matrix, meaning we do not have any correlation information. Because of this, we do a poor job suppressing the foreground. But because we decoupled the whole eigenspectrum from the data, we also avoid signal loss. Although this method did not successfully recover the EoR signal for this particular simulation, it is important that we show that there are many options for estimating a covariance matrix, and some may down-weight certain eigenmodes more effectively than others based on the spectral nature of the components in a data set.

In summary, we have shown how signal loss is caused by weighting a data set by itself, and in particular how estimated covariances can overfit EoR modes when they are coupled to data and not converged to their true forms. We have also seen that there are trade-offs between a chosen weighting method, its foreground-removal effectiveness, the number of independent samples in a data set, and the amount of resulting signal loss.

2.3 Signal Loss Mathematical Framework

In Chapter 2.2.1, we argued that the optimal quadratic estimator method has the effect of projecting out foreground modes that have a different covariance structure than the EoR signal. And in Chapter 2.2.2, we saw that *empirical* inverse covariance weighting can lead to loss via a multiplicative bias to estimates of the signal. To support these arguments, we now present two mathematical frameworks in which we use toy models in order to illustrate

these conclusions analytically.

2.3.1 A Toy Model for Inverse Covariance Weighting

In this section, we focus on the optimal quadratic estimator, Equation (2.3), and mathematically illustrate its role in estimating the power spectrum of EoR. While there exists detailed literature about quadratic estimators in general (e.g., Liu & Tegmark 2011; Trott et al. 2012; Liu et al. 2014b; Dillon et al. 2014), here we focus on two simple cases in order to outline one situation where the estimator successfully suppresses contamination and one where it does not. By describing these two cases, we hope to clarify and motivate the desire to use OQE while also understanding its limitations.

In our toy model, we specifically choose models where the data covariance is diagonal, as indeed we expect the EoR signal to be. We assume we have N data points Δ_i which are the sum of a desired signal σ_i and an undesired contaminant υ_i

$$\Delta_i = \sigma_i + \upsilon_i \tag{2.14}$$

with

$$\langle \sigma_i \rangle = 0;\ \langle \sigma_i^2 \rangle = s;\ \text{and}\ \langle \boldsymbol{\sigma}\boldsymbol{\sigma}^T \rangle = s\mathbf{I}_{N\times N} \equiv \mathbf{S}, \tag{2.15}$$

where we wish to estimate s. The contaminant in this first case has a similar structure (as the EoR) for its covariance, and is assumed uncorrelated with the signal

$$\langle \upsilon_i \rangle = 0;\ \langle \upsilon_i^2 \rangle = u;\ \langle \boldsymbol{\upsilon}\boldsymbol{\upsilon}^T \rangle = u\mathbf{I}_{N\times N} \equiv \mathbf{U}; \text{and}\ \langle \sigma_i \upsilon_j \rangle = 0. \tag{2.16}$$

With the covariance matrix given by $\mathbf{C} = \mathbf{S} + \mathbf{U}$, the estimator for s using only the quadratic part of Equation 2.3 is

$$\hat{s} = \frac{\boldsymbol{\Delta}^T\boldsymbol{\Delta}}{N} \tag{2.17}$$

and its expectation is

$$\langle \hat{s} \rangle = s + u. \tag{2.18}$$

Thus, *when the covariance structure of the contaminant is identical to the signal* ($\frac{\partial \mathbf{S}}{\partial s} = \frac{\partial \mathbf{U}}{\partial u} = \frac{\partial \mathbf{C}}{\partial s}$), the information available to the quadratic portion of the estimator to distinguish between the two is degenerate, and knowledge only of $\mathbf{C}$ and $\frac{\partial \mathbf{C}}{\partial s}$ is inadequate. In order to obtain an unbiased estimate of s, one must also use knowledge of $\mathbf{U}$. Indeed, computing the linear bias from Equation (2.6), one finds $b = u$.

Now consider a case, chosen to be very similar to the toy model in 2.2.2, in which the data again have an additive contaminant, now given by

$$\Delta_i = \sigma_i + \upsilon m_i \tag{2.19}$$

where the properties of σ_i are as before, but now υ is a random variable and m_i is a fixed function of i with

$$\langle \upsilon \rangle = 0;\ \langle \upsilon^2 \rangle = u;\ \langle \boldsymbol{\upsilon}\boldsymbol{\upsilon}^T \rangle = u\mathbf{m}\mathbf{m}^T \equiv \mathbf{U};\ \mathbf{m}^T\mathbf{m} = 1;\ \text{and}\ \langle \sigma_i \upsilon \rangle = 0. \tag{2.20}$$

Here $\mathbf{m}$ represents a mode which is correlated across many data points (i.e., we are assuming $\mathbf{U}$ need *not* be diagonal), with amplitude given by v. The normalization of $\mathbf{m}$ is a matter of convention, and can be absorbed in the variance u; the choice above will be convenient for understanding the limiting case $u \gg s$.

We can calculate the quadratic portion of the estimator explicitly by using the Sherman-Morrison identity to invert the covariance matrix. Defining

$$\xi \equiv \frac{u/s}{1+u/s}, \tag{2.21}$$

we have

$$\mathbf{C}^{-1} = \frac{1}{s}\left(\mathbf{I} - \xi \mathbf{m}\mathbf{m}^T\right) \tag{2.22}$$

and

$$\hat{s} = \frac{\mathbf{\Delta}^T(\mathbf{I} + (\xi^2 - 2\xi)\mathbf{m}\mathbf{m}^T)\mathbf{\Delta}}{N + \xi^2 - 2\xi} \tag{2.23}$$

with expectation

$$\langle \hat{s} \rangle = s + \frac{1 - 2\xi + \xi^2}{N + -2\xi + \xi^2} u. \tag{2.24}$$

It is worth observing immediately that there is no multiplicative bias on s, and that the additive bias is strictly $< u/N$.

An instructive limit is $u \gg s$, $\xi \to 1$, in which case the virtue of weighting by $\mathbf{C}^{-1}$ becomes clearer, as it becomes

$$\mathbf{C}^{-1} = \frac{1}{s}\left(\mathbf{I} - \mathbf{m}\mathbf{m}^T\right) \tag{2.25}$$

where $\mathbf{I} - \mathbf{m}\mathbf{m}^T$ is the projection operator, projecting out $\mathbf{m}$ from any vector it acts on, and further, the linear bias tends to 0 as $\xi \to 1$ (i.e., the projection is "perfect" and not "undone" by the Fisher matrix normalization).

This is the ideal case for the inverse covariance weighting performed in the PAPER analysis, where removal of contamination with a known covariance can be suppressed by a kind of projection of the offending modes. But even in this case, it is worth pointing out that the estimator still has a linear bias for finite u. We have also assumed that the contaminating mode $\mathbf{m}$ is known perfectly; the next section takes up the case where the modes are estimated from the data.

2.3.2 A Toy Model for Signal Loss

In this section, we examine a toy model for signal loss. Our goal is to derive an analytic formula for power spectrum signal loss. While this model does not apply generally to all the scenarios presented in this work, it provides some analytic intuition for how the coupling between data and an empirical covariance can result in signal loss.

The minimum-variance quadratic estimator $\widehat{P}^\alpha$ for the αth bandpower of the power spectrum is given by

$$\widehat{P}^\alpha = \frac{1}{2\mathbf{F}^{\alpha\alpha}}\mathbf{x}^t\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{x}, \tag{2.26}$$

where

$$F^{\alpha\alpha} \equiv \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\alpha\right) \tag{2.27}$$

is the αth diagonal element of the Fisher matrix. For this section only, with no loss of generality, we assume that the data $\mathbf{x}$ are real. We also assume for simplicity that $\mathbf{x}$ is the data from a single instant in time, so that it is of length N_f, where N_f is the number of frequency channels.

In our case, we do not have *a priori* knowledge of the covariance matrix. Thus, we deviate from the true minimum-variance quadratic estimator and replace $\mathbf{C}$ with $\widehat{\mathbf{C}}$, its data-derived approximation. Our estimator then becomes

$$\widehat{P}^\alpha_{\mathrm{loss}} = \frac{1}{2\widehat{\mathbf{F}}^{\alpha\alpha}}\mathbf{x}^t\widehat{\mathbf{C}}^{-1}\mathbf{Q}^\alpha\widehat{\mathbf{C}}^{-1}\mathbf{x}, \tag{2.28}$$

where

$$\widehat{F}^{\alpha\alpha} \equiv \frac{1}{2}\mathrm{tr}\left(\widehat{\mathbf{C}}^{-1}\mathbf{Q}^\alpha\widehat{\mathbf{C}}^{-1}\mathbf{Q}^\alpha\right), \tag{2.29}$$

with the label "loss" to foreshadow the fact that this will be an estimator with signal loss (i.e., a multiplicative bias of less than unity). We will now provide an explicit demonstration of this by modeling the estimated covariance as

$$\widehat{\mathbf{C}} = (1-\eta)\mathbf{C} + \eta\mathbf{x}\mathbf{x}^t, \tag{2.30}$$

where η is a parameter quantifying our success at estimating the true covariance matrix. If $\eta = 0$, our covariance estimate has perfectly modeled the true covariance and $\widehat{\mathbf{C}} = \mathbf{C}$. On the other hand, if $\eta = 1$, then our covariance estimate is based purely on the one realization of the covariance that is our actual data, and we would expect a high level of overfitting and signal loss.

Our strategy for computing the signal loss will be to insert Equation (2.30) into Equation (2.28) and to express the resulting estimator $\widehat{P}^\alpha_{\mathrm{loss}}$ in terms of $\widehat{P}^\alpha$. We begin by expressing $\widehat{\mathbf{C}}^{-1}$ in terms of $\mathbf{C}^{-1}$ using the Woodbury identity so that

$$\widehat{\mathbf{C}}^{-1} = \frac{\mathbf{C}^{-1}}{1-\eta}\left[\mathbf{I} - \frac{\eta\mathbf{x}\mathbf{x}^t\mathbf{C}^{-1}}{1+\eta(g-1)}\right], \tag{2.31}$$

where we have defined $g \equiv \mathbf{x}^t\mathbf{C}^{-1}\mathbf{x}$. Inserting this into our Fisher estimate we have

$$\widehat{F}^{\alpha\alpha} = \frac{F^{\alpha\alpha}}{(1-\eta)^2}\left[1 - \frac{\eta}{1+\eta(g-1)}\frac{h^{\alpha\alpha}}{F^{\alpha\alpha}} + \frac{1}{2}\left(\frac{\eta}{1+\eta(g-1)}\right)^2\frac{(h^\alpha)^2}{F^{\alpha\alpha}}\right], \tag{2.32}$$

where $h^\alpha \equiv \mathbf{x}^t\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{x}$ and $h^{\alpha\alpha} \equiv \mathbf{x}^t\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{x}$. Note that g, h^α, and $h^{\alpha\alpha}$ are all random variables, since they depend on $\mathbf{x}$. Inserting these expressions into our estimator gives

$$\widehat{P}^\alpha_{\rm loss} = \frac{1}{2}\frac{h^\alpha}{F^{\alpha\alpha}}\left[1 - \frac{\eta g}{1+\eta(g-1)}\right]^2\left[1 - \frac{\eta}{1+\eta(g-1)}\frac{h^{\alpha\alpha}}{F^{\alpha\alpha}} + \frac{1}{2}\left(\frac{\eta}{1+\eta(g-1)}\right)^2\frac{(h^\alpha)^2}{F^{\alpha\alpha}}\right]^{-1}. \tag{2.33}$$

Both for the purposes of analytical tractability and to provide intuition, we expand this expression to leading order in η. This approximates the limiting case where the covariance $\widehat{\mathbf{C}}$ is close to the ideal and the lossy covariance is a small perturbation. The result is

$$\widehat{P}^\alpha_{\rm loss} \approx \frac{1}{2}\frac{h^\alpha}{F^{\alpha\alpha}}\left[1 - \eta\left(g - \frac{h^{\alpha\alpha}}{F^{\alpha\alpha}}\right)\right]. \tag{2.34}$$

Taking the ensemble average of both sides and noting that the true power spectrum P^α is equal to $\langle h^\alpha\rangle/2F^{\alpha\alpha}$, we obtain

$$\langle\widehat{P}^\alpha_{\rm loss}\rangle \approx (1-\eta N_f)P^\alpha + 4\eta\frac{\mathrm{tr}(\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\alpha)}{[\mathrm{tr}(\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{Q}^\alpha)]^2} \approx (1-\eta N_f)P^\alpha, \tag{2.35}$$

where recall that N_f is the length of $\mathbf{x}$, or the number of frequency channels. In the last step we dropped the final term, since it scales as ηP^α (without the factor of N) and is therefore typically small compared to the terms that have been retained.

Recalling that P^α is the *true* power spectrum, one sees that when the covariance in the optimal quadratic estimator is naively replaced by an empirical covariance, the resulting power spectrum estimate is biased low, i.e., there is signal loss. This occurs because of couplings between $\widehat{\mathbf{C}}$ and $\mathbf{x}$, which formally means that what was originally a quadratic estimator is no longer quadratic, but contains higher-order correlations. This violates the assumptions implicit in the derivation of $F^{\alpha\alpha}$ as the normalization factor for converting unnormalized bandpowers $\frac{1}{2}\mathbf{x}^t\mathbf{C}^{-1}\mathbf{Q}^\alpha\mathbf{C}^{-1}\mathbf{x}$ into properly normalized power spectrum estimates, where the unnormalized bandpowers are assumed to be two-point (i.e., quadratic) statistics (Liu & Tegmark 2011). The result is an improperly normalized—and thus lossy—power spectrum estimate.

2.4 Error Estimation Toy Model

Our second major 21 cm power spectrum theme is error estimation, as we desire robust methods for determining accurate confidence intervals for our measurements. Two popular ways of calculating errors on a power spectrum measurement are calculating the variance of power spectrum results, and computing a theoretical error estimate based on an instrument's system temperature and observational parameters. In a perfect world, both methods would match up. However, in practice the two do not always agree due to a number of factors,

including possible non-Gaussianities in the noise properties of our instruments and possible systematics in the data.

A third option which acts as a middle ground between purely theoretical and purely empirical errors is using Gaussian error. This involves the assumption of Gaussianity, but allows the variance of the power spectrum estimator to be written as a function of the two-point estimator, or covariance. One could empirically calculate the covariance and then propagate it into an analytic expression to compute the errors, making this method fall somewhere between being fully empirical and fully modeled (see Das et al. (2011b) for an example of its implementation).

For PAPER's analysis, we choose a data-driven method of error estimation that does not rely on assumptions of Gaussianity. Namely, we compute error bars that have been derived from the inherent variance of our measurements. A common technique used to do this is bootstrapping. For pedagogical purposes, we first define the technique of bootstrapping and then illustrate one of its pitfalls through a toy model.

Bootstrapping uses sampling with replacement to estimate a posterior distribution. For example, measurements (like power spectra) can be made from different samples of data. Each of these measurements is a different realization drawn from some underlying distribution, and realizations are correlated with each other to a degree set by the fraction of sampled points that are held in common between them. Through the process of re-sampling and averaging along different axes of a data set, such as along baselines or times, we can estimate error bars for our results which represent the underlying distribution of values that are allowed by our measurements (Efron & Tibshirani 1993; Andrae 2010).

Suppose we have N different measurements targeting the same quantity (for example, N power spectrum measurements). Bootstrapping means that we form $N_{\rm boot}$ (often a large number) bootstraps, where each bootstrap is a random selection of the N measurements. Bootstraps each have dimensions of N, and the values populated into each bootstrap are drawn from the original set of measurements with replacement (i.e., every n^{th} slot in N is filled randomly for each bootstrap). Next we take the mean of each bootstrap to collapse it from an array of length N to a single number (we are interested in the mean statistic here, but any function of interest can be applied to each bootstrap as long as it's the same function for each one). The error (on the mean) is then computed as the standard deviation across all bootstraps.

We must be careful in distinguishing $N_{\rm boot}$, the number of bootstraps, from N, the number of samples, or elements, or values, that comprise a bootstrap. In the toy models presented in this section, $N_{\rm boot}$ is typically large, and the standard deviation across bootstraps (the error we are computing) converges for large $N_{\rm boot}$. Typically N is a straightforward value to set that just depends on the experiment. However, we will illustrate one case in which it is not simply the number of samples along the axis that is being re-sampled. More specifically, we will see that N depends on sample independence and may not always be straightforward to approximate.

For our toy model, suppose we have a Gaussian random signal data set of length $N = 1000$ and unity variance (zero mean). This could represent 1000 power spectrum measurements,

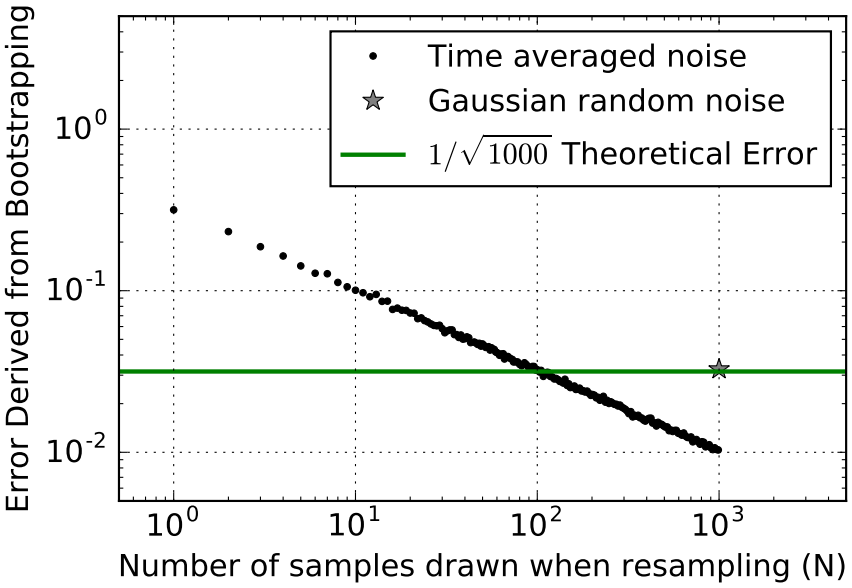

Figure 2.9: Error estimation from bootstrapping as a function of the number of elements drawn per bootstrap when sampling with replacement. The star represents the standard deviation of $N_{\text{boot}} = 500$ bootstraps, each created by drawing 1000 elements (with replacement) from a length 1000 array of a Gaussian random signal. The black points correspond to time-averaged data (correlated data) which has 100 independent samples. They illustrate how errors can be underestimated if drawing more elements than there are independent samples in the data. The estimated errors match up with the theoretical prediction only at $N = 100$.

for which we are interested in its error. We predict that the error on the mean should obey $1/\sqrt{N}$, where N is the number of samples.

We next form 500 bootstraps ($N_{\text{boot}} = 500$). To create each bootstrap, we draw N samples, with replacement, of the original data, and take the mean over the N samples. The standard deviation over the 500 bootstraps gives an error estimate for our data set. This error is indicated by the gray star in Figure 2.9 and matches our theoretical prediction (green).

One major caveat of bootstrapping arises when working with correlated data. If, for example, a data set has many repeated values inside it, this would be reflected in each bootstrap. The same value would be present multiple times within a bootstrap and also be present between bootstraps, purely because it has a more likely chance of being drawn if there are repeats of itself. Therefore, bootstrapping correlated data results in a smaller variation between bootstraps, and hence, underestimates errors. The use of a fringe-rate filter, which averages data in time to increase sensitivity, is one example which leads to a reduction in the number of independent samples, creating a situation in which errors can be underestimated. We will now show this effect using our toy model.

Going back to our toy model, we apply a sliding boxcar average to 10 samples at a time, thus reducing the number of independent data samples to $N/10 = 100$. Bootstrapping this time-averaged noise, using the same method as described earlier (drawing $N = 1000$ elements per bootstrap sample), underestimates the error by a factor of ~ 3 (black points in Figure 2.9, at $N = 1000$). This occurs because we are drawing more samples than independent ones available, and thus some samples are repeated multiple times in all bootstraps, leading to less variation between the bootstraps. In fact, the error derived from bootstrapping is a strong function of the number of elements that are drawn per bootstrap (Figure 2.9, black points), and we can both underestimate the error by drawing too many or overestimate it by drawing too few. However, since we know that we have 100 independent samples in this toy model, the error associated with drawing $N = 100$ samples with replacement does match the theoretical prediction as expected (the black points cross the green line at $N = 100$ in Figure 2.9).

This example highlights the importance of understanding how analysis techniques (such as fringe-rate filtering) can affect a common statistical procedure like bootstrapping. Bootstrapping as a means of estimating power spectrum errors from real fringe-rate filtered data requires knowledge of the number of independent samples, which is not always a trivial task. For example, computing the effective number of independent samples of fringe-rate filtered data is not as simple as counting the number of averages performed. Down-sampling a time-averaged signal is straightforward using a boxcar average, but non-trivial with a more complicated convolution function that has long tails. Hence, we do not recommend bootstrapping unless the number of independent samples along the axis that is being re-sampled is well-determined. In Chapter 3.3.1, we explain how we underestimated errors in the A15 analysis of PAPER and how our bootstrapping procedure has now changed to avoid the over-sampling of correlated data.

In summary, bootstrapping can be an effective and straightforward way to estimate errors of a data set. However, we have illustrated a situation in which bootstrapping can lead to underestimated errors and therefore underestimated power spectrum limits. We have shown that bootstrapped error depends strongly on the number of elements drawn in a bootstrap sample. Estimated errors can drop to arbitrarily small values when the number of elements drawn exceeds the effective number of independent elements. While bootstrapping is convenient because it provides a way to estimate errors from the data itself, one must assess whether certain analysis choices have compromised the method and whether a variation or an avoidance of traditional re-sampling could be preferred instead.

2.5 Bias Toy Model

In a 21 cm power spectrum, detections could be the EoR signal, but they could also be attributed to other sources of bias. Connecting a detection to EoR as opposed to noise or foreground bias is a key challenge of future 21 cm data analyses (e.g. Petrovic & Oh 2011). In this section we will discuss possible sources of bias in a measurement, as well as techniques

that can help mitigate their effects. We will also present a series of tests in a pedagogical fashion which we suggest be used to help evaluate deep limits and/or detections.

2.5.1 Foreground and Noise Bias

In Chapter 2.2, we discussed signal loss as a form of multiplicative bias to estimates of the signal. Foregrounds are another type of bias, but an additive instead of a multiplicative one. Foreground bias is perhaps one of the main factors limiting 21 cm results, as foreground signals lie $\sim$ 4-5 orders of magnitude above the cosmological signal. Though there are many techniques proposed for removing foregrounds (see e.g., Vedantham et al. 2012; Chapman et al. 2012; Parsons et al. 2012; Parsons et al. 2012; Dillon et al. 2013; Wang et al. 2013; Parsons et al. 2014; Liu et al. 2014a; Wolz et al. 2014; Liu et al. 2014b; Dillon et al. 2015a; Pober et al. 2016b; Trott et al. 2016), most experiments currently remain limited by residuals rather than noise, especially at low k.

One common method to isolate and filter foregrounds is to exploit their behavior in k-space. For a particular baseline length, there is a maximum delay imposed on sources attached to the sky, which corresponds to the light-crossing time between two antennas in a baseline. For longer baselines, this value increases, producing what is known as "the wedge" (Datta et al. 2010; Parsons et al. 2012; Vedantham et al. 2012; Pober et al. 2013; Thyagarajan et al. 2013; Liu et al. 2014a,b; Patil et al. 2017). The wedge describes a region in k-space contaminated by smooth spectrum foregrounds, bounded by baseline length (which is proportional to $k_\perp$) and delay (which is proportional to $k_\parallel$). Properties of the wedge can be used to isolate and avoid foregrounds, as done by A15, Parsons et al. (2014), Dillon et al. (2014), Dillon et al. (2015a), Jacobs et al. (2015), Beardsley et al. (2016), and Trott et al. (2016).

Although smooth-spectrum foregrounds preferentially show up at low delay, or low k-modes, their isolation within the wedge is not perfect. In deep measurements, power spectrum measurements at $k_\parallel$ values beyond the delay associated with the length of a baseline are often still contaminated at a low level. This leakage, particularly at low k's, can be attributed to convolution kernels associated with Fourier-transforming visibilities into delay-space. In other words, smooth-spectrum foregrounds appear as δ-functions in delay-space, convolved by the Fourier transform of the source spectrum, the signal chain, and the antenna response, all of which could smear out the foregrounds and cause leakage outside the wedge (e.g., Ewall-Wice et al. 2017; Kerrigan et al. 2018).

There are analysis techniques to mitigate the effects of foreground leakage and prevent information from low $k's$ from spreading to high k values. For example, narrow window functions in delay-space can be used to minimize the leakage from a particular k value into other ones (Liu et al. 2014b). In other words, one can construct an estimator using QE that forces a window function to have a minimum response to low k values. The window function used in A15 is constructed in such a way, specifically to prevent foregrounds that live at low $k's$ from contaminating higher k-modes (see Chapter 3.4).

Determining the source of positive non-EoR detections at higher k's is more difficult. In

previous power spectrum results, these detections have been explained as instrumental systematics, particularly time-variable cross talk, RFI, cable reflections, and calibration errors (A15; Parsons et al. 2014; Dillon et al. 2014; Beardsley et al. 2016; Patil et al. 2017). In the next section, we will present some tests that can help distinguish these excesses from that of EoR.

In addition to foreground bias, noise can also be responsible for positive power spectrum detections if thermal noise is multiplied by itself. Every 21 cm visibility measurement contains thermal noise that is comprised of receiver and sky noise. We expect this noise to be independent between antennas and thus we can beat it down (increase sensitivity) by integrating longer, using more baselines, etc. However, the squaring of noise can occur when cross-multiplying visibilities, which is shown by the two copies of $\mathbf{x}$ in Equation (2.8). If both copies of $\mathbf{x}$ come from the same baseline and time, it can result in power spectrum measurements that are higher than those predicted by the thermal noise of the instrument. One way to avoid this type of noise bias is to avoid cross-multiplying data from the same baselines or days. This ensures that the two quantities that go into a measurement have separate noises that don't correlate with each other. We also note that if the noise level is known, this type of bias can be subtracted off, though this procedure is argued to be dangerous (Dillon et al. 2014; Parsons et al. 2014).

Another type of noise bias can stem from the spurious cross-coupling of signals between antennas. This excess is known as instrumental crosstalk and is an inadvertent correlation between two independent measurements via a coupled signal path. Crosstalk varies on a time-scale much slower than the typical fringe-rates of sources. Because it is slow-varying, crosstalk can be suppressed using time-averages or fringe-rate filters. However, there remains a possibility that power spectrum detections that aren't the cosmological signal are caused by residual, low-level crosstalk which survived any suppression techniques.

2.5.2 Jackknife Tests

We now approach the difficult task of tracing excesses to foreground, noise, and EoR biases through a discussion of useful jackknife tests. Again, we first approach this topic pedagogically as an introduction to the related PAPER discussion in Chapter 3.4.

The jackknife is a resampling technique in which a statistic (i.e., power spectrum) is computed in subsets of the data (Quenouille 1949; Tukey 1958). These subsets are then compared to reveal systematics. In this section we define two main tests — the null test and the traditional jackknife — and explain how a power spectrum detection must pass each. We then highlight how these tests can be used to help distinguish between different sources of bias.

- **Null Test**: A null test is a type of jackknife test that removes the astronomical signal from data in order to investigate underlying systematics (see Keating et al. (2016) for examples from intensity mapping that are closely related to our current application). For example, one can divide data into two subsets by separating odd and

even Julian dates, or the first half of the observing season from the second. Subtracting the two removes signal that is common to both subsets, including foregrounds and the EoR signal. The resulting power spectrum should be consistent with thermal noise estimates; if it is not, it suggests the presence of a systematic that differs from one of the data subsets to the other (i.e., doesn't get subtracted perfectly).

- **Traditional Jackknife**: In a broader sense, it is important to perform many jackknife tests in order to instill confidence in a final result. A stable result must be steadfast throughout all jackknives no matter how the data is sliced. Jackknives can be taken along several different axes — for example, one could start with a full data set, and compute a new power spectrum every time as a day of data is removed, or a baseline is removed. This type of jackknife would reveal bias present only at certain LSTs (such as a foreground source), for example, or misbehaving baselines.

While the null test hunts for deviations from thermal noise and the jackknife tests for deviations in subsamples, they are both closely related. We can highlight the connection between the two using a toy model data set.

Suppose we have four independent measurements made along two different axes. As an example, we construct $\mathbf{x}_{1a}$, $\mathbf{x}_{1b}$, $\mathbf{x}_{2a}$, and $\mathbf{x}_{2b}$, where the numbers symbolize two different days of data and the letters represent two different baselines. Each of the measurements have dimensions of 100 time integrations and 20 frequency channels. They each have separate thermal noises constructed as a Gaussian random signal for each, and identical EoR signals.

To mimic the presence of a systematic, we add a toy sinusoid foreground, similar to the one used in Chapter 2.2.2, to only $\mathbf{x}_{2a}$ and $\mathbf{x}_{2b}$. This represents a foreground signal present in only the second day of data, but not the first. Mathematically, if $\mathbf{n}$ is noise, $\mathbf{e}$ is the EoR signal, and $\mathbf{f}$ is the foreground signal, the four measurements can be written as:

$$\mathbf{x}_{1a} = \mathbf{n}_{1a} + \mathbf{e} \tag{2.36}$$
$$\mathbf{x}_{1b} = \mathbf{n}_{1b} + \mathbf{e} \tag{2.37}$$
$$\mathbf{x}_{2a} = \mathbf{n}_{2a} + \mathbf{e} + \mathbf{f} \tag{2.38}$$
$$\mathbf{x}_{2b} = \mathbf{n}_{2b} + \mathbf{e} + \mathbf{f}. \tag{2.39}$$

We now take a jackknife along the day-axis, forming separate power spectrum estimates for each day:

$$\widehat{\mathbf{P}}_1 \propto \mathbf{x}_{1a}\mathbf{x}_{1b}^\dagger \tag{2.40}$$
$$\widehat{\mathbf{P}}_2 \propto \mathbf{x}_{2a}\mathbf{x}_{2b}^\dagger. \tag{2.41}$$

We do not perform a time-average or apply a fringe-rate filter to this toy model, since we are interested only in what jackknife tests can tell us about biases. For the same reason, we use a weighting matrix of $\mathbf{I}$ for power spectrum estimation to avoid signal loss.

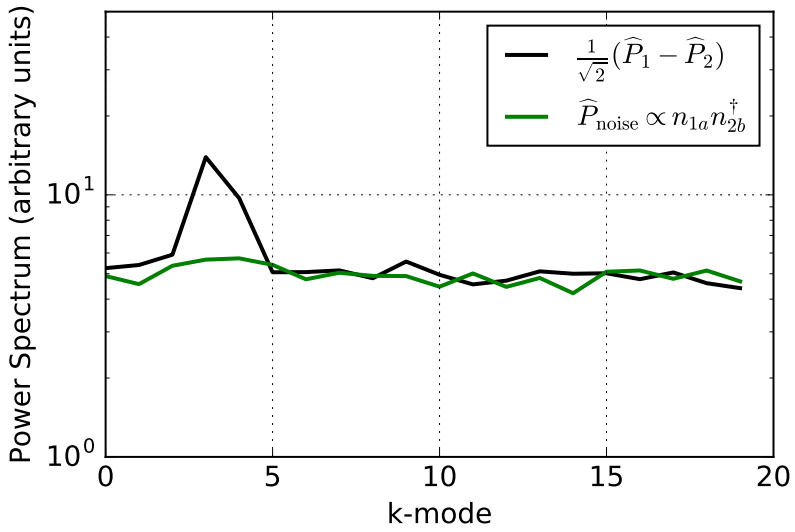

Figure 2.10: A null jackknife test shown as the power spectrum difference between two measurements (black), compared to the power spectrum of noise alone (green). Because the null test is not consistent with noise, it suggests the presence of a systematic in either $\mathbf{x}_1$ or $\mathbf{x}_2$. Null tests of clean measurements should be consistent with thermal noise.

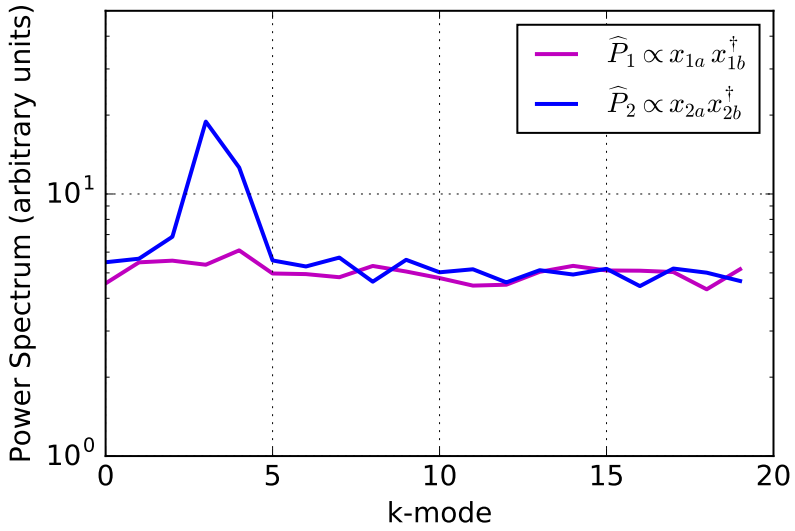

Figure 2.11: Power spectrum estimates for $\mathbf{x}_1$ and $\mathbf{x}_2$, two jackknives of the toy model. They suggest the presence of a systematic in $\mathbf{x}_2$ only, illustrating how jackknives can be used to tease out excesses. Clean measurements should remain consistent despite the jackknife taken.

To construct a null test, we difference the two power spectra, with the result shown in Figure 2.10 (black) along with the power spectrum of noise only (green). Subtracting the two estimates removes sky signal that should ideally be present in both jackknives. However, we see a clear difference between the null test and the power spectrum of noise. This signifies a non-EoR bias that is only present in either $\mathbf{x}_1$ or $\mathbf{x}_2$, but not both.

While the null test is useful for testing noise properties and the uniformity of a data set, jackknives are useful in pinpointing which data subsets are contaminated by biases and which are not; in our toy model we see that the bias exists only in $\mathbf{x}_2$ (Figure 2.11). If foreground or noise biases exist in a data set, jackknives can tease them out and provide insight into possible sources. For example, if jackknives along the time-axis reveal a bias present at a certain LST, a likely explanation would be excess foreground emission from a radio source in the sky at that time. A jackknife test involving data before and after the application of a fringe-rate filter can reveal whether crosstalk noise bias is successfully suppressed with the filter, or if similar-shaped detections in both power spectra suggest otherwise. There are many other jackknife axes of which we will not go into detail here, including baseline, frequency, and polarization. Ultimately, an EoR detection should persist through them all and a clean measurement should exhibit noise-like null spectra.

In this section we have highlighted how null tests and jackknife tests are key for determining the nature of a power spectrum detection. In Chapter 3.4 we perform some examples of these tests on PAPER-64 data in order to show that our excesses are not EoR and to identify their likely cause.

Chapter 3

PAPER-64 Case Study

3.1 Overview

In the previous chapter we have discussed three overarching 21 cm power spectrum themes: signal loss, error estimation, and bias. Understanding the subtleties and trade-offs involved in each is necessary for an accurate and robust understanding of a power spectrum result.

We now apply these lessons to a subset of data from the PAPER experiment in order to illustrate our revised analysis pipeline. We begin with a brief overview of PAPER's data processing steps prior to power spectrum estimation before delving into each theme in detail.

3.1.1 Observations

As described in Chapter 1.3.1, PAPER is a dedicated 21 cm experiment located in the Karoo Desert in South Africa. The PAPER-64 configuration consists of 64 dual-polarization drift-scan elements that are arranged in a grid layout (Figure 3.1). While every unique baseline is used for calibration, only a subset of the baselines are used for the power spectrum analysis in A15 and Chapter 4 (the three baselines used are the 30 m East/West baselines and their off-diagonal companions where two antennas are in adjacent columns and neighboring rows) and only one baseline-type is used for the "case study" demonstrations in this chapter (only the 30 m East/West baselines).

PAPER-64 conducted nighttime observations from November 2012 to March 2013. Over the course of the season, LST-coverage varied slightly, with power spectrum analyses focusing on the "cold patch" range from $\sim$ 0-8 hours when the Galaxy is below the horizon. The PAPER correlator processes a 100-200 MHz bandwidth that consists of 1024 channels, each of width 97.6 kHz. Visibilities are integrated for 10.7 s before being written to disk.

PAPER's raw data is compressed by a factor of ~ 70 through the use of RFI, delay, and delay-rate filters. More specifically, radio frequency interference is flagged at the 6σ level. Next, a low-pass delay filter is applied to all the data in order to filter out delays greater than the maximum delay allowed by the longest baseline in the array. Similarly, a low-pass

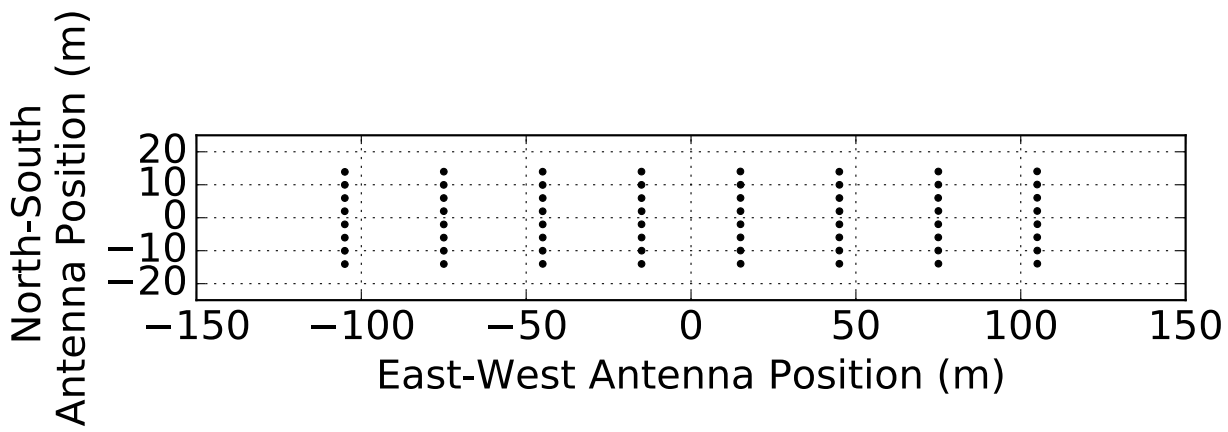

Figure 3.1: The PAPER-64 antenna layout. We use only 10 of the 30 m East/West baselines for the analysis in this chapter (i.e., a subset of the shortest horizontal spacings).

fringe-rate, or delay-rate filter is applied to limit fringe-rates to allowable scales set by the array. Finally, the data is decimated to critical Nyquist sampling rates of 493 kHz and 42.9 s. For more details about PAPER's data acquisition and compression pipelines, we refer the reader to Parsons et al. (2010) and A15.

3.1.2 Data Processing

As described in Chapters 1.2.3, 1.2.4, and 1.2.5, the primary post-processing steps of PAPER's compressed data is calibration, foreground-filtering, and fringe-rate filtering. We now give a brief overview of each, as applied to PAPER data.

We employ the package `Omnical` for redundant calibration (Zheng et al. 2014), which comprises of three steps. The first is `FirstCal`, which uses all baseline redundancies to generate a static gain solution for each antenna that will unwrap any phase wrapping between two identical baselines. We perform `FirstCal` because the next stage of `Omnical` cannot tell the difference between a phase of 0 and 2π, for example. The second step is `LogCal`, which takes the log of all the visibility equations (Equation (1.8)) and separates the real and imaginary components into two matrices. Coarse solutions are determined for both the antenna gains and "model" visibilities (one for each baseline type) simultaneously. The final step of `Omnical` is `LinCal`, which applies small perturbations to the `LogCal` solutions in an iterative fashion, honing in on the optimal solutions.

It is important to note that while `Omnical` is powerful for ensuring array redundancy, it is not able to solve for four calibration parameters - namely, the overall gain, phase, and tip/tilt of the array. For absolute calibration, we turn to a standard self-calibration routine which includes imaging Pictor A, Fornax A, and the Crab Nebula in order to fit for the overall phase solutions and the flux scale.

After calibration, we combine the XX and YY linear polarization data to form pseudo-Stokes I as defined as:

$$V_I = \frac{1}{2}(V_{XX} + V_{YY}) \tag{3.1}$$

(Moore et al. 2013).

Next, a delay-filter is used to filter out foregrounds contained inside the maximum delay set by each baseline. This is accomplished by de-convolving out our sampling function (which contains flags due to RFI) from our delay-domain visibilities using a CLEAN-like algorithm that restricts our clean components to inside the horizon limit, plus a 15 ns buffer. The Fourier-transformed clean components are then subtracted from our visibilities. This filtering process is performed on a per-baseline, per-integration basis, and we achieve a brightness suppression of ~ 4 orders of magnitude in our visibilities.

After delay-filtering, we perform a final round of RFI-removal by flagging visibilities that lie more than 3σ above the mean on a time, frequency, and baseline basis. Finally, we stack our data in LST into two data sets, alternating between even and odd Julian dates to create an "even" and "odd" LST-binned data set. A total of 124 days of data are included in the LST-binned data set.

The final step before power spectrum estimation is fringe-rate filtering. The chosen filter (which is described in the next section) is applied on a per-baseline basis and weights the fringe-rate bins on the sky by the RMS of the primary beam at that same location. A smooth filter is constructed by fitting a Gaussian to the filter shape in the fringe-rate domain. Additionally, fringe-rates below 0.2 mHz are zeroed out, effectively removing slowly-varying signals such as crosstalk. We then convolve our time-domain visibilities by the Fourier-transform of the fringe-rate filter to yield time-averaged visibilities that have gained another order of magnitude in sensitivity.

3.1.3 Case Study Data

For the case study presented in the rest of this chapter, we focus on a subset of the PAPER-64 data used in A15, namely, on LST-binned, Stokes I estimated data (Moore et al. 2013) from PAPER's 30 m East/West baselines (Figure 3.1). Hence, all data processing steps are identical to those in A15 until after the LST-binning step in Figure 3 of A15. We foreshadow the use of the full A15 data set in Chapter 4 when revising our upper limits.

The previously best published 21 cm upper limit result from A15 placed a 2σ upper limit on $\Delta^2(k)$, defined as

$$\Delta^2(k) = \frac{k^3}{2\pi^2}\hat{P}(k), \tag{3.2}$$

of $(22.4\text{ mK})^2$ in the range $0.15 < k < 0.5\,h$ Mpc^{-1} at $z = 8.4$. The need to revise this limit stems mostly from previously underestimated signal loss and underestimated error bars, both of which we address in the following sections.

For the analysis in this chapter, we use 8.1 hours of LST, namely an RA range of 0.5-8.6 hours (A15 uses a slightly longer RA range of 0-8.6 hours; we found that some early LSTs were more severely foreground contaminated). We also use only 10 baselines, a subset of the 51 total East/West baselines used in A15, in order to illustrate our revised methods. All power spectrum results are produced for a center frequency of 151 MHz using a width of 10 MHz (20 channels), identical to the analysis in A15. In the case study in this chapter, we only use one baseline type instead of the three as in A15, but Chapter 4 uses the full data set presented in A15 to revise the result and place limits on the EoR at multiple redshifts (using a straightforward and not lossy approach to avoid many of the issues presented in this chapter).

The most significant changes from A15 occur in our revised power spectrum analysis, which is explained in the rest of this chapter, but we also note that the applied fringe-rate filter is also slightly different. In A15, the applied filter was not equivalent to the optimal fringe-rate filter (which is designed to maximize power spectrum sensitivity). Instead, the optimal filter was degraded slightly by widening it in fringe-rate space. This was chosen in order to increase the number of independent modes and reduce signal loss associated with the quadratic estimator, though as we will explain in the next section, this signal loss was still underestimated. With the development of a new, robust method for assessing signal loss, we choose to use the optimal filter in order to maximize sensitivity. This filter is computed for a fiducial 30 m baseline at 150 MHz, the center frequency in our band. The filter in both the fringe-rate domain and time domain is shown in Figure 3.2.

Finally, we emphasize that the discussion that follows is solely focused on signal loss associated with empirical covariance weighting. As mentioned in Chapter 2.2, there are a number of steps in our analysis pipeline which could lead to loss, including gain calibration, delay filtering, and fringe-rate filtering, which have been investigated at various levels of detail in Parsons et al. (2014) and A15 but are clearly the subject of future work. Here we only focus on the most significant source of loss we have identified and note that Chapter 4 and other future work will consider additional sources of signal loss and exercise increased caution in reporting results.

3.2 Signal Loss

We present our PAPER-64 signal loss investigation in three parts. We first give an overview of our signal injection framework which is used to estimate loss (Chapter 3.2.1). In this framework (and as in A15), we inject simulated cosmological signals into our data and test the recovery of those signals (an approach also taken by Masui et al. (2013)). As we will see, correlations between the injected signals and the data are significant complicating factors which were previously not taken into account. Next, we describe our methodology in practice and detail how we map our simulations into a posterior for the EoR signal (Chapter 3.2.2). Finally, we build off of the previous section by experimenting with different regularization schemes on PAPER data in order to minimize loss (Chapter 3.2.3). Throughout each section,

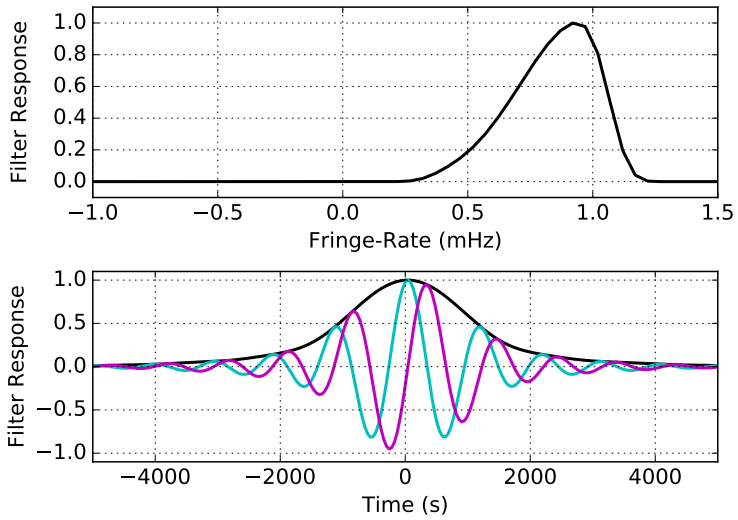

Figure 3.2: Top: the normalized optimal power-spectrum sensitivity weighting in fringe-rate space for our fiducial baseline and Stokes I polarization beam. Bottom: the time domain convolution kernel corresponding to the top panel. Real and imaginary components are illustrated in cyan and magenta, respectively, with the absolute amplitude in black. The fringe-rate filter acts as an integration in time, increasing sensitivity but reducing the number of independent samples in the data set.

we also highlight major differences from the signal loss computation used in A15.

3.2.1 Signal Loss Methodology

In short, our method for estimating signal loss consists of adding an EoR-like signal into visibility data and then measuring how much of this injected signal would be detectable given any attenuation of this signal by the (lossy) data analysis pipeline. To capture the full statistical likelihood of signal loss, one requires a quick way to generate many realizations of simulated 21 cm signal visibilities. Here we use the same method as in A15, where mock Gaussian noise visibilities (mock EoR signals) are filtered in time using an optimal fringe-rate filter to retain only "sky-like" modes. Since the optimal filter has a shape that matches the rate of the sidereal motion of the sky, this transforms the Gaussian noise into a measurement that PAPER could make. This signal is then added to the visibility data.[1]

[1]One specific change from A15 is that we add this simulated signal - which has been fringe-rate filtered once already in order to transform it into a "sky-like" signal - into the analysis pipeline before a fringe-rate filter is applied to the data (i.e., prior to the analysis step of fringe-rate filtering). Previously, the addition was done after the fringe-rate filter analysis step. This change results in an increased estimate of signal loss, likely due to the use of the fringe-rate filter as a simulator. However, this pipeline difference, while significant, is not the dominant reason why signal loss was underestimated in A15 (the dominant reason is explained in the main text in Chapter 3.2.1).

Mathematically, suppose that $\mathbf{e}$ is the mock injected EoR signal (at some amplitude level). We do not know the true EoR signal contained within our visibility data, $\mathbf{x}$, so $\mathbf{e}$ takes on the role of the true EoR signal (for which we measure its loss). Furthermore, one can make the assumption that the true EoR signal is small within our measured data, so the data vector $\mathbf{x}$ itself is representative of mostly contaminants. Using this assumption, the sum of $\mathbf{x}$ and $\mathbf{e}$, defined as $\mathbf{r}$:

$$\mathbf{r} = \mathbf{x} + \mathbf{e}, \tag{3.3}$$

can be thought of as the sum of contaminants plus EoR. The quantity $\mathbf{r}$ then becomes the data set for which we are measuring how much loss of $\mathbf{e}$ there is due to our power spectrum pipeline.

We are interested in quantifying how much variance in $\mathbf{e}$ is lost after weighting $\mathbf{r}$ and estimating the power spectrum according to QE formalism. We investigate this by comparing two quantities we call the input power spectrum and output power spectrum: $\widehat{P}_{\text{in}}$ and $\widehat{P}_{\text{out}}$, estimated using QE as

$$\widehat{P}^{\alpha}_{\text{in}} \equiv \mathrm{M}^{\alpha}_{\text{in}} \mathbf{e}^{\dagger} \mathbf{I} \mathbf{Q}^{\alpha} \mathbf{I} \mathbf{e} \tag{3.4}$$

and

$$\begin{aligned} \widehat{P}^{\alpha}_{\text{out}} &\equiv \widehat{\mathbf{P}}^{\alpha}_{r} \\ &= \mathrm{M}^{\alpha}_{r} \mathbf{r}^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r \mathbf{r}, \end{aligned} \tag{3.5}$$

where, for illustrative purposes and notational simplicity, we have written these equations with scalar normalizations M, even though for our numerical results we choose a diagonal matrix normalization using $\mathbf{M}$ as in Equation (2.9).

The quantity $\widehat{P}_{\text{in}}$, defined by Equation (3.4), is a uniformly weighted estimator of the power spectrum of $\mathbf{e}$. It can be considered the power spectrum of this particular realization of the EoR; alternatively, it can be viewed as the true power spectrum of the injected signal up to cosmic variance fluctuations. The role of $\widehat{P}_{\text{in}}$ in our analysis is to serve as a reference for the power spectrum that would be measured if there were no signal loss or other systematics. The input power spectrum is then to be compared to $\widehat{P}_{\text{out}}$, which approximates the (lossy) power spectrum estimate that is output by our analysis pipeline prior to any signal loss adjustments.

Under this injection framework, we can begin to see explicitly why there can be large signal loss. Expanding out Equation (3.5), $\widehat{P}_{\text{out}}$ becomes:

$$\begin{aligned} \widehat{P}^{\alpha}_{\text{out}} &= \mathrm{M}^{\alpha}_{r} (\mathbf{x} + \mathbf{e})^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r (\mathbf{x} + \mathbf{e}) \\ &= \mathrm{M}^{\alpha}_{a} \mathbf{x}^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r \mathbf{x} + \mathrm{M}^{\alpha}_{b} \mathbf{e}^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r \mathbf{e} \\ &+ \mathrm{M}^{\alpha}_{c} \mathbf{x}^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r \mathbf{e} + \mathrm{M}^{\alpha}_{d} \mathbf{e}^{\dagger} \mathbf{R}_r \mathbf{Q}^{\alpha} \mathbf{R}_r \mathbf{x}. \end{aligned} \tag{3.6}$$

Assuming $\mathbf{R}_r$ is symmetric, the two cross-terms (terms with one copy of $\mathbf{e}$ and one copy of $\mathbf{x}$) can be summed together as:

$$\begin{aligned}\widehat{P}^{\alpha}_{\text{out}} &= \mathrm{M}^{\alpha}_{a}\mathbf{x}^{\dagger}\mathbf{R}_r\mathbf{Q}^{\alpha}\mathbf{R}_r\mathbf{x} + \mathrm{M}^{\alpha}_{b}\mathbf{e}^{\dagger}\mathbf{R}_r\mathbf{Q}^{\alpha}\mathbf{R}_r\mathbf{e} \\ &+ 2\mathrm{M}^{\alpha}_{c}\mathbf{x}^{\dagger}\mathbf{R}_r\mathbf{Q}^{\alpha}\mathbf{R}_r\mathbf{e}. \end{aligned} \tag{3.7}$$

One of the key takeaways of this section is that the A15 analysis estimated signal loss by comparing *only* the signal-only term (second term in Equation (3.7)) with $\widehat{P}_{\text{in}}$, whereas in fact the cross-term (third term in Equation (3.7)) can substantially lower $\widehat{P}_{\text{out}}$. In order to investigate the effect of each of these terms on signal loss, all three components are plotted in Figure 3.3 for two cases: empirically estimated inverse covariance weighting ($\mathbf{R}_r \equiv \widehat{\mathbf{C}}_r^{-1}$) and uniform weighting ($\mathbf{R}_r \equiv \mathbf{I}$). We will now go into further detail and examine the behavior of this equation in three different regimes of the injected signal - very weak (left ends of the P_{in} axes in Figure 3.3), very strong (right ends), and in between (middle portions).

Small injection: In this regime, the cross-terms (red) behave as noise averaged over a finite number of samples. Output values are Gaussian distributed around zero, spanning a range of values set by the injection level. This is because $\widehat{\mathbf{R}}_r$ is dominated by the data $\mathbf{x}$, avoiding correlations with $\mathbf{e}$ that can lead to solely negative power (explained further below). In fact, for the uniformly weighted case, the cross-term $\mathrm{M}^{\alpha}_{x}\mathbf{x}^{\dagger}\mathbf{I}\mathbf{Q}^{\alpha}\mathbf{I}\mathbf{e}$ is well modeled as a symmetric distribution with zero mean and width $\sqrt{\widehat{\mathbf{P}}_e}\sqrt{\widehat{\mathbf{P}}_x}$. We also note that in this regime, $\widehat{\mathbf{P}}_r$ (black) approaches the data-only power spectrum value (gray) as expected.

Large injection: When the injected signal is much larger than the measured power spectrum, the data-only components can be neglected as they are many orders of magnitude smaller. We include a description of this regime for completeness in our discussion, but note that the upper limits that we compute are typically not determined by simulations in this regime (i.e., in using an empirical weighting scheme we've assumed the data to be dominated by foregrounds rather than the cosmological signal). However, it is useful as a check of our system in a relatively simple case. As we can see from Figure 3.3, the cross-terms (red) are small in comparison to the signal-only term (green). Here only does the signal-only term used in A15 dominate the total power output. We again see that, in the empirical inverse covariance weighted case, the cross-terms behave as noise (positive and negative fluctuations around zero mean). This is for the same reason as at small injections — here $\widehat{\mathbf{C}}_r$ is dominated by the signal $\mathbf{e}$. The cross-correlation can again be modeled as a symmetric distribution of zero mean and width $\sqrt{\widehat{\mathbf{P}}_e}\sqrt{\widehat{\mathbf{P}}_x}$.

In between: When the injected signal is of a similar amplitude to the data by itself, the situation becomes less straightforward. We see that the weighted injected power spectrum component mirrors the input power indicating little loss (i.e., the green curve follows the dotted black line), eventually departing from unity when the injected amplitude is well above the level of the data power spectrum. However, in this regime the cross-term (red) has nearly the same amplitude, but with a negative sign. As explained below, this negativity is the

result of cross-correlating inverse covariance weighted terms. This negative component drives down the $\widehat{P}_{\rm out}$ estimator (black). Again, we emphasize that in A15, signal loss was computed by only looking at the second term in Equation (3.7) (green), which incorrectly implies no loss at the data-only power spectrum level. Ignoring the effect of the negative power from the cross-terms is the main reason for underestimating power spectrum limits in A15.

The source of the strong negative cross-term is not immediately obvious, however it is an explainable effect. When $\mathbf{R}_r$ is taken to be $\widehat{\mathbf{C}}_r^{-1}$, the third term of Equation (3.7) is a cross-correlation between $\widehat{\mathbf{C}}_r^{-1}\mathbf{x}$ and $\widehat{\mathbf{C}}_r^{-1}\mathbf{e}$. As shown in Switzer et al. (2015), this cross-correlation term is non-zero, and in fact negative in expectation. This negative cross-term power arises from a coupling between the inverse of $\widehat{\mathbf{C}}_r$ and $\mathbf{x}$. Intuitively, we can see this by expanding the empirical covariance of $\mathbf{r} = \mathbf{x} + \mathbf{e}$:

$$\begin{aligned}\widehat{\mathbf{C}}_r &= \langle \mathbf{r}\mathbf{r}^\dagger \rangle_t \\ &= \langle \mathbf{x}\mathbf{x}^\dagger \rangle_t + \langle \mathbf{x}\mathbf{e}^\dagger \rangle_t + \langle \mathbf{e}\mathbf{x}^\dagger \rangle_t + \langle \mathbf{e}\mathbf{e}^\dagger \rangle_t, \end{aligned} \tag{3.8}$$

where we can neglect the first term because $\mathbf{x}$ is small (i.e., the large negative cross-term power in the left panel of Figure 3.3 occurs when the injected amplitude surpasses the level of the data-only power spectrum). Without loss of generality, we will assume an eigenbasis of $\mathbf{e}$, so that $\langle \mathbf{e}\mathbf{e}^\dagger \rangle_t$ is diagonal. The middle two terms, however, can have power in their off-diagonal terms due to the fact that, when averaging over a finite ensemble, $\langle \mathbf{x}\mathbf{e}^\dagger \rangle_t$ is not zero. As shown in Appendix C of Parsons et al. (2014), to leading order the inversion of a diagonal-dominant matrix like $\widehat{\mathbf{C}}_r$ (from $\langle \mathbf{e}\mathbf{e}^\dagger \rangle_t$) with smaller off-diagonal terms results in a new diagonal-dominant matrix with negative off-diagonal terms. These off-diagonal terms depend on both $\mathbf{x}$ and $\mathbf{e}$. Then, when $\widehat{\mathbf{C}}_r^{-1}$ is multiplied into $\mathbf{x}$, the result is a vector that is similar to $\mathbf{x}$ but contains a residual correlation to $\mathbf{e}$ from the off-diagonal components of $\widehat{\mathbf{C}}_r^{-1}$. The correlation is negative because the product $\widehat{\mathbf{C}}_r^{-1}\mathbf{x}$ effectively squares the $\mathbf{x}$-dependence of the off-diagonal terms in $\widehat{\mathbf{C}}_r^{-1}$ while retaining the negative sign that arose from the inversion of a diagonal-dominant matrix.

In general: Another way to phrase the shortcoming of the empirical inverse covariance estimator is that it is not properly normalized. Signal loss due to couplings between the data and its weightings arise because our unnormalized quadratic estimator from Equation (2.8) ceases to be a quadratic quantity, and instead contains higher order powers of the data. However, the normalization matrix $\mathbf{M}$ is derived assuming that the unnormalized estimator is quadratic in the data. The power spectrum estimate will therefore be incorrectly normalized, which manifests as signal loss. We leave a full analytic solution for $\mathbf{M}$ for future work, since our simulations already capture the full phenomenology of signal loss and have the added benefit of being more easily generalizable in the face of non-Gaussian systematics.

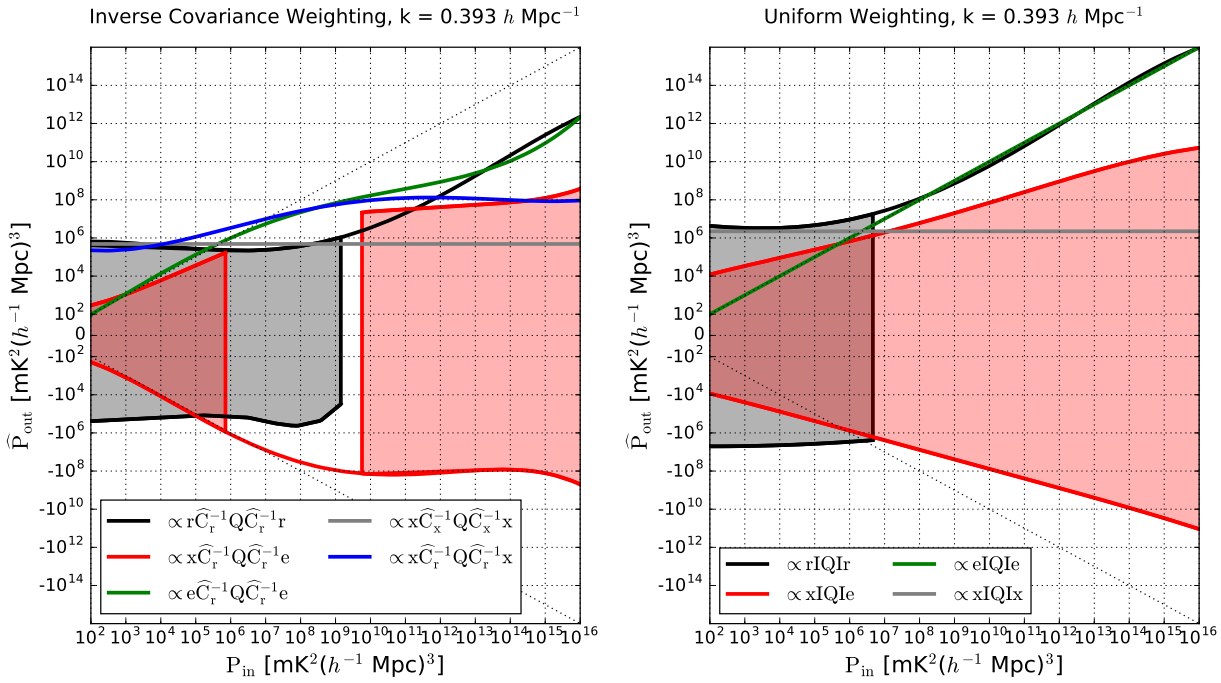

Figure 3.3: Illustration of the power spectrum amplitude of five different power spectrum terms, each a function of visibility data (**x**), simulated injected EoR signal (**e**), or both (**r**). This figure shows how these quantities behave as the power level of the injected EoR signal increases (along the x-axis). The details of the simulation used to generate the figure is explained in Chapter 3.2.2; here we sample a larger $P_{\rm in}$ range and fit smooth polynomials to our data points to make an illustrative example. We emphasize that the output power spectrum in black ($\widehat{P}_{\rm out} = \widehat{\mathbf{P}}_r$) approximates the (lossy) power spectrum estimate that is output by our analysis pipeline prior to any signal loss adjustments. Roughly speaking, it can be compared to the input signal level ($P_{\rm in}$) to estimate the amount of signal loss. Left: Empirical inverse covariance weighting is used in power spectrum estimation, as done in Ali et al. (2015). The dotted diagonal black line indicates perfect 1:1 input-to-output mapping (no signal loss). The gray horizontal line is the power spectrum value of data alone, $\widehat{\mathbf{P}}_x$ (it does not depend on injected power). The green signal-signal component is the term used in Ali et al. (2015) to estimate signal loss. It is significantly higher than $\widehat{\mathbf{P}}_r$ (black) when the cross-terms (red) are large and negative (black = green + red + blue). In the regime where cross-correlations between signal and data are not dominant (small and large $P_{\rm in}$), the cross-terms have a noise-like term with width $\sqrt{\widehat{\mathbf{P}}_e}\sqrt{\widehat{\mathbf{P}}_x}$. However, at power levels comparable to the data (the middle region), the cross-terms can produce large, negative estimates due to couplings between **x** and **e** which affect $\widehat{\mathbf{C}}_r$. This causes the difference between the green curve (which exhibits negligible loss at the data-only power spectrum value) and the black curve (which exhibits ~ 4 orders of magnitude of loss). Right: The same power spectrum terms illustrated for the uniform weighted case.

3.2.2 Signal Loss in Practice

We now shift our attention towards computing upper limits on the EoR signal for the fringe-rate filtered PAPER-64 data set in a way that accounts for signal loss. While our methodology outlined below is independent of weighting scheme, here we demonstrate the computation using empirically estimated inverse covariance weighting ($\mathbf{R} \equiv \widehat{\mathbf{C}}^{-1}$), the weighting scheme used in A15 which leads to substantial loss.

One issue to address is how one incorporates the randomness of $\widehat{P}_{\rm out}$ into our signal loss corrections. A different realization of the mock EoR signal is injected with each bootstrap run, causing the output to vary in three ways — there is noise variation from the bootstraps, there is cosmic variation from generating multiple realizations of the mock EoR signal, and there is a variation caused by whether the injected signal looks more or less "like" the data (i.e., how much coupling there is, which affects how much loss results).

For each injection level, the true $P_{\rm in}$ is simply the average of our bootstrapped estimates $\widehat{P}_{\rm in}$, since $\widehat{P}_{{\rm in},\alpha}$ is by construction an unbiased estimator. Phrased in the context of Bayes' rule, we wish to find the posterior probability distribution $p(P_{\rm in}|\widehat{P}_{\rm out})$, which is the probability of $P_{\rm in}$ given the uncorrected/measured power spectrum estimate $\widehat{P}_{\rm out}$. Bayes' rule relates the posterior, which we don't know, to the likelihood, which we can forward model. In other words,

$$p(P_{\rm in}|\widehat{P}_{\rm out}) \propto \mathcal{L}(\widehat{P}_{\rm out}|P_{\rm in})\, p(P_{\rm in}), \tag{3.9}$$

where $\mathcal{L}$ is the likelihood function defined as the distribution of data plus signal injection ($\widehat{P}_{\rm out}$) given the injection $P_{\rm in}$. We construct this distribution by fixing $P_{\rm in}$ and simulating our analysis pipeline for many realizations of the injected EoR signal consistent with this power spectrum. The resulting distribution is normalized such that the sum over $\widehat{P}_{\rm out}$ is unity, and the whole process is then repeated for a different value of $P_{\rm in}$.

The implementation details of the injection process require some more detailed explanation. In our code, we add a new realization of EoR to each independent bootstrap of data (see Chapter 3.3.1 for a description of PAPER's bootstrapping routine) with the goal of simultaneously capturing cosmic variance, noise variance, and signal loss. To limit computing time we perform 20 realizations of each $P_{\rm in}$ level. We also run 50 total EoR injection levels, yielding $P_{\rm in}$ values that range from $\sim 10^5\,{\rm mK}^2\ (h^{-1}\ {\rm Mpc})^3$ to $\sim 10^{11}\,{\rm mK}^2\ (h^{-1}\ {\rm Mpc})^3$, resulting in a total of 1000 data points on our $P_{\rm in}$ vs. $\widehat{P}_{\rm out}$ grid.

Going forward, we treat every k-value separately in order to determine an upper limit on the EoR signal per k. We bin our simulation outputs along the $P_{\rm in}$ axis (one bin per injection level) and, since they are well-approximated by a Gaussian distribution in our numerical results, we smooth the distribution of $\widehat{P}_{\rm out}$ values by fitting Gaussians for each bin based on its mean and variance (and normalize them). Stitching all of them together results in a 2-dimensional transfer function — the likelihood function in Bayes' rule, namely $\mathcal{L}(\widehat{P}_{\rm out}|P_{\rm in})$. We then have a choice for our prior, $p(P_{\rm in})$, and we choose to invoke a Jeffreys prior (Jaynes 1968) because it is a true uninformative prior for a parameter space using

Bayesian probability.

The Jeffreys prior is defined as:

$$p(P_{\rm in}) \propto \sqrt{\left\langle \left(\frac{\partial \mathcal{L}}{\partial P_{\rm in}} \right)^2 \right\rangle}, \tag{3.10}$$

where

$$\mathcal{L} = \ln p(\widehat{P}_{\rm out} | P_{\rm in}), \tag{3.11}$$

recalling that in our framework $P_{\rm in}$ is the power spectrum of the EoR signal (uniformly weighted), and $\widehat{P}_{\rm out}$ is the weighted output power spectrum of the data plus EoR.

Since, for a single injection amplitude, our bootstrapped $\widehat{P}_{\rm out}$ values are well-approximated by a Gaussian distribution, we can write:

$$p(y|x) = \frac{1}{\sigma(x)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\bar{y}(x)}{\sigma}\right)^2}, \tag{3.12}$$

simplifying our notation so that $x = P_{\rm in}$, $y = \widehat{P}_{\rm out}$, σ is the standard deviation of $\widehat{P}_{\rm out}$, and $\bar{y}$ is the mean of $\widehat{P}_{\rm out}$. Using Equations (3.12) and (3.11), the quantity inside the expectation value of Equation (3.10) becomes:

$$\begin{aligned} \left(\frac{\partial \mathcal{L}}{\partial x}\right)^2 &= \frac{1}{\sigma^2}\left(\frac{\partial \sigma}{\partial x}\right)^2 - \left(\frac{2(y-\bar{y})}{\sigma^3}\right)\frac{\partial \sigma}{\partial x}\frac{\partial \bar{y}}{\partial x} - \left(\frac{2(y-\bar{y})^2}{\sigma^4}\right)\left(\frac{\partial \sigma}{\partial x}\right)^2 \\ &+ \left(\frac{(y-\bar{y})^2}{\sigma^4}\right)\left(\frac{\partial \bar{y}}{\partial x}\right)^2 + \left(\frac{2(y-\bar{y})^3}{\sigma^5}\right)\frac{\partial \sigma}{\partial x}\frac{\partial \bar{y}}{\partial x} + \left(\frac{(y-\bar{y})^4}{\sigma^6}\right)\left(\frac{\partial \sigma}{\partial x}\right)^2. \end{aligned} \tag{3.13}$$

Taking the expectation value then removes all terms with odd powers of $(y-\bar{y})$ because those Gaussian moments evaluate to zero. Additionally, the second moment can be simplified since $\langle (y-\bar{y})^2 \rangle = \sigma^2$ and the fourth moment can be simplified since $\langle (y-\bar{y})^4 \rangle = 3\sigma^4$. Finally, after some additional simplification the Jeffreys prior becomes:

$$p(x) \propto \sqrt{\frac{1}{\sigma^2}\left(2\left(\frac{\partial \sigma}{\partial x}\right)^2 + \left(\frac{\partial \bar{y}}{\partial x}\right)^2\right)}. \tag{3.14}$$

When we simulate our full injection framework, we note that the prior is set to zero outside our injection range. For the injections that we do sample, we can simply fit analytic functions to the mean and standard deviations of $\widehat{P}_{\rm out}$ ($\bar{y}$ and σ) as functions of $P_{\rm in}$. An example of the typical shape of these functions for the PAPER-64 analysis is shown in Figure 3.4, though in practice we fit solutions for every k-value and simulation independently.

We also show the typical shape of the Jeffreys prior used in our analysis in Figure 3.5, as computed by Equation (3.14). Most noticeably, it is not constant with $P_{\rm in}$, meaning a uniform prior, which is often used for simplicity, is informative in our application. Therefore,

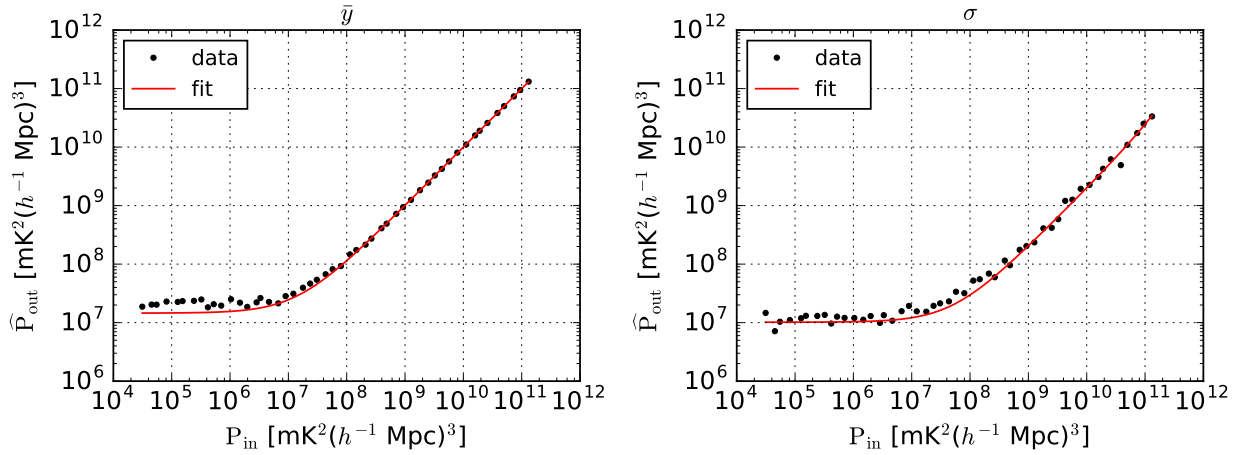

Figure 3.4: An illustrative example (for the PAPER-64 analysis using uniform weighting and $k = 0.393\,h$ Mpc^{-1}) of how the mean of P_{out} (left) and standard deviation of P_{out} (right) behave as a function of P_{in}. Polynomials are fit to each (red) to describe how $\bar{y}$ and σ evolve with x (injection level), respectively, for the computation of the Jeffreys prior as defined in Equation (3.14). The polynomial fits for this example are $y = (-5.1\times10^{-15})x^2+x+(1.5\times10^{7})$ and $y = (5.0 \times 10^{-13})x^2 + 0.2x + 10^7$ for $\bar{y}$ and σ, respectively.

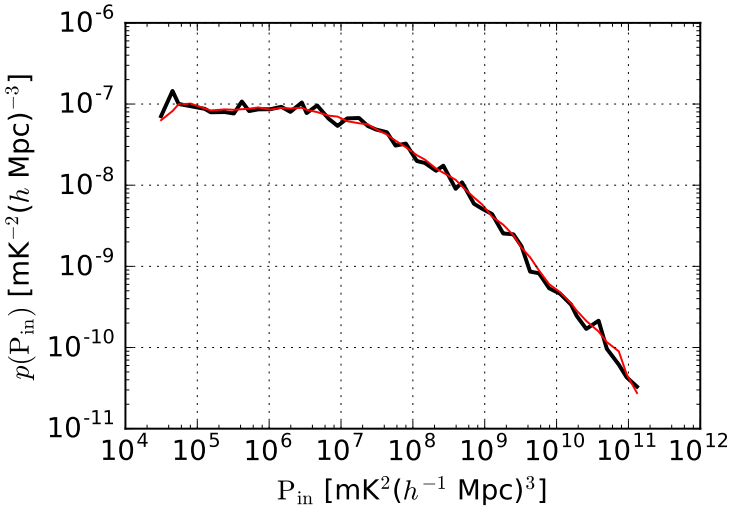

Figure 3.5: An example of the typical Jeffreys prior shape for the PAPER-64 analysis as computed by Equation (3.14) (black). We smooth the prior using a sliding boxcar average over every 5 injection levels (red). Most noticeably, the Jeffreys prior is not constant with P_{in}, meaning a uniform prior would be an informative prior.

due to its objective nature we choose to use a Jeffreys prior in our analysis, multiplying our likelihood functions by Equation (3.14) before computing posterior distributions.

Finally, our transfer functions are shown in Figure 3.6 for both the weighted (left) and unweighted (right) cases. Our bootstrapped power spectrum outputs are shown as black points and the colored heat-map overlaid on top is the likelihood function modified by our prior. Although we only show figures for one k-value, we note that the shape of the transfer curve is similar for all k's. We then invoke Bayes' interpretation and re-interpret it as the posterior $p(P_{\rm in}|\widehat{P}_{\rm out})$ where we recall that $\widehat{P}_{\rm out}$ represents a (lossy) power spectrum. To do this we make a horizontal cut across at the data value $\widehat{\mathbf{P}}_x$ (setting $\widehat{P}_{\rm out} = \widehat{\mathbf{P}}_x$), shown by the gray solid line, to yield a posterior distribution for the signal. We normalize this final distribution and compute the 95% confidence interval (an upper limit on EoR).

By-eye inspection of the transfer function in Figure 3.6 gives a sense of what the signal loss result should be. The power spectrum value of our data, $\widehat{\mathbf{P}}_x$ is marked by the solid gray horizontal lines. From the left plot (empirically estimated inverse covariance weighting), one can eyeball that a data value of $10^5\,{\rm mK}^2\ (h^{-1}\ {\rm Mpc})^3$, for example, would map approximately to an upper limit of $\sim 10^9\,{\rm mK}^2\ (h^{-1}\ {\rm Mpc})^3$, implying a signal loss factor of $\sim 10^4$.

The loss-corrected power spectrum limit for empirically estimated inverse covariance weighted PAPER-64 data is shown in Figure 3.7 (solid red), which we can compare to the original lossy result (dashed red). Post-signal loss estimation, the power spectrum limits are higher than both the theoretical noise level (green) and uniform-weighted power spectrum (which is shown three ways: black and gray points are positive and negative power spectrum values, respectively, with 2σ error bars from bootstrapping, the solid blue is the upper limit on the EoR signal using the full signal injection framework, and the shaded gray is the power spectrum values with thermal noise errors). We elaborate on this point in the next section, as well as investigate alternate weighting schemes to inverse covariance weighting, with the goal of finding one that balances the aggressiveness of down-weighting contaminants and minimizing the loss of the EoR signal.

3.2.3 Minimizing Signal Loss

With a signal loss formalism established, we now have the capability of experimenting with different weighting options for $\mathbf{R}$. Our goal here is to choose a weighting method that successfully down-weights foregrounds and systematics in our data without generating large amounts of signal loss as we have seen with the inverse covariance estimator. We have found that the balance between the two is a delicate one and requires a careful understanding and altering of empirical covariances.

We saw in Chapter 2.2.4 how limiting the number of down-weighted eigenmodes (i.e., flattening out part of the eigenspectrum and effectively decoupling the lowest-valued eigenmodes, which are typically EoR-dominated, from the data) can help minimize signal loss. We experiment with this idea on PAPER-64 data, dialing the number of modes that are down-weighted from zero (which is equivalent to identity-weighting, or the uniform-weighted case) to 21 (which is the full inverse covariance estimator). The power spectrum results

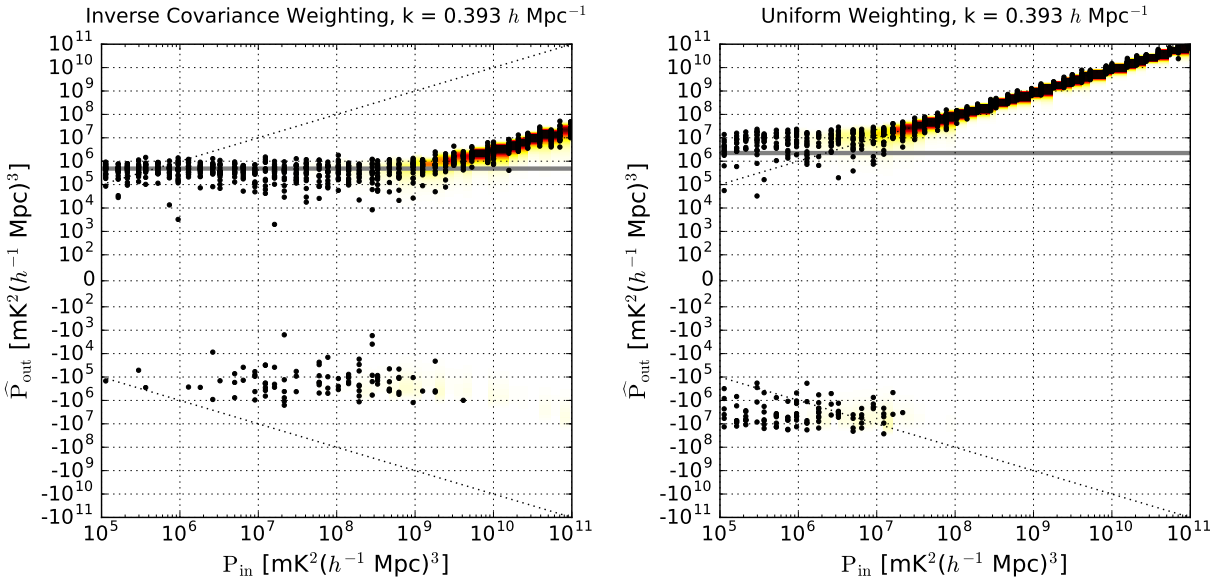

Figure 3.6: Signal loss transfer functions showing the relationship of P_{in} and $\widehat{P}_{\mathrm{out}}$, as defined by Equations (3.4) and (3.5). Power spectra values (black points) are generated for 20 realizations of $\mathbf{e}$ per signal injection level. Since our $\widehat{P}_{\mathrm{out}}$ values are well-approximated by a Gaussian distribution, we fit Gaussians to each injection level based on the mean and variance of the simulation outputs. This entire likelihood function is then multiplied by a Jeffreys prior for $p(P_{\mathrm{in}})$, with the final result shown as the colored heat-maps on top of the points. Two cases are displayed: empirically estimated inverse covariance weighted PAPER-64 data (left) and uniform-weighted data (right). The dotted black diagonal lines mark a perfect unity mapping, and the solid gray horizontal line denotes the power spectrum value of the data $\widehat{\mathbf{P}}_x$, from which a posterior distribution for the signal is extracted. From these plots, it is clear that the weighted case results in ~ 4 orders of magnitude of signal loss at the data-only power spectrum value, whereas the uniform-weighted case does not exhibit loss. The general shape of these transfer functions are also shown by the black curves in Figure 3.3 for comparison.

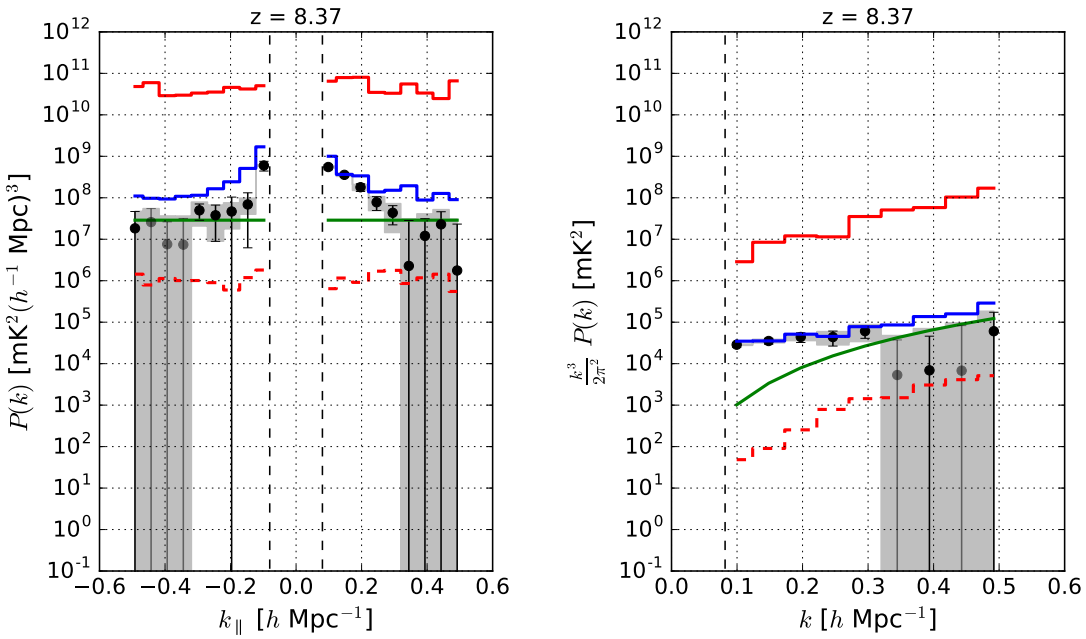

Figure 3.7: A power spectrum of a subset of PAPER-64 data illustrating the use of empirical inverse covariance weighting. The solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework using empirical inverse covariance weighting. Shown for comparison is the lossy limit prior to signal loss estimation (dashed red). The theoretical 2σ thermal noise level prediction based on observational parameters is in green, whose calculation is detailed in Chapter 3.3.2. Additionally, the power spectrum result for the uniform weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), the 2σ upper limit on the EoR signal using our full signal injection framework (solid blue), and the measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. In this example, we see that the lossy power spectrum limit is ~ 4 orders of magnitude too low when using empirical inverse covariance weighting.

for one k-value, both before and after signal loss estimation, are shown in the top panel in Figure 3.8. We see that the amount of signal loss increases as weighting becomes more aggressive (dashed red). In other words, more EoR-dominated fluctuations are being overfit and subtracted as more modes are down-weighted. We also find that the power spectrum upper limit, post-signal loss estimation, increases with the number of down-weighted modes (solid red). The more modes we use in down-weighting, the stronger the coupling between the weighting and the data, and the greater the error we have in estimating the power spectrum. Switzer et al. (2013) took a similar approach in determining the optimal number of modes to down-weight in GBT data, finding similar trends and noting that removing too few modes is limited by residual foregrounds and removing too many modes is limited by large error bars and signal loss.

Optimistically, we expect there to be a "sweet spot" as we dial our regularization knob; a level of regularization where weighting is beneficial compared to uniform weighting (blue). In other words, we would like a weighting scheme that down-weights eigenmodes that predominantly describe foreground modes, but not EoR modes. We see in Figure 3.8 that this occurs roughly when only the ~ 3 highest-valued eigenmodes are down-weighted and the rest are given equal weights (though for the case shown, weighting only slightly outperforms uniform weighting). For a similar discussion on projecting out modes (zeroing out eigenmodes, rather than just ignoring their relative weightings as we do in this study), see Switzer et al. (2013).

We also saw in Chapter 2.2.4 how adding the identity matrix to the empirical covariance can minimize signal loss. We experiment with this idea as well, shown in the bottom panel of Figure 3.8. The dashed red and solid red lines represent power spectrum limits pre and post-signal loss estimation, respectively, as a function of the strength of $\mathbf{I}$ that is added to $\widehat{\mathbf{C}}$, quantified as a percentage of $\mathrm{Tr}(\widehat{\mathbf{C}})\mathbf{I}$ added to $\widehat{\mathbf{C}}$. We parameterize this "regularization strength" parameter as γ, namely $\widehat{\mathbf{C}} \equiv \widehat{\mathbf{C}} + \gamma \mathrm{Tr}(\widehat{\mathbf{C}})\mathbf{I}$. From this plot we see that only a small percentage of $\mathrm{Tr}(\widehat{\mathbf{C}})$ is needed to significantly reduce loss. We expect that as the strength of $\mathbf{I}$ is increased (going to the left), both the red curves will approach the uniform-weighted case. We also notice that the post-signal loss limit hovers around the uniform-weighted limit for a large range of regularization strengths and while an overall trend from high-to-low signal loss is seen as the strength increases, there does not appear to be a clear "minimum" that produces the least loss.

In addition to our thermal noise prediction (green) and uniform-weighted power spectrum limit (blue), one additional horizontal line is shown in Figure 3.8 in both panels and represents a third regularization technique. This line (black) denotes the power spectrum value, post-signal loss estimation, for inverse variance weighting (multiplying an identity matrix element-wise to $\widehat{\mathbf{C}}$). This result is single-valued and not a function of the horizontal axis. We see that all three regularization schemes shown (solid red top panel, solid red bottom panel, black) perform similarly at their best (i.e., when ~ 3 eigenmodes are down-weighted in the case of the top panel's solid red curve). However, for the remainder of this chapter, we choose to use the weighting option of $\widehat{\mathbf{C}} + 0.09\,\mathrm{Tr}(\widehat{\mathbf{C}})\mathbf{I}$, or $\gamma = 0.09$, which we will denote as $\widehat{\mathbf{C}}_{\mathrm{eff}}$. We choose

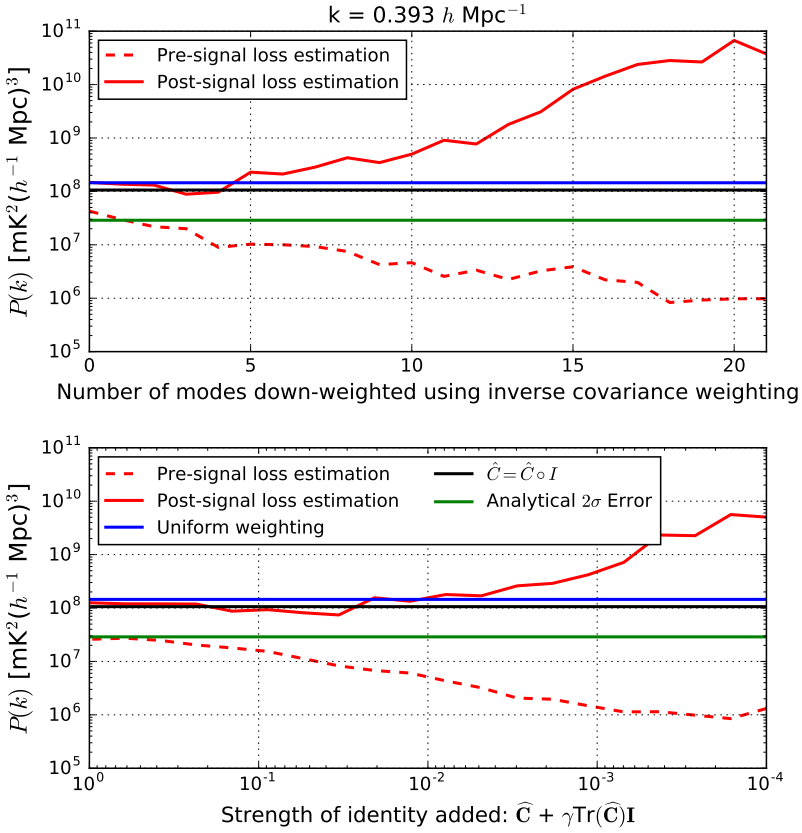

Figure 3.8: Power spectra 2σ upper limits for $k = 0.393\,h$ Mpc^{-1} for fringe-rate filtered PAPER-64 data. Top: Values are shown before (dashed red) and after (solid red) signal loss estimation via our signal injection framework as a function of number of eigenmodes of $\widehat{\mathbf{C}}$ that are down-weighted. This regularization knob is tuned from 0 modes on the left (i.e., unweighted) to 21 modes on the right (i.e., the full inverse covariance estimator). ~ 4 orders of magnitude of signal loss results when using empirically estimated inverse covariance weighting. Bottom: Power spectrum upper limits before (dashed red) and after (solid red) signal loss estimation as a function of identity added to the empirical covariance. This regularization knob is tuned from $\gamma = 10^{-4}$ on the right (i.e., very little regularization) to $\gamma = 1$ on the left (see main text for the definition of γ). Also plotted in both panels for comparison are 2σ power spectrum upper limits for the uniform-weighted case (blue) and inverse variance weighted case (black); both are after signal loss estimation. Finally, a theoretical prediction for noise (2σ error) is plotted as green. In the PAPER-64 analysis in this chapter, we choose to use a regularization scheme of $\widehat{\mathbf{C}}_{\mathrm{eff}} \equiv 0.09\,\mathrm{Tr}(\widehat{\mathbf{C}})\mathbf{I} + \widehat{\mathbf{C}}$ ($\gamma = 0.09$) as a simple example of regularization that minimizes loss, and note that the power spectrum limits using this type of regularization are roughly constant across a large range of values of γ.

this weighting scheme merely as a simple example of regularizing PAPER-64 covariances, noting that the power spectrum upper limit remains roughly constant for a broad range of values of γ.

It is important to note that our signal injection methodology for assessing loss makes the assumption that we know the true signal's strength and structure. Realistically, these details about the EoR signal are unknown and our signal loss framework is limited by our simulations. Therefore, while this chapter employs this methodology as an example of one way of estimating loss, Chapter 4 uses uniform weightings in order to produce more trustworthy, straightforward power spectrum limits that do not suffer from loss.

The power spectrum result for our subset of PAPER-64 data (using only one baseline separation type, 10 baselines, and $\widehat{\mathbf{C}}_{\text{eff}}$) using the analysis presented in this chapter is shown in Figure 3.9. Again, the solid red curve represents our upper limit on the EoR signal using the full signal injection framework. The uniform weighted case is shown as the black and gray points, which correspond to positive and negative power spectrum values respectively (with 2σ errors bars from bootstrapping). It is also shown as an upper limit using the signal injection framework (solid blue), which is interestingly larger than the errors computed from bootstrapping, likely because the full injection framework takes into account additional sample variance whereas the bootstrapped errors do not. Finally, the gray shaded regions combine the measured uniform weighted power spectrum values with thermal noise errors. We show this power spectrum result as one example of how a simple regularization of an empirical covariance matrix can minimize signal loss, though we also note that this weighting does not produce more stringent limits than the uniform weighted case, thus further motivating uniform-weighting for our revised results in Chapter 4.

In this section we have shown three simple ways of regularizing $\widehat{\mathbf{C}}$ to minimize signal loss using PAPER-64 data. There are many other weighting schemes that we leave for consideration in future work. For example, one could estimate $\widehat{\mathbf{C}}$ using information from different subsets of baselines. For redundant arrays this might mean calculating $\widehat{\mathbf{C}}$ from a different but similar baseline type, such as the ~ 30 m diagonal PAPER baselines (instead of the horizontal E/W ones). Alternatively, covariances could be estimated from all baselines except the two being cross-multiplied when forming a power spectrum estimate. This method was used in Parsons et al. (2014) (a similar method was also used in Dillon et al. (2015a)) in order to avoid suppressing the 21 cm signal, and it is worth noting that the PAPER-32 results are likely less impacted from the issue of signal loss underestimation because of this very reason (however, they are affected by the error estimation issues described in Chapter 3.3.2, so we also regard those results as suspect and superseded by those in Chapter 4).

Another possible way to regularize $\widehat{\mathbf{C}}$ is to use information from different ranges of LST. For example, one could calculate $\widehat{\mathbf{C}}$ with data from LSTs where foregrounds are stronger (earlier or later LSTs than the "foreground-quiet" range typically used in forming power spectra) — doing so may yield a better description of the foregrounds that we desire to down-weight, especially if residual foreground chromaticity is instrumental in origin and stable in time. Fundamentally, each of these examples are similar in that they rely on a computation of $\widehat{\mathbf{C}}$ from data that is similar but not exactly the same as the data that is

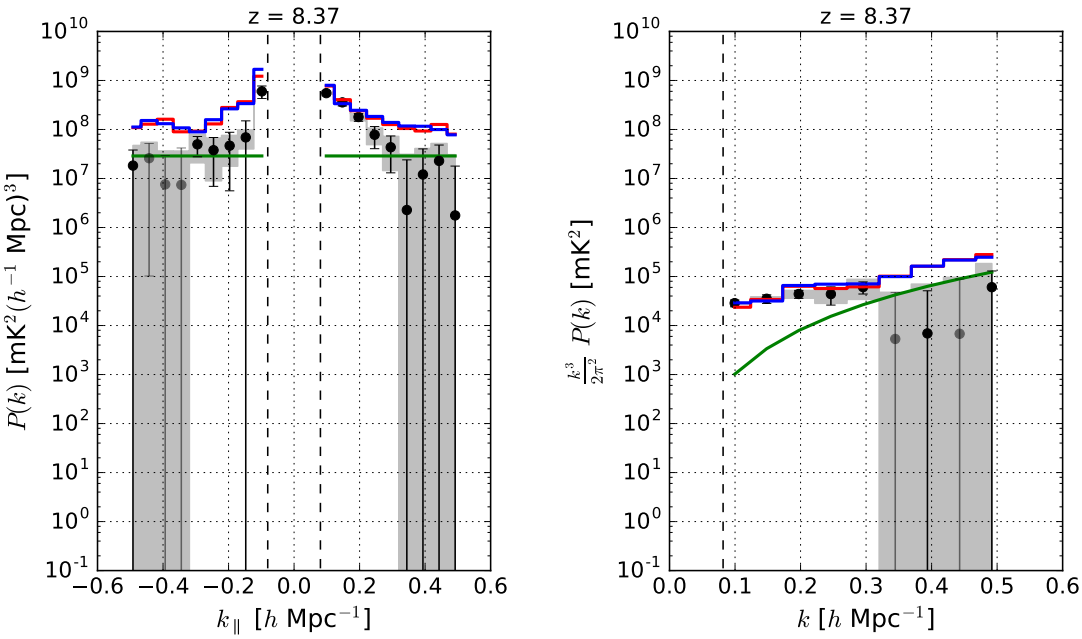

Figure 3.9: A power spectrum of a subset of PAPER-64 data illustrating the use of $\widehat{\mathbf{C}}_{\text{eff}}$ to minimize signal loss. The solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework. The theoretical 2σ thermal noise level prediction based on observational parameters is in green. Additionally, the power spectrum result for the uniform weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), the 2σ upper limit on the EoR signal using our full signal injection framework (solid blue), and the measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. This power spectrum result does not use the full data set's sensitivity as in Ali et al. (2015) and Chapter 4, though we include all analysis changes which have mostly stemmed from revisions regarding signal loss, bootstrapping, and the theoretical error computation. We see that the regularization scheme used here produces limits similar to the unweighted limits.

being down-weighted. Ideally this would be effective in down-weighting shared contaminants yet avoid signal loss from overfitting EoR modes in the power spectrum data set itself.

In Chapter 3.2, we have detailed several aspects of signal loss in PAPER-64: how the loss arises, how it can be estimated from an injection framework, and ways it can be minimized. We again emphasize that these lessons learned about signal loss are largely responsible for shaping our revised analysis of PAPER data. In the remainder of this chapter, we will transition to other aspects of our analysis that have been revised since A15.

3.3 Error Estimation

Underestimated signal loss is the main reason for the revision of the power spectrum limits from A15. It is interesting to note that — had all the other aspects of the original analysis been correct — the underestimated limits may have been more easily caught. Unfortunately, two related power spectrum components, namely the error bars on the power spectrum data points and the theoretical noise prediction, were also calculated incorrectly.

In this section, we summarize multiple inconsistencies and errors that have been found since the previous analysis in terms of error estimation. We first describe updated methods regarding bootstrapping, which determines the error bars on our limits. We then highlight an updated calculation for the theoretical noise sensitivity of PAPER-64 and illustrate how our revised calculation has been verified through simulations.

3.3.1 Bootstrapping

In Chapter 2.4 we described the method of bootstrapping, which is used to determine confidence intervals on our power spectrum measurements. We highlighted one caveat of bootstrapping that arises when working with correlated data and showed how errors can be underestimated when data is correlated along the bootstrapping axis.

This is the precisely how errors were underestimated in PAPER-64. Because of fringe-rate filtering, which averages data in time to increase sensitivity, PAPER-64 data is correlated along the time axis. Hence, there are fewer independent samples after filtering, thus decreasing the variance of the bootstraps.

More specifically, the PAPER-64 pipeline outputs 20 bootstraps (over baselines), each a 2-dimensional power spectrum that is a function of k and time. In A15, a second round of bootstrapping occurred over the time axis, and a total of 400 bootstraps were created in this step, each comprised of randomly selected values sampled with replacement (i.e., each of these bootstraps contained the same number of values as the number of time integrations, which, at ~ 700, greatly exceeds the approximate number of independent samples after fringe-rate filtering). Means were then taken of the values in each bootstrap. Finally, power spectrum limits were computed by taking the mean and standard deviation over all the bootstraps. We emphasize again that in this previous analysis, the number of elements sampled per bootstrap greatly exceeded the number of independent LST samples, underestimating errors. A random

draw of 700 measurements from this data set has many repeated values, and the variance between hundreds of these random samples is smaller than the true underlying variance of the data.

Given our new understanding of the sensitivity of bootstraps to the number of elements sampled, we have removed the second bootstrapping step along time entirely and now simply bootstrap over the baseline axis. Power spectrum 2σ errors (computed from bootstrap variances) with and without this bootstrapping change for a fringe-rate filtered noise simulation are shown in Figure 3.10 in black and gray, respectively. The estimates are uniformly weighted in order to disentangle the effects of bootstrapping from signal loss. As shown in the figure, when more elements are drawn for each bootstrap than the number of independent samples (by over-sampling elements along the time axis), repeated values begin to crop up and the apparent variation between bootstraps drops, resulting in limits (gray) below the predicted noise level (green). Using the revised bootstrapping method, where bootstrapping only occurs over the baseline axis, the limits (black) are shown to agree with the analytic prediction for noise. While Figure 3.10 implies that errors, computed prior to our bootstrapping change (gray), are underestimated by a factor of ~ 5 in mK^2 for the noise simulation (whose creation details are outlined in the next section), in practice this factor is lower for the case of real data (a factor of ~ 3 in mK^2 instead), possibly due to the data being less correlated in time than the fringe-rate filtered noise in the simulation.

In addition to learning how sample independence affects bootstrapped errors, we have made three additional changes to our bootstrapping procedure since A15, summarized here:

- A second change to our bootstrapping procedure is that we now bootstrap over baseline cross-products, instead of the baselines themselves. In the previous analysis, baselines were bootstrapped prior to forming cross power spectra, and using this particular ordering of operations (bootstrapping, then cross-multiplication) yields variances that have been found to disagree with predicted errors from bootstrapping using simulations. On the contrary, bootstrapping over cross power spectra ensures that we are estimating the variance of our quantity of interest (i.e., the power spectrum). This change, while fundamental in retaining the integrity of the bootstrapping method in general, alters the resulting power spectrum errors by factors of < 2 in practice.

- In A15, individual baselines were divided into five independent groups, where no baselines were repeated in each group. Then, baselines within each group were averaged together, and the groups were cross-multiplied to form power spectra. This grouping method was used to reduce computational time, however upon closer examination it has been found that the initial grouping introduces an element of randomness into the final measurements — more specifically, the power spectrum value fluctuates depending on how baselines are assigned into their initial groups. Our new approach removes this element of randomness at the cost of computational expense, as we now perform all baseline cross-products.

- Finally, the last change from the A15 method is that our power spectrum points (pre-

viously computed as the mean of all bootstraps), are now computed as the power spectrum estimate resulting from not bootstrapping at all. More specifically, we compute one estimate without sampling, and this estimate is propagated through our signal loss computation (this estimate is $\widehat{\mathbf{P}}_x$). The difference between taking the mean of the bootstrapped values and using the estimate from the no-bootstrapping case is small, but doing the latter ensures that we are forming results that reflect the estimate preferred by all our data.

In summary, we have learned several lessons regarding bootstrapping and have revised our analysis procedure in order to determine error bars that correctly reflect the variance in our power spectrum estimates. Bootstrapping can be an effective and straightforward way to estimate errors of a data set, however, bootstrapping as a means of estimating power spectrum errors from real fringe-rate filtered data requires knowledge of the number of independent samples, which is not always a trivial task. We have thus avoided this issue by removing one of our bootstrap axes, as well as updated several other details of our procedure to ensure accurate re-sampling and error estimation.

3.3.2 Theoretical Error Estimation

One useful way of cross-checking measured power spectrum values and errors is to compute a theoretical estimation of thermal noise based on observational parameters. Although a theoretical model often differs from true errors, it is helpful to understand the ideal case and the factors that affect its sensitivity. Upon re-analysis of PAPER-64, we have discovered that this estimate was also underestimated in previous analyses.

To compute our theoretical noise estimate, we use an analytic sensitivity calculation. Through detailed studies using several independently generated noise simulations, what we found was that our simulations all agreed but were discrepant with the previous calculations. The analytic calculation is only an approximation and attempts to combine a large number of pieces of information in an approximate way; however, when re-considering some of the approximations, the differences were large enough (factors of 10 in some cases) to warrant a careful investigation. What follows here is an accounting of the differences which have been discovered. We note that our theoretical error estimate, which is plotted as the solid green curve in many of the previous power spectrum plots in this thesis, is computed with these changes accounted for.

The noise prediction $n(k)$ (Parsons et al. 2012; Pober et al. 2013) for a power spectral analysis of interferometric 21 cm data, in temperature-units, is:

$$N(k) = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{\sqrt{2 N_{\rm lst} N_{\rm seps}}\, t_{\rm int} N_{\rm days} N_{\rm bls} N_{\rm pols}}. \tag{3.15}$$

We will now explain each factor in Equation (3.15) and highlight key differences from the numbers used in A15.

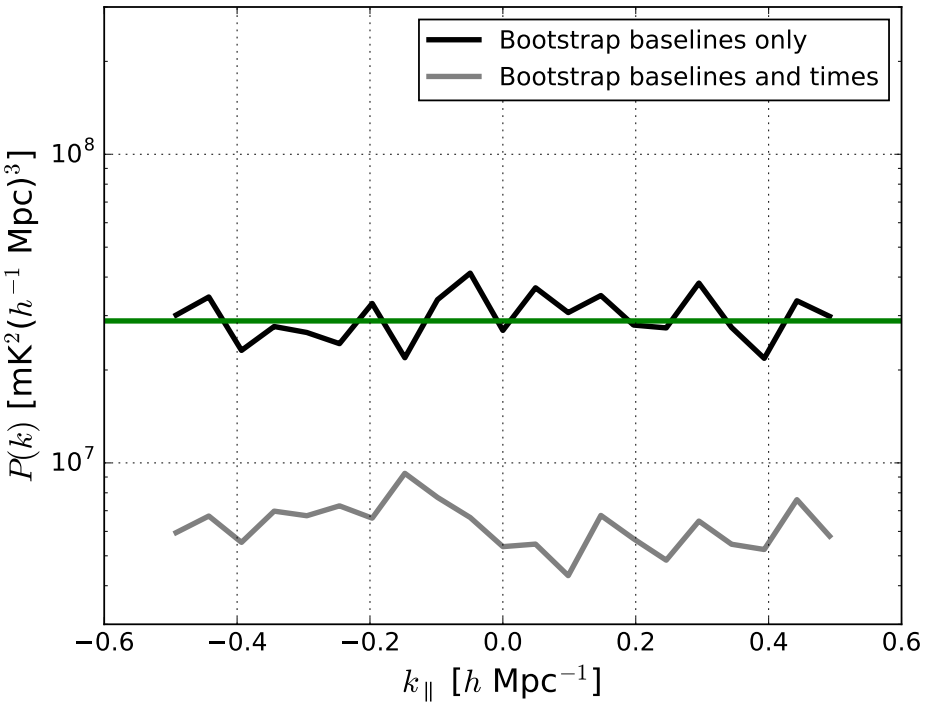

Figure 3.10: 2σ power spectrum errors (from bootstrap variances) for a noise simulation (computed via Equation (3.20) using PAPER-64 observing parameters) using two different bootstrapping methods. The noise is fringe-rate filtered and a weighting matrix of $\mathbf{I}$ (uniform-weighted) is used in order to disentangle the effects of bootstrapping from signal loss. The bootstrapping method used in Ali et al. (2015) is shown in gray, where bootstrapping occurs along both the baseline and time axes. This underestimates errors by sampling more values than independent ones in the data set (fringe-rate filtering reduces the number of independent samples along time). We use the method illustrated by the black curve in our updated analysis, where bootstrapping only occurs along the baseline axis. We find that these revised limits agree with the 2σ analytic prediction for noise (green).

- X^2Y: Conversion factors from observing coordinates (angles on the sky and frequency) to cosmological coordinates (co-moving distances). For $z = 8.4$, $X^2Y = 5 \times 10^{11}\, h^{-3}$ Mpc^3 str^{-1} GHz^{-1}.

- $\Omega_{\rm eff}$: The effective primary beam area in steradians (Parsons et al. 2010; Pober et al. 2012). The effective beam area changes with the application of a fringe-rate filter, since different parts of the beam are up-weighted and down-weighted. Using numbers from Table 1 in Parsons et al. (2016), $\Omega_{\rm eff} = 0.74^2/0.24$ for an optimal fringe-rate filter and the PAPER primary beam.

- $T_{\rm sys}$: The system temperature is set by:

$$T_{\rm sys} = 180\left(\frac{\nu}{0.18}\right)^{-2.55} + T_{\rm rcvr}, \tag{3.16}$$

  where ν are frequencies in GHz (Thompson et al. 2001). We use a receiver temperature of 144 K, yielding $T_{\rm sys} = 431$ K at 150 MHz. This is lower than the $T_{\rm sys}$ of 500 K used in A15 because of several small miscalculation errors that were identified[2].

- $\sqrt{2}$: This factor in the denominator of the sensitivity equation comes from taking the real part of the power spectrum estimates after cross-multiplying two independent visibility measurements. In A15, a factor of 2 was mistakenly used.

- $N_{\rm lst}$: The number of independent LST bins that go into a power spectrum estimation. The sensitivity scales as the square root because we integrate incoherently over time. For PAPER-64, $N_{\rm lst} = 8$.

- $N_{\rm seps}$: The number of baseline separation types (where baselines of a unique separation type have the same orientation and length) averaged incoherently in a final power spectrum estimate. For the analysis in this chapter, we only use one type of baseline (PAPER's 30 m East/West baselines). However, both the updated limits in Chapter 4 and the sensitivity prediction in Figure 3.11 use three separation types ($N_{\rm seps} = 3$) to match A15.

- $t_{\rm int}$: Length of an independent integration of the data. It is crucial to adapt this number if filtering is applied along the time axis (i.e., a fringe-rate filter). We compute the effective integration time of our fringe-rate filtered data by scaling the original integration time t_i using the following:

$$t_{\rm int} = t_i \frac{\int 1\, df}{\int w^2(f)\, df}, \tag{3.17}$$

  where $t_i = 43$ seconds, $t_{\rm int}$ is the fringe-rate filtered integration time, w is the fringe-rate profile, and the integral is taken over all fringe-rates. For PAPER-64, this number is $t_{\rm int} = 3857$ s.

[2]For example, there was a missing a square root in going from a variance to a standard deviation.

- $N_{\rm days}$: The total number of days of data analyzed. In A15, this number was set to 135. However, because we divide our data in half (to form "even" and "odd" data sets, or $N_{\rm datasets} = 2$), this number should reflect the number of days in each individual data set instead of the total. Additionally, this number should be adjusted to reflect the actual number of cross-multiplications that occur between data sets ("even" with "odd" and "odd" with "even", but not "odd" with "odd" or "even" with "even" in order to avoid noise biases). Finally, because our LST coverage is not 100% complete (it doesn't overlap for every single day), we incorporate a root-mean-square statistic in computing a realistic value of $N_{\rm days}$. Our expression therefore becomes:

$$N_{\rm days} = \sqrt{\langle N_i^2 \rangle} \sqrt{(N_{\rm datasets}^2 - N_{\rm datasets})} \tag{3.18}$$

  where i indexes LST and frequency channel over all data sets (Jacobs et al. 2015). For PAPER-64, our revised estimate of $N_{\rm days}$ is ~ 47 days.

- $N_{\rm bls}$: The number of baselines contributing to the sensitivity of a power spectrum estimate. In A15, this number was the total number of 30 m East/West baselines used in the analysis. However, using the total number of baselines ($N_{\rm bls\_total} = 51$) neglects the fact that the A15 analysis averages baselines into groups for computational speed-up when cross-multiplying data. Our revised estimate for the parameter is:

$$N_{\rm bls} = \frac{N_{\rm bls\_total}}{N_{\rm gps}} \sqrt{\frac{N_{\rm gps}^2 - N_{\rm gps}}{2}}, \tag{3.19}$$

  where, in the A15 analysis, $N_{\rm gps} = 5$. Each baseline group averages down linearly as the number of baselines entering the group ($N_{\rm bls\_total}/N_{\rm gps}$) and then as the square root of the number of cross-multiplied pairs $\left(\sqrt{\frac{N_{\rm gps}^2 - N_{\rm gps}}{2}}\right)$. A revised A15 analysis should therefore use $N_{\rm bls} \sim 32$ instead of 51, and this change is taken into account in Figure 3.11. However, the analysis throughout this thesis no longer averages baselines into groups ($N_{\rm gps} = 1$). For the subset of data presented in this chapter, $N_{\rm bls} = 10$.

- $N_{\rm pols}$: The number of polarizations averaged together. For the case of Stokes I, $N_{\rm pols} = 2$.

An additional factor of $\sqrt{2}$ is gained in sensitivity when folding together positive and negative k's to form $\Delta^2(k)$.

Our revised sensitivity estimate for the A15 analysis of PAPER-64 is shown in Figure 3.11. Together, the revised parameters yield a decrease in sensitivity (higher noise floor) by a factor of ~ 7 in mK^2.

To verify our thermal noise prediction, we form power spectra estimates using a pure noise simulation. We create Gaussian random noise assuming a constant $T_{\rm rcvr}$ (translated into $T_{\rm sys}$ via Equation (3.16)) but accounting for the true $N_{\rm days}$ as determined by LST

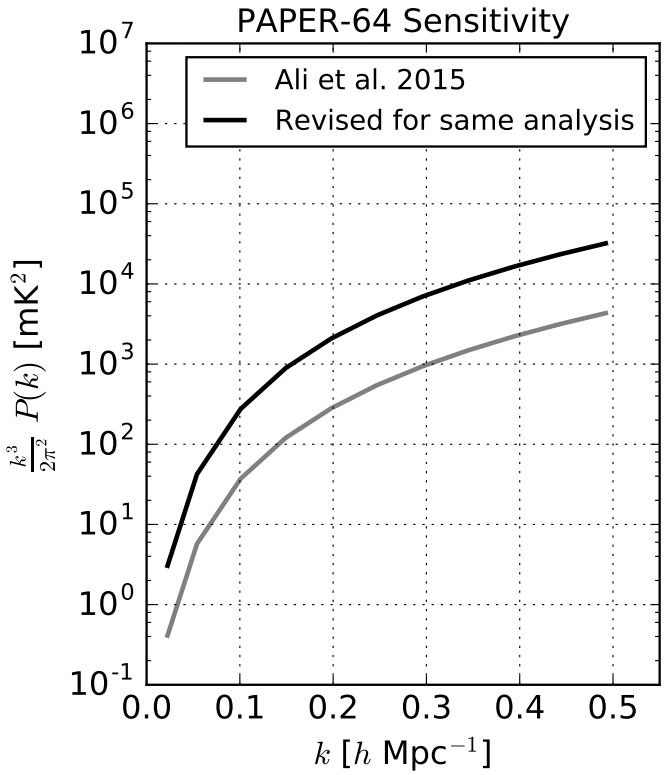

Figure 3.11: An updated prediction for the thermal noise level of PAPER-64 data (black) is shown in comparison to previously published sensitivity limits (gray), both computed for the parameters and methods used in Ali et al. (2015). Major factors that contribute to the discrepancy are $\Omega_{\rm eff}$, $N_{\rm days}$ and $N_{\rm bls}$, as in Equation (3.15) and described in Chapter 3.3.2, which when combined decreases our sensitivity (higher noise floor) by a factor of ~ 7 in mK^2.

sampling counts for each time and frequency in the LST-binned data. We convert $T_{\rm sys}$ into a root-mean-square variance statistic using:

$$T_{\rm rms} = \frac{T_{\rm sys}}{\sqrt{\Delta\nu \Delta t N_{\rm days} N_{\rm pols}}}, \tag{3.20}$$

where $\Delta\nu$ is the channel spacing, Δt is the integration time, $N_{\rm days}$ is the number of daily counts for a particular time and frequency that went into our LST-binned set, and $N_{\rm pols}$ is the number of polarizations (2 for Stokes I). This temperature sets the variance of the Gaussian random noise.

Power spectrum results for the noise simulation, which uses our full power spectrum pipeline, are shown in Figure 3.12. We highlight that the bootstrapped data (black and gray points, with 2σ error bars) and thermal noise prediction (solid green) show good agreement, as bootstrapping provides an accurate estimate of the noise variance. However, the limits from the full signal loss framework (weighted and unweighted in red and blue, respectively) are inflated, likely due to the additional inclusion of sample variance that comes from the EoR simulations. While the noise simulation provides an important indicator about the accuracy of our theoretical noise calculation, we note that the calculation did not take into account additional sources of error associated with earlier analysis steps (for example, Trott & Wayth (2017) show how calibration specifically can add errors to visibilities). Additionally, we recommend that future work investigate possible error correlations between baseline pairs and any interaction effects between signal and noise that may effect error calculations. Because of these reasons, we therefore interpret our noise prediction as the sensitivity floor for our measurements.

3.4 Bias

In Chapter 2.5 we highlighted some common sources of bias that can show up as power spectrum detections and imitate an EoR signal. We discussed the importance of using jackknife and null tests for instilling confidence in an EoR detection, as well as for identifying other sources of biases. Here we demonstrate methods used by PAPER-64 to mitigate foreground and noise bias and we perform null tests in order to characterize the stability and implications of our results.

3.4.1 Mitigating Bias

We briefly discuss one way we mitigate foreground leakage in a power spectrum estimate, and two ways we suppress noise biases. These methods are not novel to this analysis but here we frame them in the context of minimizing false (non-EoR) detections.

Tailoring window functions is one way to suppress foreground biases (similar discussions to the following one are in Liu et al. (2014b) and A15). As alluded to in Chapter 2.2, we have a choice for the normalization matrix $\mathbf{M}$ in Equation (2.9). For the analysis of PAPER-64

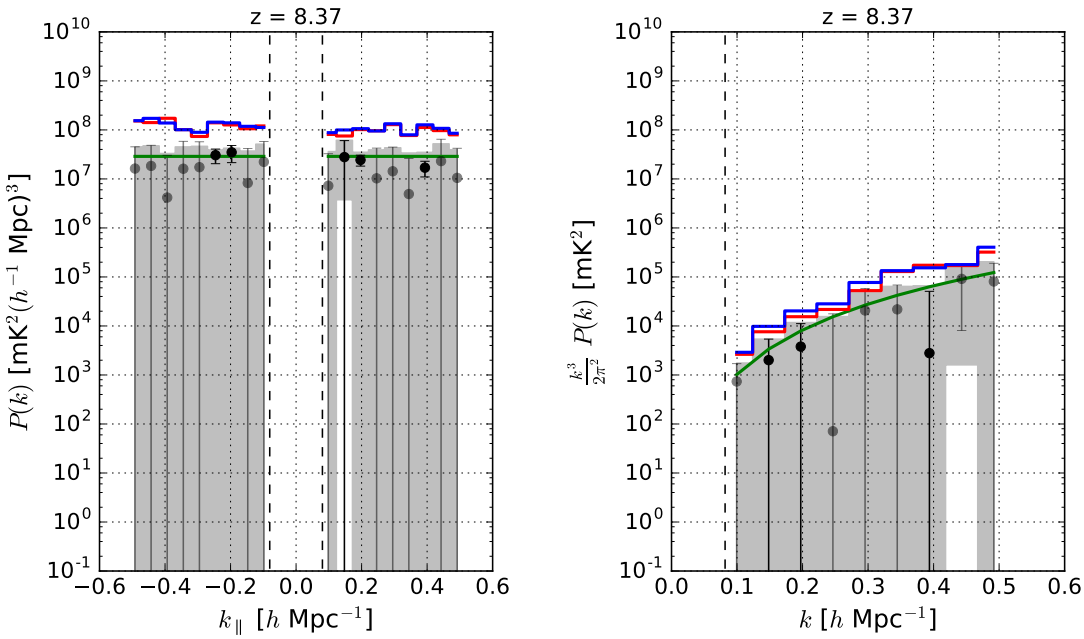

Figure 3.12: The power spectrum for a noise simulation that mimics the noise level of a subset of PAPER-64 data, where the solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework using $\widehat{\mathbf{C}}_{\text{eff}}$. The theoretical 2σ thermal noise level prediction based on observational parameters (calculated by Equation (3.15)) is in green. Additionally, the power spectrum result for the uniform weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), the 2σ upper limit on the EoR signal using our full signal injection framework (solid blue), and the measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. We highlight that the bootstrapped data points and thermal noise prediction show good agreement, while the limits from the full injection framework (red and blue) are inflated due to the additional inclusion of sample variance that comes from the injection simulations.

data, we compute $\mathbf{M}$ using the matrix $\mathbf{G}$ (which would be the Fisher matrix if $\mathbf{R} \equiv \mathbf{C}^{-1}$), defined as:

$$\mathbf{G}^{\alpha\beta} = \frac{1}{2}\mathrm{tr}[\mathbf{R}\mathbf{Q}^{\alpha}\mathbf{R}\mathbf{Q}^{\beta}] \tag{3.21}$$

where $\mathbf{R}$ is the data-weighting matrix and α and β are wavebands in $k_{\parallel}$. We take the Cholesky decomposition of $\mathbf{G}$, decomposing it into two lower triangular matrices (which is possible since $\mathbf{G}$ is Hermitian):

$$\mathbf{G} = \mathbf{L}\mathbf{L}^{\dagger}. \tag{3.22}$$

Next, we construct $\mathbf{M}$:

$$\mathbf{M} = \mathbf{D}\mathbf{L}^{-1} \tag{3.23}$$

where $\mathbf{D}$ is a diagonal matrix. In doing so, our window function, defined as $\mathbf{W} = \mathbf{M}\mathbf{G}$ (see Equation (2.10)), becomes

$$\mathbf{W} = \mathbf{D}\mathbf{L}^{\dagger}. \tag{3.24}$$

Because of the nature of the lower triangular matrix, this window function has the property of preventing the leakage of foreground power from low-k to high-k modes. Specifically, we order the elements in $\mathbf{G}$ in such a way so that power can leak from high-k modes to low-k modes, but not vice versa. Since most foreground power shows up at low-k's, this method ensures a window function that retains clean, noise-dominated measurements while minimizing the contamination of foreground bias. This tailored window function was used in the A15 analysis, however throughout this chapter, we use a diagonal $\mathbf{M}$ for simplicity.

In addition to mitigating foreground bias at high-k's, two other sources of bias that we actively suppress in the PAPER-64 analysis are noise bias associated with the squaring of thermal noise and noise bias from crosstalk. In order to avoid the former, we filter out certain cross-multiplications when forming $\widehat{q}$ in Equation (2.8). Namely, the PAPER-64 data set is divided into two halves: even Julian dates and odd Julian dates. Our data vectors are then $\mathbf{x}_{even,1}$ for the "even" data set and baseline 1, $\mathbf{x}_{odd,1}$ for the "odd" data set and baseline 1, etc. We only form $\widehat{q}$ when the two copies of $\mathbf{x}$ come from different groups and baselines, never multiplying "baseline 1" with "baseline 1", for example, in order to prevent the squaring of the same thermal noise.

To mitigate crosstalk bias, which appears as a static bias in time, we apply a fringe-rate filter that suppresses fringe-rates of zero. Figure 3.2 shows that the filter response is zero for such static signals. The effect of filtering out zero fringe-rates on power spectrum results is shown in A15. Most notably, even without accounting for signal loss, the crosstalk bias at all k's is very strong compared to the removed case.

3.4.2 Jackknife and Null Tests

As shown in Figure 3.9, our illustrative PAPER-64 power spectrum shows biases above the predicted noise level, particularly at low-k values. As discussed in Chapter 2.5.1, this bias is most likely attributable to foreground leakage.

Here we perform three null tests on PAPER-64 data that aim to isolate systematics in the data and verify that our biases are not attributable to EoR. Similar to in Chapter 2.5.2, we take jackknives along different axes of the data set to produce multiple power spectra. We then difference them (i.e., the null test) to tease out excess variances.

The three results are shown in Figure 3.13. Each test displays the differenced power spectrum between two halves of a jackknife, where the plotted points are the differenced power spectrum values, and the plotted errors are the bootstrapped errors of the two data set halves added in quadrature. The expected thermal noise level (gray shaded regions) is the thermal noise of each data set added in quadrature as well. Constructing the tests as such ensures that we are probing whether the variances of each data set differ by an amount consistent with the thermal noise. We use uniform weightings for all tests.

We take jackknives along three different axes:

- Baselines: We split our data set into two halves, where each contains half of the total baselines used in the analysis. No baselines are repeated between the two data sets.
- Sidereal Hour: We split our data set into two halves based on LST, namely the first half (LSTs 0.5-4.5 hours) and second half (LSTs 4.5-8.6 hours).
- Day: We split our data set into even and odd Julian dates. We form power spectra for each separately, allowing the cross-multiplication of "even" with "even", for example, for this null test only. If the same sky signal is in both the "even" and "odd" data sets, we expect it to cancel out.

In investigating Figure 3.13, we focus on three main possibilities — whether the data points and error bars are consistent with thermal noise ("passing"), whether the error bars are consistent with zero but not consistent with thermal noise ("passing but has an additional variance"), or whether the error bars are not consistent with zero at all ("failing"). We examine each case in the context of our results below.

Firstly, all three null tests display data points that lie within the thermal noise gray band for $k > \pm 0.25\,h$ Mpc^{-1}. In addition, all three null tests show error bars consistent with the thermal noise level for those same k's. This implies that the two jackknife halves do not differ by an amount greater than the thermal noise (i.e., the baselines making up the two jackknives either do not contain bias, or contain similar amounts of bias; we suspect it is the former though more thorough jackknives along this axis are needed to make this conclusion). We deem these as "passing" null tests for the specific jackknives taken (again, dividing up the data in a different way along the same axes may not yield the same results, so more thorough testing is needed to be sure).

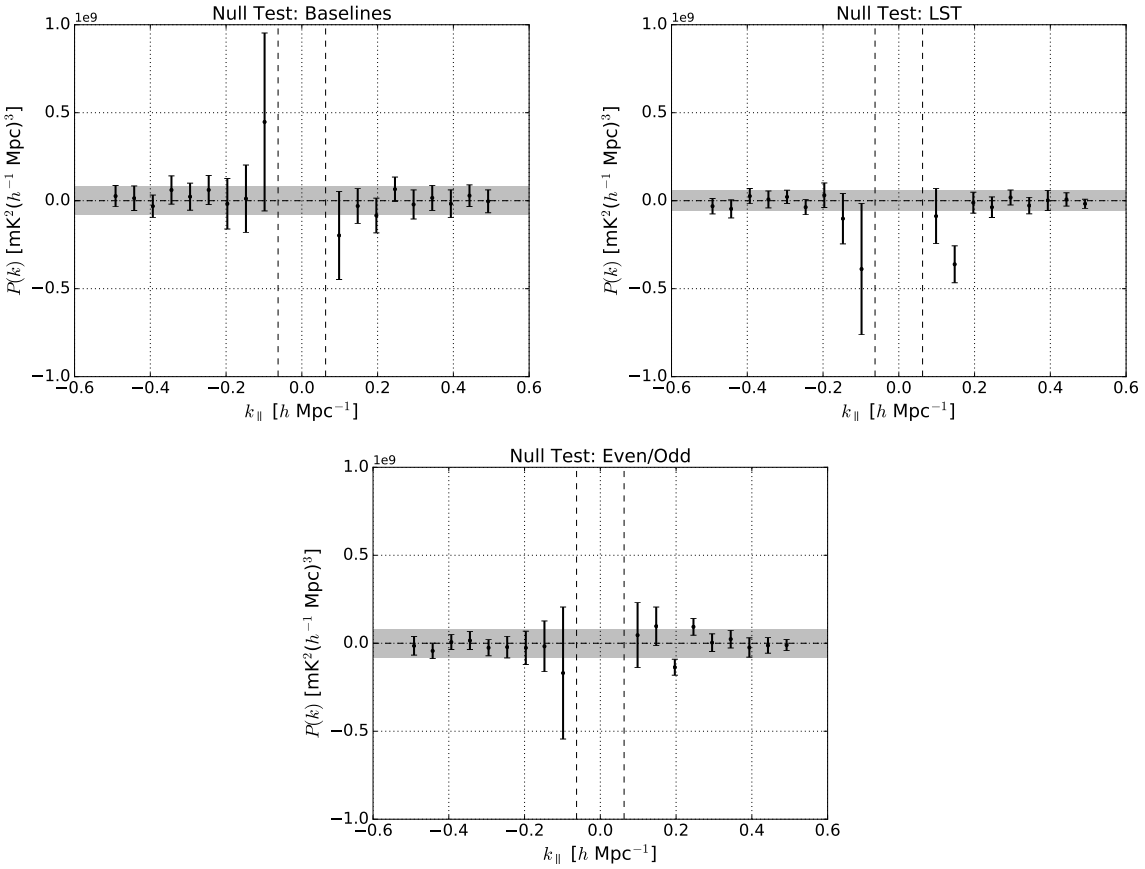

Figure 3.13: Differenced power spectrum results (with 2σ bootstrapped errors) for three null tests, where a jackknife is taken along the baseline axis (top left), LST axis (top right), and even/odd Julian date axis (bottom). The results shown are unweighted (no signal loss), where the power spectrum values plotted are computed from the difference between two power spectra produced on either side of the jackknife axis. The gray shaded region in each plot is the estimated 2σ theoretical noise limit given the parameters of each test. We find that there are no significant systematics for $k > \pm 0.2\,h$ Mpc^{-1} for all three tests. However, we find that all tests exhibit an extra variance at k-values near the horizon ($k \sim \pm 0.1\,h$ Mpc^{-1}), likely due to foreground-noise coupling terms when foregrounds are brightest. Additionally, we find that the LST null test is not fully consistent with zero, implying a bias that is LST dependent and likely caused by varying foregrounds.

The second null test possibility (error bars consistent with zero but not with thermal noise) is displayed by the k-values just outside the horizon ($k \sim \pm 0.1\,h$ Mpc^{-1}) for all three tests. This indicates an additive noise component that is increasing our errors. More specifically, although we expect each cross-multiplication that is used in power spectrum estimation to have independent noise, there is still the possibility of a noise-foreground coupling term that can introduce power. This is because cross-multiplications produce four additive terms — a signal-squared term (where "signal" includes both foregrounds and EoR), two cross-terms between the signal and noise, and one noise-only term. When differencing two power spectra (each with their own four terms), we expect the signal-only term to subtract out for a "passing" null test, and we expect the noise-only terms to be consistent with thermal noise. While the cross-terms have a mathematical expectation value of zero, in practice we are limited by our number of samples (90 cross-products for this analysis times ~ 8 independent LST samples). Combined with the fact that the foregrounds are so bright, the finite ensemble of the couplings can introduce extra variance that varies with foreground strength. It is therefore not surprising that this effect is largest at k-values just outside the horizon, where we expect foregrounds to be brightest post-delay filtering.

Lastly, a third null test result is an error bar not consistent with zero. This is the case for the LST null test at $k \sim -0.1\,h$ Mpc^{-1} and $k \sim 0.15\,h$ Mpc^{-1}, as well as the even/odd test at $k \sim 0.2\,h$ Mpc^{-1} and $k \sim 0.25\,h$ Mpc^{-1}. In such a case, the two jackknife halves differ by an amount greater than the thermal noise (i.e., the data point is not in the thermal noise band), yet they are each constrained tightly by individual error bars that when combined, are not consistent with zero. This result implies that there exists a low level bias that is LST-dependent and Julian date-dependent. The former is likely caused by residual foregrounds that vary in LST, and it is not surprising that this type of bias occurs near the horizon limit. The latter requires further investigation that we leave to future work.

In this section we have presented the first jackknife and null tests from the PAPER experiment. Unsurprisingly, they imply that our measurements are biased by foregrounds, and not the EoR signal (a clean detection of EoR would have passed all three tests). While simple, these tests outline a framework that can be used by future measurements. The 21 cm community is beginning to recognize the importance of these types of tests (Pober et al. 2016a) in characterizing power spectra at the EoR level, and it is clear that future results will require more substantial and thorough investigations of this nature.

3.5 Summary

Although current 21 cm published power spectrum upper limits lie several orders of magnitude above predicted EoR levels, ongoing analyses of deeper sensitivity data sets from PAPER, MWA, and LOFAR, as well as next generation instruments like HERA, are expected to continue to push towards EoR sensitivities. As the field progresses towards a detection, we have shown that it is crucial for future analyses to have a rigorous understanding of signal loss in an analysis pipeline and be able to accurately and robustly calculate

both power spectrum and theoretical errors.

In particular, in Chapters 2 and 3 we have investigated the subtleties and tradeoffs of common 21 cm power spectrum techniques on signal loss and error estimation, which can be summarized as follows:

- Substantial signal loss can result when weighting data using empirically estimated covariances due to couplings with the data realizations (Chapter 2.2). Loss of the 21 cm signal is especially significant the fewer number of independent modes that exist in the data. Hence, there exists a trade-off between sensitivity driven time-averaging techniques such as fringe-rate filtering and signal loss when using empirically estimated covariances.

- Signal injection and recovery simulations can be used to quantify signal loss (Chapter 3.2.1). However, a signal-only simulation (i.e., comparing a uniformly weighted vs. weighted power spectrum of EoR only) can underestimate loss by failing to account for correlations between the data and signal which can be large and negative.

- Errors that are estimated via bootstrapping can be underestimated if samples in the data set are significantly correlated (Chapter 3.3.1). However, if the number of independent samples in a data set is well-determined, bootstrapping is a simple and accurate way of estimating errors.

As a consequence of our investigations, we have also used a subset of PAPER-64 data to make a new power spectrum analysis. This serves as an illustrative example of using a signal injection framework, correctly computing errors via bootstrapping, and accurately estimating thermal noise. Our revised PAPER-64 limits are presented in Chapter 4, which supersede all previously published PAPER limits. Because of the many challenges associated with signal loss and its estimation as described in this chapter, there we use a straightforward power spectrum estimation approach that is not lossy. However, the main reasons for a previously underestimated limit (Ali et al. 2018) and ways in which our new analysis differs can still be summarized by the following:

- Signal loss, previously found to be $< 2\%$ in A15, was underestimated by a factor of >1000 for the case of empirically estimated inverse covariance weighting. Using a regularized covariance weighting method can minimize loss (Chapter 3.2.3), however, because a regularized weighting method is not as aggressive as the former, it produces limits that are still higher than the lossy empirical inverse covariance limits. Underestimated signal loss therefore represents the bulk of our revision.

- Power spectrum errors, originally computed by bootstrapping, were underestimated for the data by a factor of ~ 2 in mK due to oversampling data whose effective number of independent samples was reduced from fringe-rate filtering (Chapter 3.3.1). Several other errors were also found regarding error estimation, though with smaller effects.

- Several factors used in an analytic expression to predict the noise-level in PAPER-64 data were revised, yielding a decrease in predicted sensitivity level by a factor of ~ 3 in mK (Chapter 3.3.2). We note that our sensitivity prediction is revised by a factor less than our overall power spectrum result, implying that if taken at face value, the theoretical prediction for noise in A15 was too high for its data points.

The future of 21 cm cosmology is exciting, as new experiments have sensitivities that expect to reach and surpass EoR levels, improved foreground mitigation and removal strategies are being developed, and simulations are being designed to better understand instruments. On the power spectrum analysis side, robust signal loss simulations and precise error calculations will play critical roles in accurate 21 cm results. With strong foundations being established now, it is safe to say that we can expect to learn much about reionization and our early Universe in the coming years.

Chapter 4

PAPER-64 Revised Results

4.1 Introduction

As described in the re-analysis in Chapter 3, signal loss (the unintentional removal of the target cosmological signal) in the empirical covariance inversion method was discovered in previous PAPER analyses (Ali et al. 2018). More specifically, in A15 this signal loss resulted from the use of empirically estimated covariance matrices as a weighting matrix in the Quadratic Estimator (QE) during power spectrum estimation. An empirically estimated covariance matrix contains terms related to the data, and this dependence induces higher order (i.e., non-quadratic) terms in a QE. Applying the Optimal Quadratic Estimator (OQE) normalization despite these terms then gives the wrong power level (i.e., signal loss). This effect is described more thoroughly in Chapter 3. Chapter 3 also describes how the amount of underestimated signal loss in the A15 analysis was further obfuscated by similarly underestimated uncertainties (from both analytic noise estimates and bootstrapped error bars). While Chapter 3 presents a detailed look at the origin of these issues, it does not deliver a revised analysis for the same data. In this chapter we present revised limits on the 21 cm power spectrum using an independently developed pipeline which conservatively has had many steps removed in light of the issues discovered (see Figure 4.1).

Specifically, we aim to make improvements in two areas. First, we use the independently developed pipeline `simpleDS`[1] which has minimal cross common code with the original PAPER pipeline built for A15 and extended by Chapter 3. Second, this analysis reduces the number of pipeline steps. The basic concept of the delay spectrum is retained, with a power spectrum measurement coming from each type of baseline; however several steps have been removed and others replaced. In particular, while the re-analysis described in Chapter 3 focuses almost exclusively on the final power spectrum estimation stage, here the intermediate stages (like foreground filtering) have been re-examined.

We expand the subset of data analyzed in Chapter 3 to the full PAPER-64 data set as used in A15 in order to make our revised power spectrum analysis. Data collection and processing

[1] `github.com/mkolopanis/simpleDS`

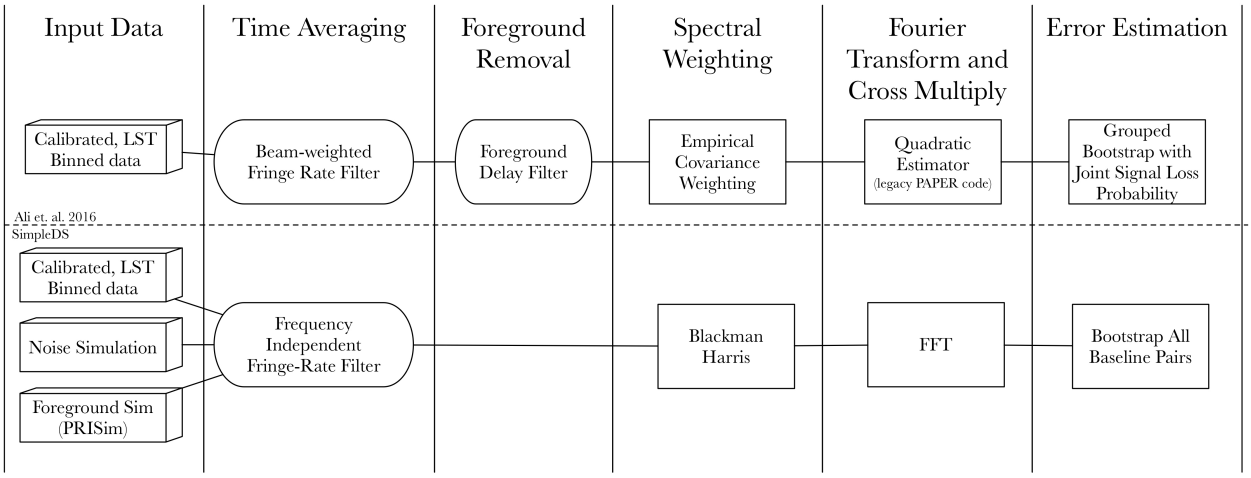

Figure 4.1: Comparison between the prior PAPER analysis by Ali et al. (2015) and `simpleDS`. Our frequency independent fringe-rate filter has a smoother delay response compared to the one used in Ali et al. (2015) and Chapter 3 in order to reduce leakage of foreground power outside the wedge. Additionally, the delay filter for foreground removal has been omitted from this analysis to keep the pipeline as simple as possible. While the foreground removal technique should not affect cosmological signals outside the wedge (Parsons & Backer 2009; Parsons et al. 2012; Parsons et al. 2014), recent works have shown that the use of this filter does not produce a statistically significant reduction in power at high delay modes (Kerrigan et al. 2018). Also, we find that the Fourier-transform used to go from frequency to delay is not dynamic range limited when including foreground signals. Most importantly, in order to avoid signal loss during power spectrum estimation, we use a uniformly weighted Fast Fourier-Transform (FFT) estimator instead of the empirical inverse covariance weighted OQE used in previous PAPER analyses.

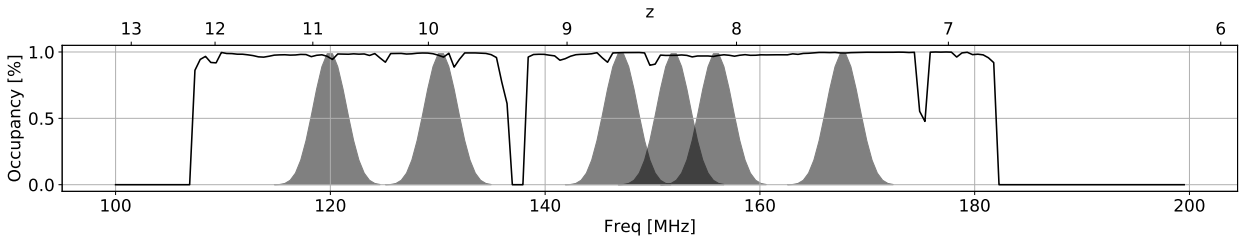

Figure 4.2: The six frequency bands used in this analysis plotted over the relative number of total binned days in a frequency bin (e.g., the relative fraction of total days used in LST binning). Redshift bands are denoted by the Blackman-Harris window functions used during the Fourier-transform from frequency to delay in order to reduce foreground leakage to high delays. All frequency bands used in this analysis have been shown, including the $z = 8.37$ (151.6 MHz) band analyzed in Chapter 3 and Ali et al. (2015). This redshift bin is included in order to properly compare with previous works, but it is worth noting the information obtained from this bin is not entirely independent from the two redshift bins with which it overlaps.

steps prior to LST-binning remain identical to previous analyses (see Chapters 3.1.1 and 3.1.2). We analyze the same three ~ 30 m baselines as in A15 and form estimates for five independent (and six total) redshift bands. In selecting our bands, we note that a practical limitation in selection comes from a desire to avoid including channels with significant RFI flagging. Bands with the most continuous spectral sampling span the redshift range 7.5 to 11, and we select redshift ranges that are approximately coeval, i.e., bandwidths over which limited evolution of the 21 cm signal is expected. Constraints from EDGES *high* suggest that the evolution is probably slower than a Δz (total change in redshift) of 1-2 (Monsalve et al. 2017) which corresponds to a spectral width of 26 MHz at frequency 200MHz and 12 MHz at frequency 100MHz. We adopt a band size of 10 MHz as a conservative size.

This band size allows us to choose a number of spectral windows with very little to no RFI flagging. The specific windows chosen here are centered on $z = 10.87$, 9.93, 8.68, 8.13, and 7.48 (119.7, 130.0, 146.7, 155.6, and 167.5 MHz respectively). As a validation check, we also include a reprocessing of the $z = 8.37$ bin centered at 151.6 MHz which was analyzed in A15 and Chapter 3. These bands are illustrated visually in Figure 4.2.

This chapter is organized in three main parts. We first present major differences between the `simpleDS` pipeline and that of A15, including changes in the fringe-rate and foreground filtering steps of our analysis pipeline and an investigation into the redundancy of PAPER-64 which motivated the removal of certain contaminated data prior to power spectrum analysis. Then, we present our multi-redshift power spectrum results supported by a discussion into systematics through the interpretation of jackknife and null tests. Finally, we present 21 cm upper limits within the context of the broader field.

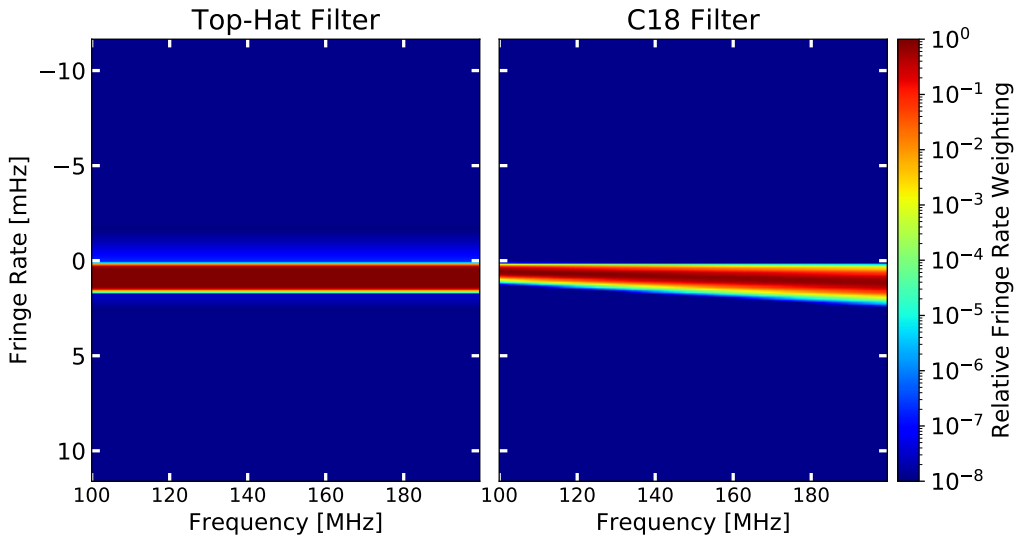

Figure 4.3: A comparison of the Top-Hat fringe-rate filter (TH, left) and the filter used in Chapter 3 (right) in the fringe-rate, frequency domain. The Chapter 3 filter (and the Ali et al. (2015) filter) varies with frequency and this spectral variation can cause additional structure when performing a delay transform of the visibilities. In the interest of simplicity in this analysis, we choose to perform time-averaging with the Top-Hat filter.

4.2 A Simplified Pipeline

In this section we describe the differences in analysis steps prior to power spectrum estimation between this work and A15, including time-averaging, foreground removal techniques, and flagging on redundant baselines (see Figure 4.1 for a visual representation of pipeline differences).

4.2.1 Time-averaging

The LST-binned PAPER-64 data have been averaged into 43 s bins, a timescale which is short compared to the $\sim$ 3500 s fringe coherence time of 30 m baselines. Here, as in past PAPER analyses, we choose to perform time-averaging by convolving the time stream with a windowing function. This function is defined as a filter in fringe-rate space (the Fourier dual to time) which can be tuned to maximize sensitivity to sky-like modes and exclude slowly varying systematics. Parsons & Backer (2009) show that a fringe-rate corresponds to sky-like rates of motion which map geometrically to rings on the sky. Parsons et al. (2016) then show that a fringe-rate filter (FRF) can be defined with weights corresponding to the instrument's primary beam power integrated along the line of constant fringe-rate. Applying an FRF with this weighting provides an optimal coherent integration in time for a baseline of a given length.

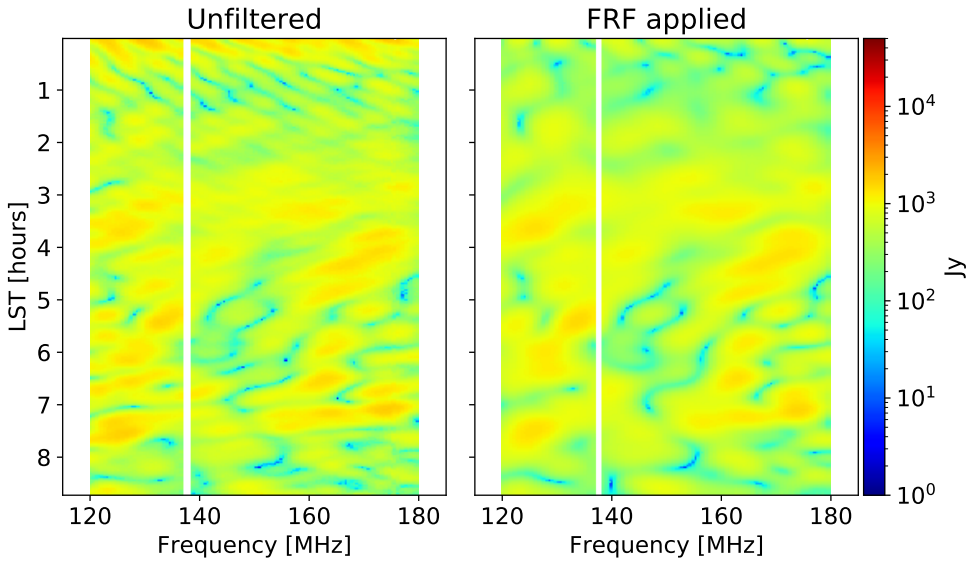

Figure 4.4: LST and frequency waterfalls of a representative baseline taken from the even LST-binned set before (left) and after (right) application of the Top-Hat FRF. The baseline illustrated is the antenna pair (1,4). The application of the fringe-rate filter removes very fast fringe modes but preserves the structure of sky-like modes.

Previous PAPER analyses have used variations on such a filter. For example, A15 formed the beam-weighted filter, fit a Gaussian in fringe-rate space, and then artificially increased the width of the Gaussian to provide easy parameterization across the PAPER bandpass and decrease the effective integration time. A similar Gaussian fit was also used and discussed in the PAPER-64 case study in Chapter 3, but the width of the fit was not increased in that analysis (i.e., the optimal filter was used).

However, as can be seen in the right panel of Figure 4.3, this filter is frequency dependent. In particular, the maximum fringe-rate range probed by a baseline increases linearly with frequency. This spectral dependence is inherently smooth; however in implementation this filter may introduce additional structure during the delay transform; further investigation is needed to find the best approach for mitigating this effect. We also recall that the use of these "aggressive" fringe-rate filters is shown to contribute to signal loss when used in combination with certain weighting techniques (Chapter 2.2.3).

Therefore, as a simplification to avoid potential signal loss and reduce contamination of high delay modes, we adopt a Top-Hat filter (left panel, Figure 4.3) that weights all fringe-rates evenly across frequency, similar to the filter used in Parsons et al. (2012). The maximum fringe-rate passed by our filter is set by the highest frequency included in the data set; the lowest fringe-rate is chosen to exclude known common mode signals with zero fringe rates.

Common mode signals are sky signals which vary on time scales longer than would be

expected from an ideal interferometer (Ali et al. 2015). Such common modes were previously referred to as "crosstalk" — however, these signals may not necessarily result from signals observed in one antenna and leaked to another but rather any time-independent signal which is observed by all antennas. We exclude common mode signals here by setting the minimum fringe-rate included in the filter to 3.5×10^{-5} Hz; this excludes all modes with periods longer than ∼45 minutes. We also filter out all negative fringe-rates.

Waterfalls of a representative baseline before and after the application of the fringe-rate filter are shown in Figure 4.4. The application of the fringe-rate filter removes very fast fringe modes but preserves the structure of Eastward moving sky-like modes.

4.2.2 Foreground Removal

To mitigate foreground contamination during power spectrum estimation, PAPER analyses have used a wide-band iterative deconvolution algorithm (WIDA), often referred to as a "clean-like" iterative deconvolution algorithm (Chapter 3.1.2). This algorithm relies on the underlying mathematics of CLEAN as described in Högbom (1974) to remove delay components from PAPER data inside of some range of delays. This type of deconvolution and its specific application to radio data is described in Parsons & Backer (2009). The WIDA was used in Parsons et al. (2012), Parsons et al. (2014), Jacobs et al. (2015), A15, Kerrigan et al. (2018), and Chapter 3.

We choose to omit this filtering technique in our simplified analysis. While the technique should not affect cosmological signals outside the user-defined range of delays to clean (Parsons & Backer 2009; Parsons et al. 2012; Parsons et al. 2014, and explored further in Kerrigan et al. 2018), recent works have also shown that the use of this filter, in certain situations, does not produce statistically significant reduction of power at high delay modes (Kerrigan et al. 2018). Since our analysis aims to focus on upper limits set at high delay modes, we omit this step in the interest of simplicity. Even without any attempt to remove foregrounds from the visibility data, we find that our delay transform used to estimate the cosmological power spectrum is not limited by the inherent dynamic range of the transform.

4.2.3 Flagging on Redundancy

We have described how we simplified both the time-averaging and foreground filtering steps of our pipeline. Finally, before estimating the power spectrum of our data, we conduct statistical tests on our observations to determine the degree to which our baselines are redundant. We do this because the per-baseline delay spectra described in Parsons et al. (2012) can be averaged across all baseline cross-multiples only for perfectly redundant baselines. While it is unrealistic to assume the PAPER baselines are perfectly redundant, this analysis can help identify extreme outliers which should not be used in power spectrum estimation.

In Figure 4.5, we show the 8 x 8 antenna configuration used in the PAPER-64 deployment, which was chosen to increase power spectrum sensitivity through having many copies of the same baseline (i.e., redundant observations). Each of the three baseline vectors shown in

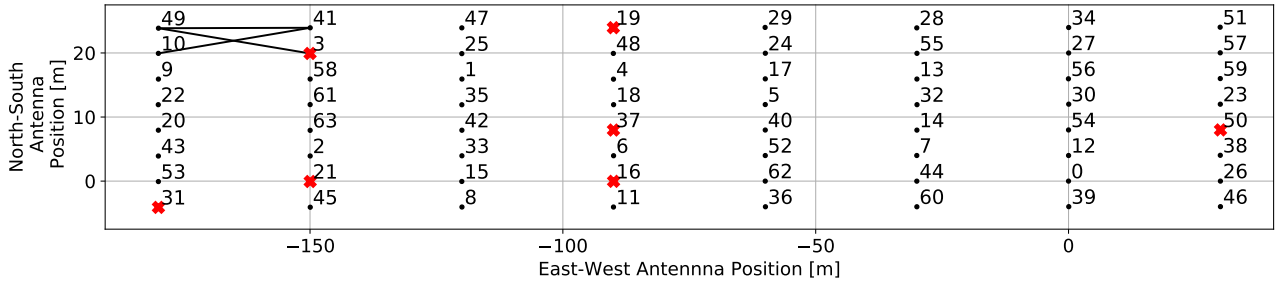

Figure 4.5: The antenna positions of PAPER-64. Highlighted in black are the three baseline types used in this analysis. These baselines consist of East-West baselines from adjacent antenna columns with no row separation (e.g., 49-41, 1-4, 0-26), baselines with one column separation and one positive Northward row separation (e.g., 10-41, 1-48, 0-38), and baselines with one column separation and one negative Northward row separation (e.g., 49-3, 1-18, 0-46). A red 'x' denotes antennas which have been flagged from the analysis. Reasons for flagging include previously known spectral instability (19, 37, and 50), low number of counts in LST-binning (3 and 16), and suspected non-redundant information (21 and 31).

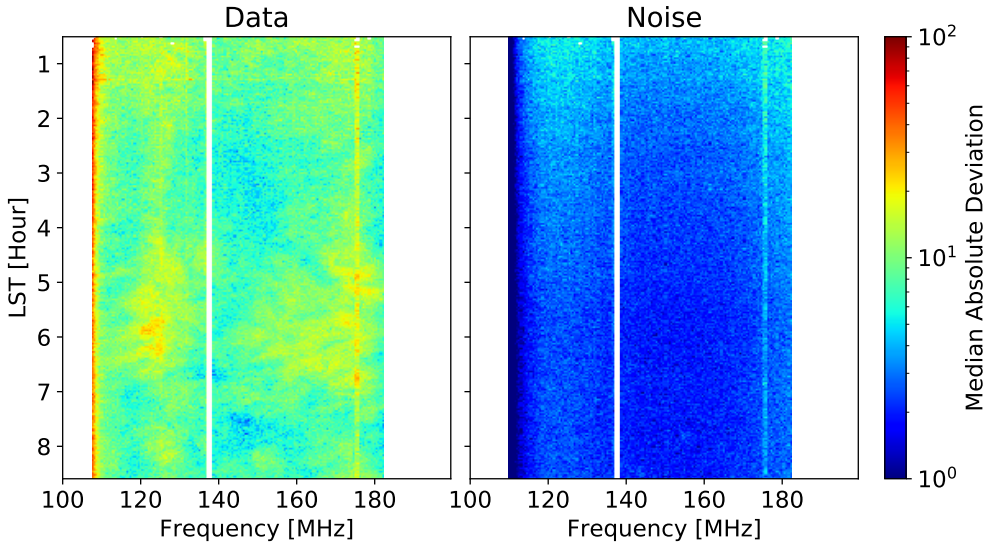

Figure 4.6: A representative Median Absolute Deviation (MAD) for both data (left) and a noise simulation (right) computed for each time and frequency observed by PAPER in the LST range $00^h30^m00^s - 08^h36^m00^s$. The data shown here corresponds to strictly 30 m East-West baselines. For perfectly redundant sky measurements the individual baseline measurements will only differ by thermal noise. The large amplitude of deviations observed on the left illustrates that there is a significant amount of non-redundant information in the data.

black (which we use for our revised power spectrum analysis in this chapter) are sampled many times across the grid-like array. Rather than average baselines together, we cross-multiply all redundant pairs and then bootstrap average for an estimate of variance (Chapter 3.3.1).

A first test of the array's redundancy is to compare the measured variation between baselines with that expected due to thermal noise, using Equation (3.15) to generate noise for each time and frequency for each baseline. As a measure of variance between baselines we take the Median Absolute Deviation (MAD) of the visibility amplitude across redundant baselines for each frequency and time, defined as

$$\mathrm{MAD}(t,\nu) = median\left(\left||V_{i,j}(t,\nu)| - median\left(V(t,\nu)\right)\right|\right) \tag{4.1}$$

where the median visibility amplitude is taken at each time and frequency across the redundant baseline group.

The MAD for both data and our noise simulation is shown in Figure 4.6. For perfectly redundant sky measurements the individual baseline measurements will only differ by thermal noise. We see that some measurements have a MAD consistent with thermal noise — however the larger deviations observed at other frequencies and times illustrate a significant amount of non-redundant information in the data.

We then use the MAD to estimate the significance of each baseline's deviation from the median baseline measurement using the modified z-score $M_z(t,\nu)$ defined as

$$M_z(t,\nu) = 0.6745\frac{|V_{i,j}(t,\nu) - median\left(V(t,\nu)\right)|}{MAD} \tag{4.2}$$

which can be thought of as the number of standard deviations away from the median each data point is. The 0.6745 scaling factor is introduced to normalize the modified z-score for a large number of samples (Iglewicz & Hoaglin 1993).

The waterfalls of z-scores (in time and frequency) are averaged in LST and a maximum is taken along the frequency dimension to find a single worst case modified z-score for each baseline. A histogram of these z-scores for every baseline in this analysis is shown in Figure 4.7. As suggested in Iglewicz & Hoaglin (1993), any baseline with a modified z-score of at least 3.5 is identified as a potential outlier.

Using this metric, only the baseline $(21, 31)$ is identified as an outlier. With only one outlier, it is difficult to tell if this deviation is due to a single defective antenna or the physical baseline. We take the most conservative approach and flag both antennas 21 and 31 from all remaining analysis. Although no other baselines qualify as outliers using our metric, the existence of scores above 2 in general indicates an amount of non-redundancy inconsistent with thermal noise from the baselines in this analysis and may affect the interpretation of our final power spectrum estimates.

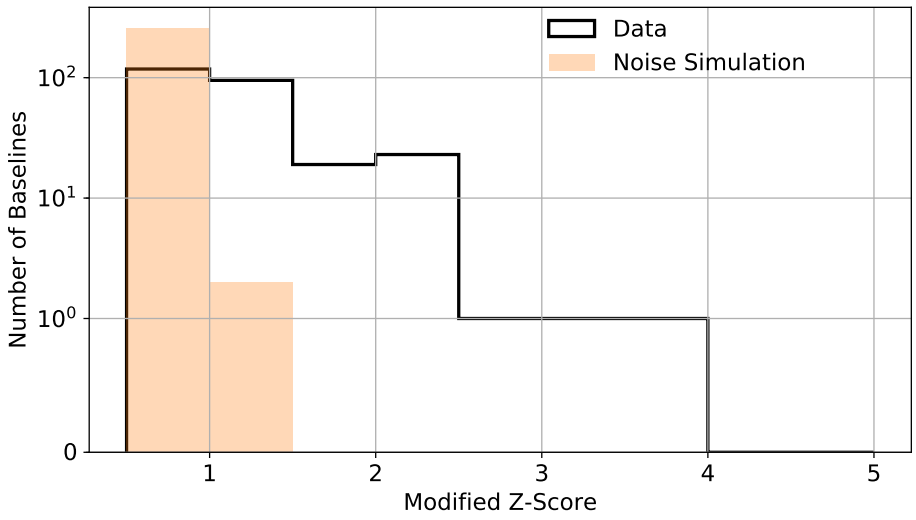

Figure 4.7: A histogram of modified z-scores of data (black) and input noise simulation (orange) suggests that a cut on baselines with z-score larger than 3.5 will safely avoid statistically outliers (e.g., non-redundant baselines). This is consistent with the recommendation from Iglewicz & Hoaglin (1993). Using this metric, only the baseline (21,31) is a statistically significant outlier. Since it is unclear which of the two antennas may be contributing to this non-redundancy we flag both antennas in all further analysis.

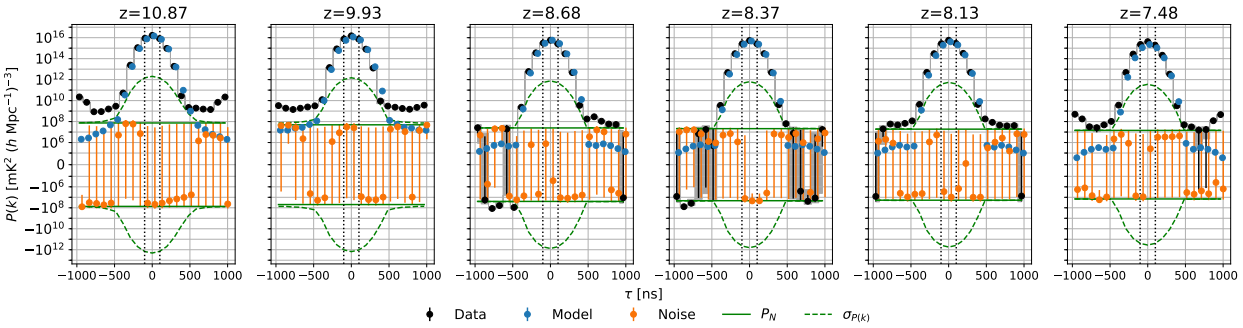

Figure 4.8: Power spectrum estimates computed for the observed data (black), simulated noise (orange), and simulated foreground observation (blue). Error bars on data points are the bootstrapped 2σ uncertainty. The solid green line indicates the theoretical thermal noise estimate for each redshift bin, and the dashed green line includes a modeled foreground error (derived in Kolopanis et al. (*in prep.*)). Gray shaded regions are the foreground dependent uncertainties plotted around each data point. The vertical black dotted lines indicate the horizon/wedge/light travel time for a 30 m baseline. We find that the simulated noise is consistent with the theoretical thermal noise predictions (orange vs. solid green). At delay $\tau = 0$ ns, both the data and foreground simulation show good agreement in the total power observed; generally, the power at all delays inside the horizon agrees between the two simulations within a factor of ~ 5. The simulated data set also shows some power leakage outside the horizon, consistent with the power observed by PAPER out to ~ 400 ns. The PAPER data also show numerous statistically significant detections beyond 400 ns, however, which are not predicted by the foreground simulation. To investigate the origin of these signals, jackknife and null tests are performed.

4.3 Multi-Redshift Power Spectrum Results

In this section, we present power spectrum results for all six of our redshift bands. We show three principal products in our figures: our observed data, a simulated foreground observation, and a noise-only simulation. As a high level overview, the simulated foreground observation is created using the simulator PRISim (Thyagarajan et al. 2015a,b), which performs a full-sky visibility calculation matching the observing parameters of PAPER's LST-binned data set. For more details about the construction and accuracy of the sky simulation, we refer the reader to Kolopanis et al. (*in prep.*). Power spectra of the other two products, data and noise-only, are produced similarly to those in Chapter 3, but with an unweighted, non-lossy estimator.

Figure 4.8 shows the delay power spectrum estimates for our three products: the observed data (black), the PRISim simulated foreground observation (blue), and the noise-only simulation (orange). Within delay modes between $\sim \pm 400$ ns, both the observed and simulated data illustrate similar shapes. This suggests that the statistically significant detections of power observed in PAPER immediately outside the horizon limits are consistent with fore-

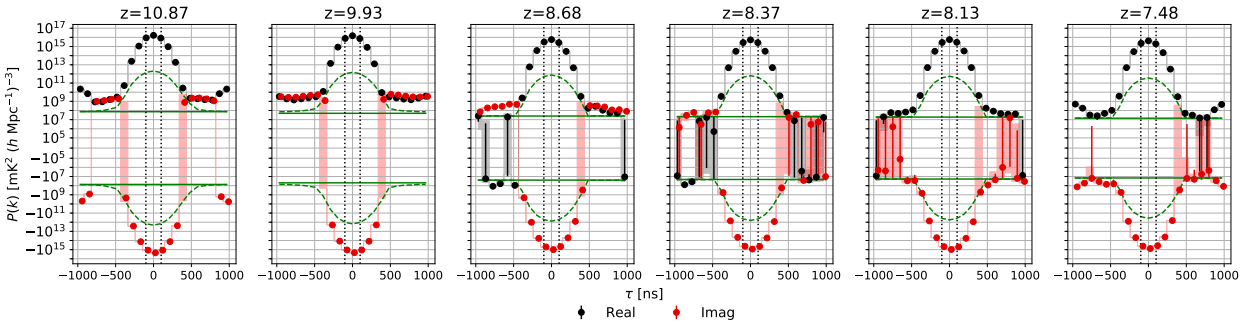

Figure 4.9: The real (black) and imaginary (red) components of the power spectrum. The red shaded region is the foreground dependent theoretical error drawn around the imaginary components; all other lines are the same as in Figure 4.8. There are statistically significant imaginary components at $|\tau| < 400$ ns, generally at a power level which is $\sim 20\%$ of the real components at the same delay. This may result from non-redundancies in calibration or baseline orientation. At delay modes $|\tau| > 400$ ns, the imaginary component of the power spectrum displays comparable power to the real part. This is especially prominent in, but not isolated to, the two highest redshift bins (lowest frequencies). The statistically significant imaginary power is indicative of some non-redundant information during power spectrum estimation, systematic biases introduced during data analysis or calibration, or residual contaminants like improperly flagged RFI.

ground signals (as suggested by the study of foreground subtraction applied to PAPER data in Kerrigan et al. 2018). At larger delays, however, the PAPER power spectra are a mix of statistically significant detections and null results. The most statistically significant detections at high delays are seen to occur at the lowest frequencies. In the next few subsections, we present several analyses designed to help determine the cause of the statistically significant detections at high delays seen in the PAPER observations.

4.3.1 Investigation of High Delay Detections

We have seen that the PAPER power spectrum results exhibit statistically significant detections at high delays and aim to understand their origin. The power spectrum is computed by cross-multiplying different baseline pairs within redundant baseline groups. Ideally, this cross-multiplication of complex-valued delay spectra will result in any sky-like power being confined to the real part in the power spectrum, leaving the imaginary part dominated by noise. However, effects can leak real sky power into the imaginary part of the spectrum. A perfectly calibrated array with non-redundant baselines — for example, with slightly different antenna positions — will cause two nominally "redundant" baselines to have slightly different phases. The imaginary parts of these cross-multiplied visibilities will therefore not cancel out, and non-zero power will be seen in the imaginary component of the power spectrum estimate. The same effect would come from a perfectly redundant but imperfectly

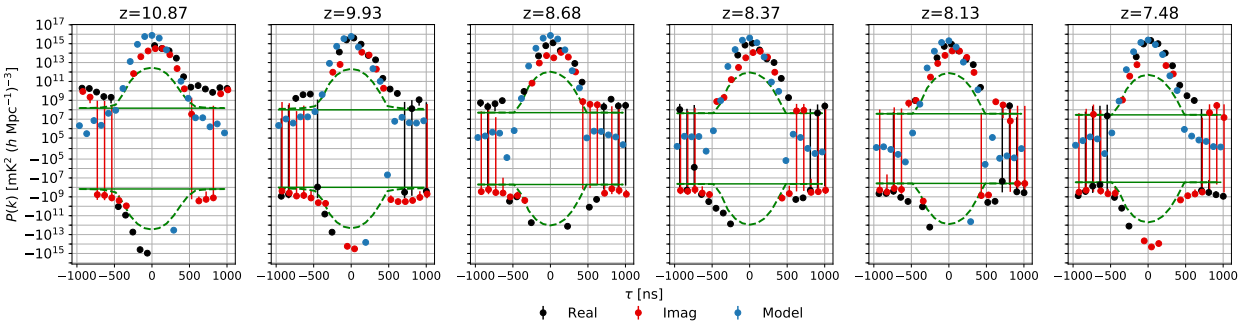

Figure 4.10: Null tests constructed by splitting the LST range ($[00^h30^m00^s, 08^h36^m00^s]$) in half (at 04^h30^m), making two power spectrum estimates and differencing the result. Real (black) and imaginary (red) parts are both shown, along with the null test results when applied to the simulated data (blue). All noise estimates shown are as described in Figure 4.8. Such a null test should remove isotropic cosmological signals, leaving behind anything with dependence on sidereal time. Statistically significant detections in the real part suggest power varying across the sky while significant imaginary power suggests a time dependence to phase calibration errors. The observed variations are consistent with the simulation up to delays of 400 ns. The detections at higher delay modes indicate a large LST dependence which is inconsistent with cosmological power.

calibrated array. In a sense, the amount of statistically significant power in the imaginary component of the power spectrum, compared to power in the real part, is a measure of the net redundancy and calibration quality of the array.

A comparison of the real and imaginary parts of the power spectrum is shown in Figure 4.9. The statistically significant imaginary components at $|\tau| < 400$ ns (red) are generally at a power level which is $\sim 20\%$ of the real components (black) at the same delay. This may result from non-redundancies in calibration or baseline orientation.

At delay modes $|\tau| > 400$ ns, the imaginary component of the power spectrum displays comparable power to the real part. This is especially prominent in the two highest redshift bins (lowest frequencies), but is observable across the entire band. The disagreement between the imaginary component (red) and the estimated foreground error (dashed green) is indicative of some non-redundant information, systematic biases introduced by data analysis or calibration steps, or residual contaminants like improperly flagged RFI.

4.3.2 Null Tests

While the presence of imaginary power suggests at least some presence of calibration error or non-redundancy, it does not fully explain the origin of the excess power at delays greater than 400 ns. Calibration errors, as long as they do not introduce spectral structure, should not necessarily scatter power to high delays. Null tests — i.e., differences between power spectra of different data selections — can provide hints at the origin of these detections

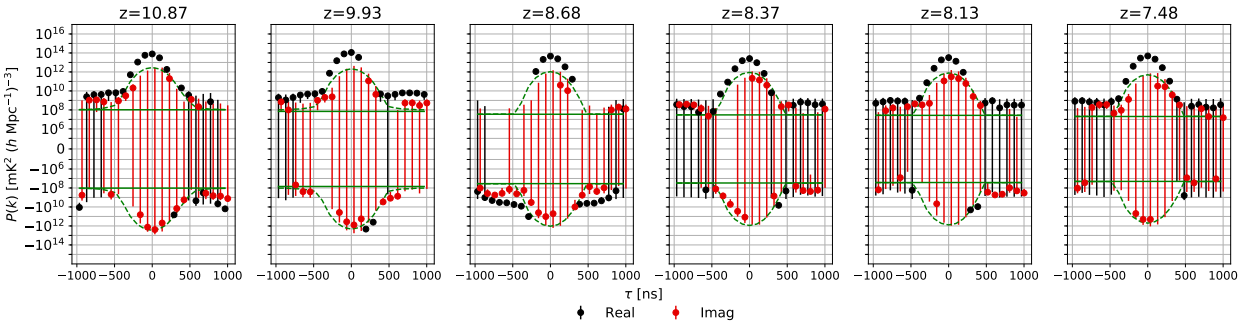

Figure 4.11: Jackknife test constructed by splitting the data set into even and odd Julian days. Plotted here is the difference between the power spectra from these two sets. We use the same color scheme as Figure 4.10. While the largest difference in the LST null test (Figure 4.10) was on the order of 10% of the measured value, here differences are less than 1% at delays less than 400 ns, and the imaginary points are nearly all consistent with predicted error bars. At delays larger than 400 ns, statistically significant detections in the three highest redshift bands are at comparable levels to the power spectrum values in Figure 4.8. This may be the result of a corrupt day of data that is present in only one set of the even or odd data (positive value for even, negative values for odd), which is mitigated during the cross-multiplication of these sets during power spectrum estimation.

(Chapter 2.5.2).

For example, differencing the power spectra of two distinct ranges of sidereal time will remove isotropic cosmological signals but leave behind signals with strong dependence on sidereal time (like foregrounds). Dividing our data set in half by LST into ranges $[00^h30^m00^s, 04^h30^m00^s)$ and $[04^h30^m00^s, 08^h36^m00^s)$ creates two sets of roughly equal sensitivity. The resulting differenced power spectrum is shown in Figure 4.10, along with a matching calculation for the foreground simulation. The two are broadly consistent at delays less than 400 ns, i.e., they have the same sign and a similar amplitude. Galactic synchrotron emission and bright point sources (like Fornax A and Pictor A) are the most obvious contenders for strong variability. We also see that the significant power seen in modes well beyond the horizon (for example, the strong positive offset at redshift 9.93 seen in the Figure 4.8 power spectrum) is reflected in this null test.

Additionally, we see that the imaginary component of the power spectrum null test is comparable to the real component at most delay modes across all redshifts. This suggests a sidereal time dependence of phase differences between baselines. In particular, note that the strong bias seen at redshift 9.93 is associated with a strong imaginary bias, implying a phase difference between baselines. Such an LST dependence of the imaginary component might be expected for non-redundancy (slightly different sky seen by nominally redundant baselines) or repeatable differences in calibration which depend on the sky configuration (for example, obtaining one calibration solution when Fornax is transiting and a different one for when Pictor dominates). This kind of variation in redundant calibration with sky flux density is

shown in Joseph et al. (2018). We note that this picture of non-redundancy strengthens the earlier hints provided by the z-score analysis in Section 4.2.3, which suggested that redundancy was particularly low around 120-130 MHz (redshifts 9 and 10).

A second easily constructed null test is to difference power spectra made from only the even and odd binned data sets. Recall that these sets were constructed by separating even and odd numbered days during LST-binning. A significant difference in this test would therefore be suggestive of a variation at the night-to-night level, as these two sets are otherwise expected to have identical sky signals with different realizations of noise.

The resulting even/odd differenced power spectra for each redshift band are shown in Figure 4.11. Across all redshifts, there are points well beyond the horizon which are inconsistent with both the analytic thermal noise and the propagated uncertainty. However there are two important differences between this test and the LST null test. First, the overall amplitude of the differenced even/odd power spectrum is much less. Within the wedge, the difference in amplitude is at most a few $\times\ 10^{13}$, or less than 0.1% of the original power spectrum amplitude. Second, the imaginary power spectrum is consistent with noise across most modes. This is particularly notable within the wedge where even a small percent difference would drive a significant deviation. This suggests that whatever causes the small but detectable difference between even and odd is not attributable to a phase difference between baselines.

The two highest redshift bins again show the most significant differences at high delay; for example, the high delay modes at redshift 9.93 reach 10%-20% of the power spectrum amplitude. This suggests the presence of a signal contaminating a single day which is averaged into the LST-binned data set. Examples of such a systematic are improperly flagged RFI, a low amplitude signal not detected before cross-multiplication, or large transient gain isolated to a single night.

Finally, another interesting feature can be seen in the redshift 8.68 bin in Figure 4.11. There we see a consistent bias which was not present in the mean power spectrum (Figure 4.8). However, there is a similarly shaped bias in the *imaginary* part of the mean power spectrum. A plausible hypothesis is that, in this part of the spectrum, phase error between baselines is larger in one of the even or odd LST-binned sets than the other. However, there is no clear significant difference in redundancy seen in the z-score/MAD analysis, so further evidence would be required to support this conclusion.

4.3.3 Null Test Discussion

Our two null tests provide evidence that the foregrounds, which vary significantly as a function of LST, are likely the cause of some of the residual power detected at high delays. There is also some evidence that suggests significant phase differences exist between nominally redundant baselines which introduce non-redundant signals into the power spectrum estimates.

The presence of highly significant detections in the even/odd null test also suggests there may be some net non-redundant signal between the two LST-binned data sets. These

detections are significant compared to the propagated error bar ($\sim 10\sigma$ to $\sim 100\sigma$ inside the horizon) but represent a small fraction of the total power observed ($\leq 1\%$ of the power in Figure 4.8). However the agreement between the imaginary part of the power spectrum with the foreground-dependent error suggests that each of the even/odd sets has internally redundant baselines but the data sets themselves are slightly different.

Both the null tests discussed in this work, along with the presence of a significant fraction ($\sim 20\%$) of power leaking from the real to imaginary component of the power spectrum, indicate the presence of non-redundant and non-isotropic signals. The latter is not surprising since this analysis is performed on data without a foreground subtraction step. Additionally, the sky varies with LST as the galaxy and other sources rise and set over an observation. In some places, particularly at low frequencies, this power couples to larger delays, presumably because of instrumental spectral structure. The even/odd null test suggests that this spectral structure potentially varies in time while the imaginary component suggests that the spectral structure is not the same across nominally redundant baselines.

In summary, many of our tests and statistics suggest that non-redundant information contaminates the nominally redundant PAPER baselines — the distribution of modified z-scores in Figure 4.7, the presence of power in the imaginary component of the power spectrum, and the non-null results from multiple null tests all indicate that the data varies between baselines at a level larger than what is expected from thermal noise.

4.3.4 Possible Future Directions

It is clear that additional jackknife tests would help reveal more specific origins of our systematics. For example, more jackknife tests along LST could be used to identify and possibly remove residual RFI, along with night-to-night variations as identified in the even/odd null test. Because the variation we detected is significant enough to be observable after differencing data averaged over the entire season, a single culprit could potentially be further tracked down by performing tests with smaller sets of binned days (or by performing a null test that differences data from the first and second half of the observing season). This would provide information about the stability of antennas and the observations over the life of the PAPER experiment. Unfortunately, returning to the raw visibility data set is outside the scope of this analysis.

While we have identified non-redundancies as a likely cause of our failed null tests, their exact origins have not been pinpointed. Two obvious ways for non-redundancies to happen are variations in antenna positions and variations in beam patterns. In theory, an element like PAPER should produce a symmetric beam, though this is not true in practice. One simple test for non-redundancy due to beam differences would be to test for deviations from symmetry by recording observations with antennas rotated by 180°. Differencing the 0°and 180°data sets would highlight abnormalities in the beam response to the sky. For an ideal, symmetric beam, all sky signal will cancel and leave thermal noise fluctuations at all times; however imperfections in beam responses will not cancel, resulting in a net signal in the visibility data and constraints on the level of beam-to-beam variation. We leave these

possible tests as suggested future work for upcoming analyses.

4.4 21 cm Upper Limits

Finally, we use the PAPER-64 data to place upper limits on the 21 cm signal using the dimensionless power spectrum: $\Delta^2(k) = \frac{|k|^3}{2\pi^2} P(|k|)$. To convert from interferometric delay to cosmological co-moving wavenumber, we assume WMAP-9 year cosmology in "little-h" units ($H_0 = 100\ h\,\mathrm{km/s/Mpc}$). These power spectra are shown in Figure 4.12.

As a summary and comparison of progress across the field, we also report from each published power spectrum the lowest upper limits achieved by various instruments in the k-ranges reported by each instrument (Figure 4.13). To encapsulate the results of the work in this chapter, the most sensitive limit is reported from the range $0.3 < k < 0.6\ h\ \mathrm{Mpc}^{-1}$, where both null tests pass for most k-modes in each redshift bin. These limits on the 21 cm power spectrum from reionization are $(1440\ \mathrm{mK})^2$, $(1850\ \mathrm{mK})^2$, $(290\ \mathrm{mK})^2$, $(190\ \mathrm{mK})^2$, $(360\ \mathrm{mK})^2$, $(290\ \mathrm{mK})^2$ at redshifts $z = 10.87,\ 9.93,\ 8.68,\ 8.37,\ 8.13$, and 7.48, respectively.

Our results represent increased (less stringent) upper limits compared to prior limits published by the PAPER instrument (by a factor of ~ 10 in mK). They also exceed the expected amplitude of a fiducial `21CMFAST`[2] model by a factor of ~ 100 in mK (Mesinger et al. 2011). However, these results represent the most robust results from the PAPER experiment, replacing results from both PAPER-32 (Parsons et al. 2014; Jacobs et al. 2015; Moore et al. 2017), which used a different covariance estimation technique but have not been subjected to a rigorous re-analysis à la Chapter 3, and PAPER-64 (Ali et al. 2015; Ali et al. 2018). Any constraints on the spin temperature of hydrogen made by Pober et al. (2015) and Greig et al. (2015) based on the previously published upper limits should also be disregarded. Though these measurements do not place significant constraints on the IGM temperature, the analysis presented in those two papers remains relevant to any future limits on the 21 cm power spectrum at levels similar to the original results of A15.

In conclusion, we have re-analyzed the PAPER-64 data first presented in A15 and presented 21 cm power spectra and uncertainties in five independent (six total) redshift bins. These estimates are made using an independently developed pipeline which skips foreground subtraction and simplifies time-averaging. Simulations of noise and foregrounds are used to build a basic picture of internal consistency. The resulting power spectra are consistent with noise across much of the spectrum, but above redshift 9 (below 130 MHz) they demonstrate a statistically significant excess of power. Null tests support a picture where power spectrum detections are caused by foregrounds modulated by spectrally dependent deviations from redundancy or calibration error. In particular, the z-scores and imaginary power tests suggest that residuals could be the result of some net non-redundant signal or imperfections in baseline placement at the $\sim 1\%$ level for the 30 m baselines analyzed in this work.

Future analyses of highly redundant sky measurements will require strict comparisons between nominally redundant samples before cross-multiplication to ensure effects like these

[2] github.com/andreimesinger/21cmFAST

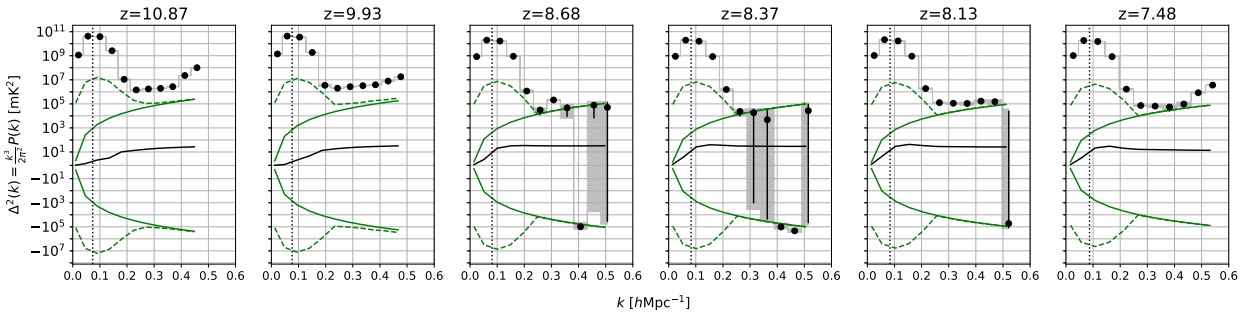

Figure 4.12: The dimensionless power spectrum $\Delta^2(k)$ derived from the PAPER-64 observations (black). All error bars represent 2σ uncertainties. Also plotted are theoretical thermal noise limits estimated from Equation (3.15) (solid green) and a foreground-dependent variance estimate (dashed green and gray shaded; derived in Kolopanis et al. (*in prep.*). The black solid line represents a fiducial `21cmFAST` model of reionization. The horizon line (vertical dotted black) has been transformed from the maximum signal delay between antennas to cosmological co-moving scales using Equations 12 and 13 of Liu et al. (2014a).

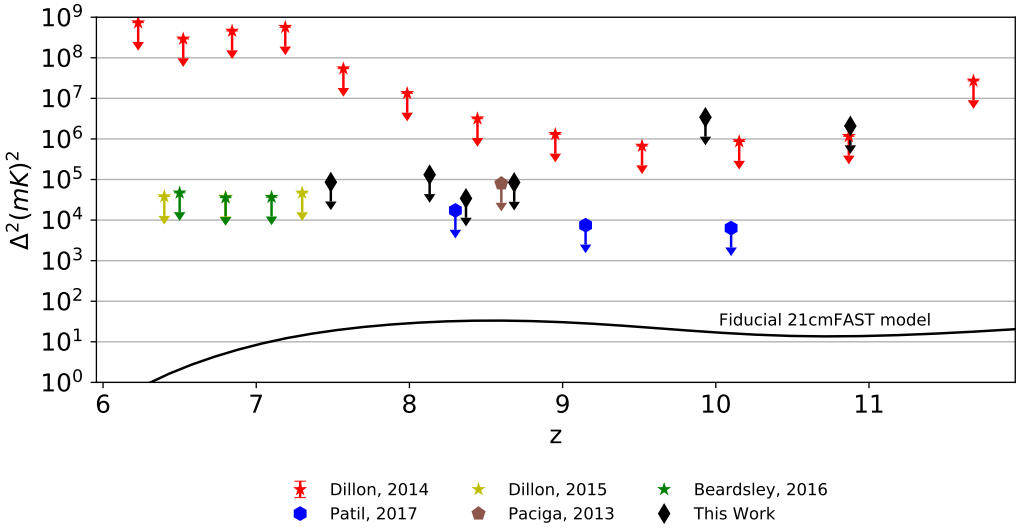

Figure 4.13: A comparison of the lowest limits achieved by various instruments in the k-ranges reported by each instrument. The results reported here are taken in the range $.3 \leq k \leq .6$ h Mpc^{-1}. Data is taken from the MWA (stars; Dillon et al. (2014); Dillon et al. (2015a); Beardsley et al. (2016)), the GMRT (pentagon; Paciga et al. (2013)), LOFAR (hexagons; Patil et al. (2017)), and PAPER (diamonds; this work). We include the $z = 8.37$ redshift bin analyzed in Ali et al. (2015), although it is worth noting this redshift bin is not entirely independent from the $z = 8.13$ and 8.68 bins, as can be inferred from the overlapping window functions in Figure 4.2. For reasons described throughout this thesis, these PAPER results supersede all previous PAPER limits.

can be mitigated. Also, further jackknives and comparisons of data should be performed before or as part of LST-binning to detect likely contributions to excess. Precise antenna placement will also be required to ensure baselines designed to be redundant do not introduce signal in the imaginary component of the power spectrum.

To date, as shown in Figure 4.13, all power spectrum estimates have been reported as upper limits. However, to discern and characterize the physics of reionization, high significance detections of the 21 cm power spectrum are necessary. Next generation radio telescopes, like the fully realized 350 element configuration of HERA (Pober et al. 2014; DeBoer et al. 2017; Liu & Parsons 2016) and the future Square Kilometre Array (SKA; Mellema et al. (2013)), are predicted to be able to make these detections and place stringent constraints on reionization.

Chapter 5

PAPER-128

5.1 Overview

The PAPER experiment expanded out from 64 antennas to 128 antennas in 2013 and observed for two additional seasons before retiring. The first season began in November 2013 and lasted until March 2014. The second began in July 2014 and ended in January 2015. The PAPER-128 configuration consists of 112 core antennas arranged in a grid layout (7 rows and 16 columns), with neighboring East/West spacings being 15 m and neighboring North/South spacings being 4 m (Figure 5.1). Additionally, 16 outrigger antennas were placed in strategic locations in order to form long baselines and increase uv-plane sampling. These outrigger antennas are not used for the power spectrum analysis presented in this chapter, but are useful for imaging analyses.

In general, the signal chain of PAPER-128 is similar to that of PAPER-64. However, one major change is the addition of receiverators on site, which house the receivers used to amplify and filter the antenna signals. Prior to this change, the receivers were located inside a cooled shipping container along with the rest of the DSP system. With the addition of more antennas, however, the receivers were moved outside the container to save space.

Although PAPER-128 doubled in number of antennas, the data collected by this array is typically found to be lower in quality than that of PAPER-64. There are many reasons for this, including general wear and tear on the instrument and the addition of the receiverators (which had no monitoring system and, as we will see, is one of the main reasons for missing and corrupted PAPER-128 data). Because of these issues, PAPER-128 requires the development of novel techniques in order to filter out contaminated data products prior to analysis. Using the entire season of data, without any filtering, would result in a power spectrum analysis severely dominated by systematics and (non-EoR) detections. Thus, one of the unique challenges facing PAPER-128 is how to automatically and accurately detect and remove bad data (i.e., misbehaving baselines, dead receiverators, criss-crossed signal paths, etc.) in order to curate a data set as free of systematics as possible.

In this chapter, we present methods developed for the detection of corrupt data in PAPER-128. These methods represent the first routinely-used "quality assurance" steps

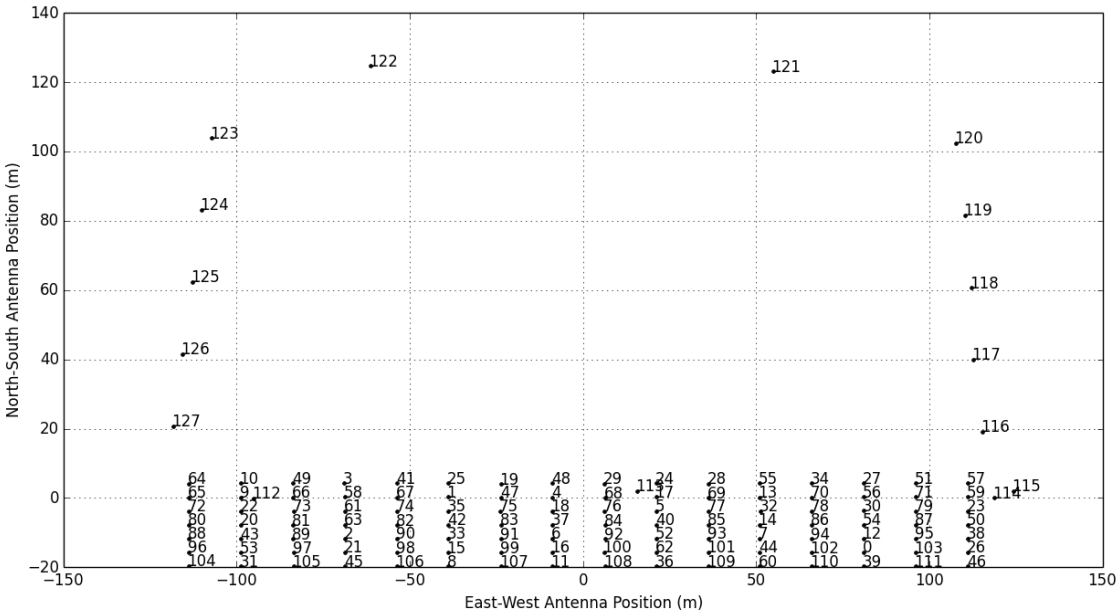

Figure 5.1: The PAPER-128 antenna layout. There are 112 antennas arranged in a grid layout which are used for power spectrum analyses. The addition of 16 outrigger antennas is used to increase *uv*-plane sampling for imaging analyses. We focus solely on the 30 m East/West baselines in our analysis.

	Epoch 1	**Epoch 2**
Dates	Nov 2013 - Jan 2014 JDs 6617-6673	Jan 2014 - March 2014 JDs 6678-6724
LST Range	1h-9h	4h-12h
Number of Baselines	51	79
Number of Days	38	40
Flagged Days	6647,6662,6664,6665,6673	6692,6702,6704,6706,6717,6730
Flagged Antennas	3,7,8,16,19,20,27,28 34,53,56,84,85,96,100 **FX2**: 2,10,15,22,31,33,42, 43,47,58,64,72,91,97,105,107	3,7,16,27,34,56, 57,84,85,100,110

Table 5.1: An overview of the properties of PAPER-128 data analyzed in this chapter. The number of baselines and days is indicative of the final numbers used in the power spectrum analysis, post-flagging. Epoch 1 has many more flagged antennas due to the failure of correlator FX2.

for the PAPER experiment. They also represent the first generation of data assessment techniques that are currently being expanded upon for incorporation into HERA's real-time processing system.

Additionally in this chapter, we process two epochs of PAPER-128 data (both from the first season) and show power spectrum results for each. We do not show results from the second season of data due to increased hardware failure that was experienced at the end of PAPER-128's deployment. As such, the first season of PAPER-128 represents the bulk of the array's sensitivity. Table 5.1 gives an overview of the PAPER-128 data set analyzed in this chapter.

5.2 Quality Assurance

As we saw in Chapter 4, post-processing of PAPER data relies heavily on the array's redundancy and any sources of non-redundancy have the potential to corrupt redundant calibration and power spectrum estimation. These sources of error — which span from the failure of analog or digital components to improper feed installation to the accidental deletion and subsequent recovery of data products — manifest themselves in corrupted data that can be found primarily along two main axes: the time-axis and the antenna-axis.

We next present metrics that we have developed to locate contaminated data along these axes. We use the results of these metrics to remove specific days of data and specific antennas from our analysis prior to calibration in order to ensure robust calibration that is

not influenced by outlier data.

5.2.1 Flagging Julian Dates

To better assess the variation of PAPER-128 data across its two-year deployment (which consists of 7+ individual epochs of data, where an epoch is characterized by the shut-down and subsequent re-starting of the correlator), we first investigate one-dimensional slices of compressed data that span the entire time axis but only one frequency channel. We choose channel 100 for this analysis, as its corresponding frequency of 150 MHz lies in the middle of our band (similar to PAPER-64, PAPER-128's bandwidth consists of 203 frequency channels ranging from 100 to 200 MHz).

We look at the visibility amplitude of each day of data in an epoch as a function of LST, shown as the different colors in Figure 5.2. We designate a reference day to each epoch (a manually chosen "good" day), and flag days that differ from the reference day by more than 20%. After flagging, visibilities show good day-to-day agreement across an epoch (bottom row of Figure 5.2.

An example of data from a particularly "bad" day is shown in Figure 5.3. It is clear that Julian date 2456692 differs dramatically from reference day 2456680, as it contains many corrupted files throughout the day due to failure somewhere in the signal chain. This failure mode is unfortunately often due to the accidental deletion and erroneous recovery of data (much of the season's data was accidentally deleted prior to any of the work in this chapter), as well as hardware failures (e.g., loose cable connections) that result in lost sky fringes. From Table 5.1, we see that a total of 38 days pass our flagging metric for Epoch 1 and 40 days pass for Epoch 2 (both from PAPER-128's first observing season), and a total of five and six days are flagged for each epoch, respectively.

5.2.2 Flagging Antennas

There are a few types of failure modes for data associated with specific antennae. The first is if feeds are accidentally rotated by 90 degrees (or equivalently, the cables for the XX and YY polarizations are swapped). Consequently, visibilities involving a "cross-polarized" antenna exhibit visibility amplitudes that are more weakly correlated (lower amplitude) than what is expected for an XX or YY visibility (and similarly, XY and YX visibilities exhibit too high of an amplitude, or too strong of a correlation). In order to locate potentially cross-polarized antennas, we use the following metric:

$$M_i = \frac{\sum_{j,\nu,t} |V_{ij}|}{(N-1)}, \tag{5.1}$$

where the visibility for baseline ij is summed for every j^{th} antenna, frequency ν, and time t, and N is the number of antennas. We compute this metric for all four polarizations (XX,

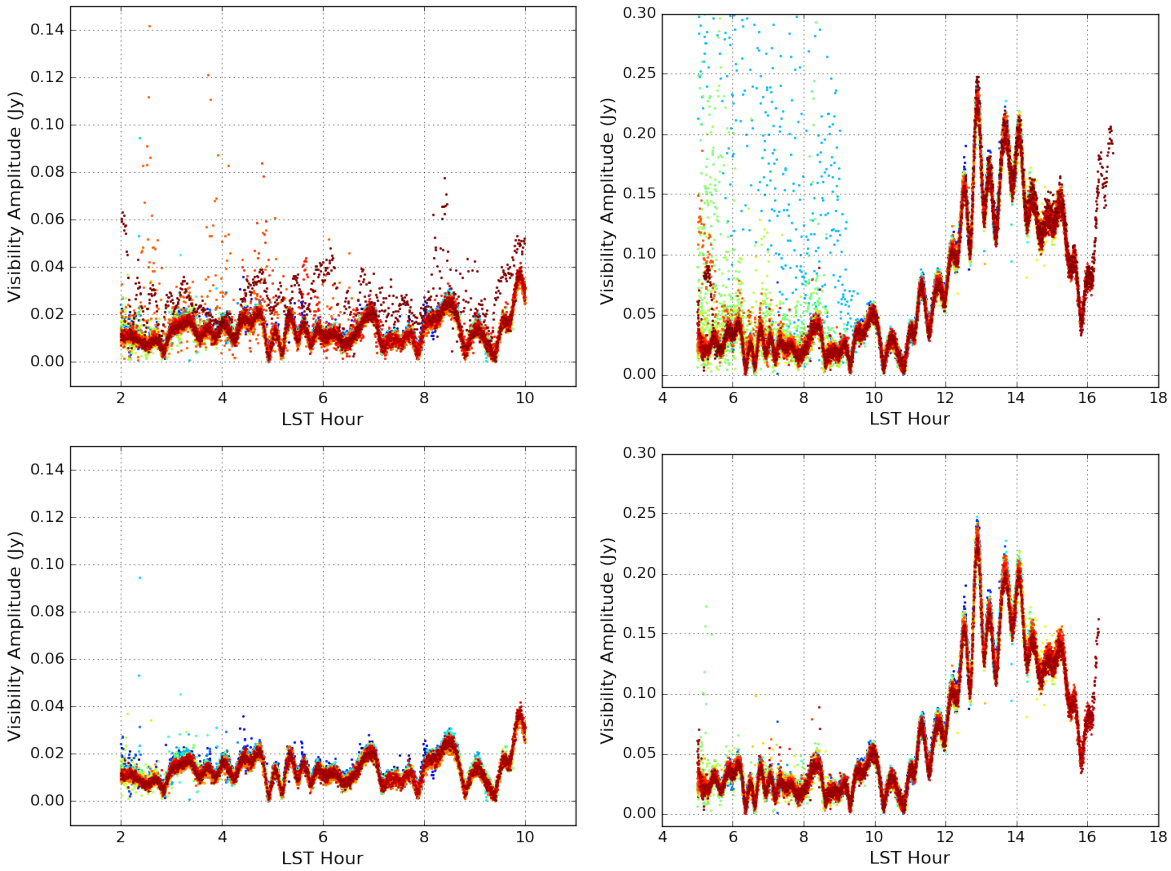

Figure 5.2: Visibility amplitudes as a function of LST for different Julian days of data (colors). The left column shows the data before (top) and after (bottom) the flagging of outlier days for Epoch 1. The right column shows similar data for Epoch 2. After flagging, visibilities show good day-to-day agreement across an epoch.

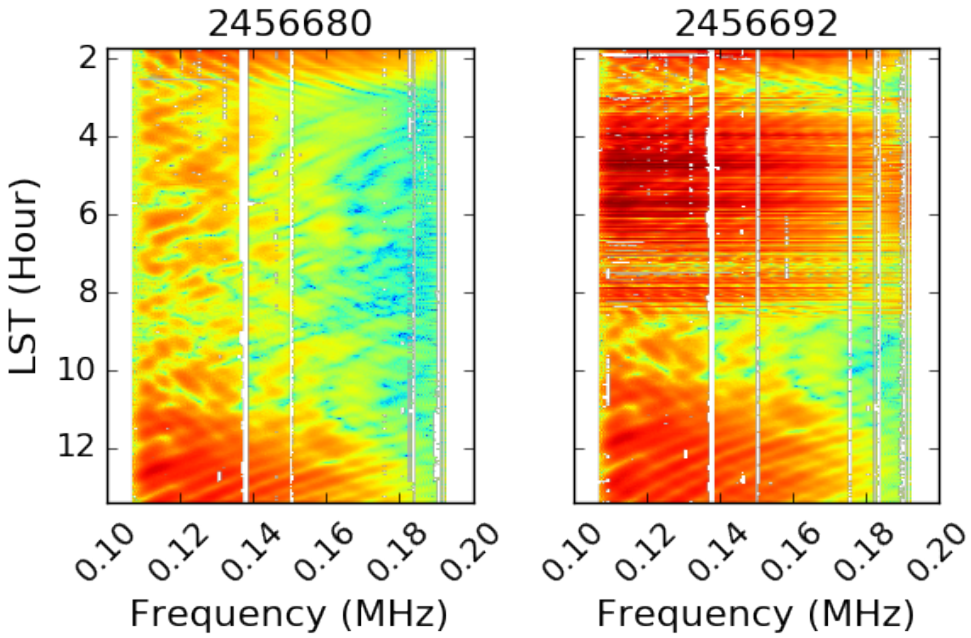

Figure 5.3: Waterfall plots of visibility amplitudes for a "good" reference day (left) in Epoch 2 and "bad" day (right). We exclude corrupted data for specific days found with our metric.

XY, YX, YY). Next, for each night of data we form the polarization fraction quantity:

$$P_i = \frac{M_i^{XY} + M_i^{YX}}{M_i^{XX} + M_i^{YY}}. \tag{5.2}$$

We expect the numerator of this quantity to be higher than the denominator for cross-polarized antennae. In practice, we form polarization fractions for every antenna and flag those with high P_i values as being cross-polarized. We note that this metric works best for long baselines, where XX and XY visibilities are expected to have large differences in signal-to-noise over short distance scales. For short baselines (and large-scales), astrophysical polarization has been shown to be present in all polarization quantities, including the linear ones, thus bringing P_i closer to unity (Lenc et al. 2016).

Using this metric, we find six antennas to be cross-polarized in PAPER-128 observations (antennas 26, 34, 38, 46, 50, and 72). An example of visibilities containing antenna 26 is shown in Figure 5.4, and it is clear that the "X" and "Y" polarizations are swapped for baselines involving that antenna. We fix this issue, re-naming polarization states where necessary, for all the compressed PAPER-128 data before performing other quality assessment tests and calibration.

Another failure mode is an antenna that exhibits low amplitude, a fairly common side effect of equipment malfunction. For example, low power can result from a malfunctioning amplifier or resistance somewhere along the signal chain. A particularly drastic example of power failure is when an entire correlator (FX2) was accidentally shut-off, cutting off

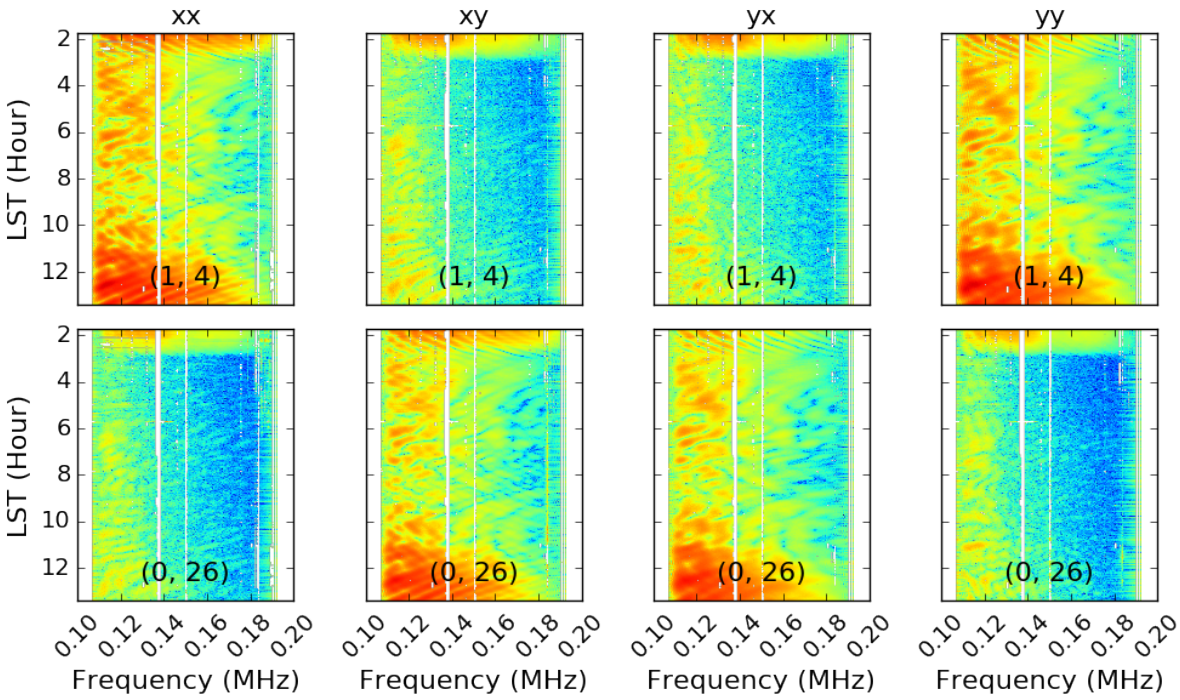

Figure 5.4: Waterfall plots of visibility amplitudes for four different polarizations and two different baselines. Antenna 26 is found to be cross-polarized because its feed was rotated by 90 degrees, and hence its "X" and "Y" polarization states are mislabeled. Equation (5.2) captures visibility amplitudes like the ones shown here in order to automatically detect cross-polarized antennae.

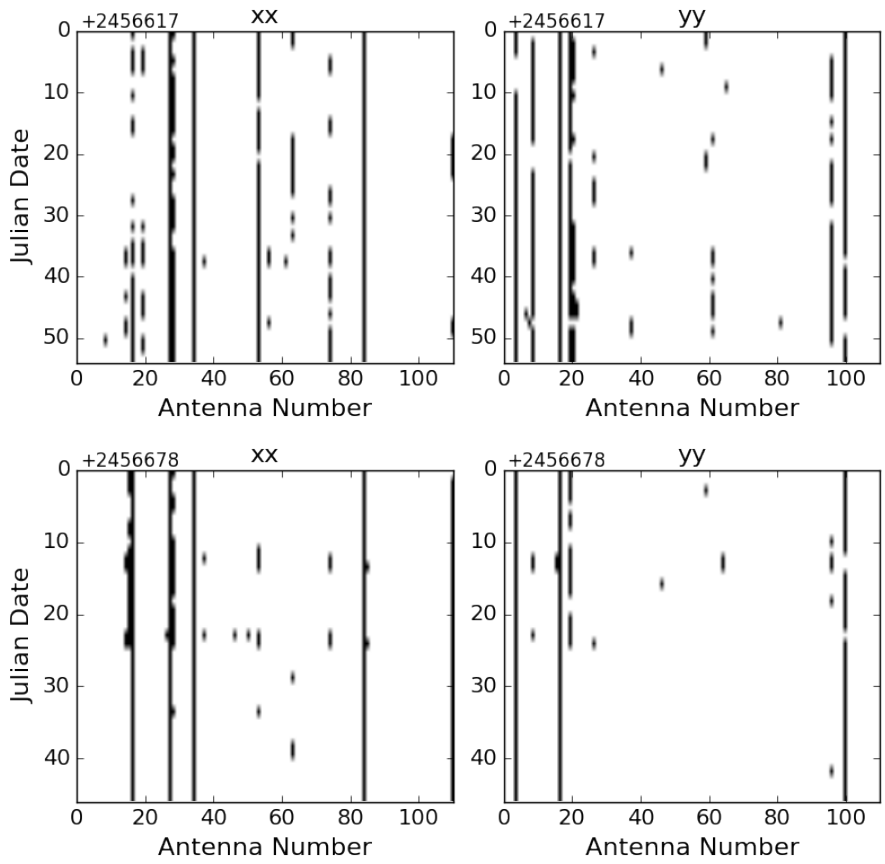

Figure 5.5: Flagged antennas, found using Equation (5.1), are marked in black for each antenna number (x-axis) and Julian date (y-axis). The left column shows flags for XX polarization, and the right column shows flags for YY polarization. The top row shows flags for Epoch 1 and the bottom row shows flags for Epoch 2. We remove antennas that are flagged greater than 50% of the time per epoch.

connections to 16 antennas. The loss of those 16 antennas only affected Epoch 1, and the correlator was then re-started for Epoch 2.

In order to find antennas with unusually low signals, we again employ Equation (5.1), computing the metric for every antenna. Comparing M_i across the entire array can reveal which antennas are consistently misbehaving. We then flag antennas with M_i values that are low by at least 1σ. Figure 5.5 shows these antenna flags across both epochs, and we remove all data associated with antennas that are flagged greater than 50% of the time. Then, we combine the flags for the XX and YY polarizations for each epoch.

Almost all of the flagged antennas listed in Table 5.1 are found via the metric described above. However, we do note that a handful are found manually by visually inspecting visibility data and `Omnical` results. While our metric does a good job identifying antennas with low amplitudes, we found that there were certain antennas (usually one-off instances)

with additional issues that do not fall under any of our previous metrics and would require a more sophisticated metric to be able to find automatically.

Finally, we highlight the importance of flagging antennas prior to redundant calibration by showing `FirstCal` phase solutions for a single antenna without flagging and with flagging (Figure 5.6, top). `Omnical` χ^2 results are also shown (bottom). These results depict how the quality of redundant calibration (the stability of the calibration solutions and level of redundancy) depends on the behavior of the antennas. It is evident that the `FirstCal` solutions are unstable from file to file without any initial antenna flagging (top left). Similarly, higher χ^2 values means that the `Omnical` model visibilities differ substantially from the gain-corrected measured visibilities, meaning redundant calibration is poor. It is obvious that pre-calibration flagging metrics are crucial in order to robustly calibrate PAPER-128 data.

5.3 Data Processing

We process both epochs of PAPER-128 data, closely following the PAPER-64 processing steps outlined in Chapter 3.1.2. One specific difference from the PAPER-64 pipeline is that we aggressively flag RFI prior to delay-filtering in order to prevent low levels of spectral structure from leaking into the EoR window. Because the overall quality of PAPER-128 data is in general lower than PAPER-64, we found that aggressive flagging methods are worthwhile in order to maximize signal-to-noise. To accomplish this, we first flag calibrated visibilities on a per-frequency, per-integration basis based on OMNICAL χ^2 values (using a 6σ deviation cutoff). This masks potentially problematic data identified from the redundant calibration solutions. In addition, we manually flag ten known channels containing RFI.

Delay-filtering proceeds identically to the original PAPER-64 pipeline. We then bin both the foreground-containing and foreground-removed data by LST into separate data sets. For both these data sets, we form Stokes I using Equation (3.1).

Recalling that redundant calibration only solves for internal calibration parameters, absolute calibration remains needed. In the PAPER-64 analysis, a self-calibration method was used to solve for overall gain and phase calibration solutions (immediately after redundant calibration) (Ali et al. 2015). Trusting these solutions, we can then calibrate the PAPER-128 data to the calibrated PAPER-64 data. To do this, we use foreground-containing, LST-binned data for both PAPER-64 and PAPER-128 (where the PAPER-64 data set is already calibrated) and look at a matching fiducial baseline for each. We align both data sets in LST and compute the ratio of the two for every time integration and frequency channel. We then average these ratios over time to yield a bandpass solution.

Mathematically, if the visibility data for PAPER-64 and PAPER-128 are written as the following complex numbers:

$$V_{64} = ae^{i\phi_a} \tag{5.3}$$

$$V_{128} = be^{i\phi_b}, \tag{5.4}$$

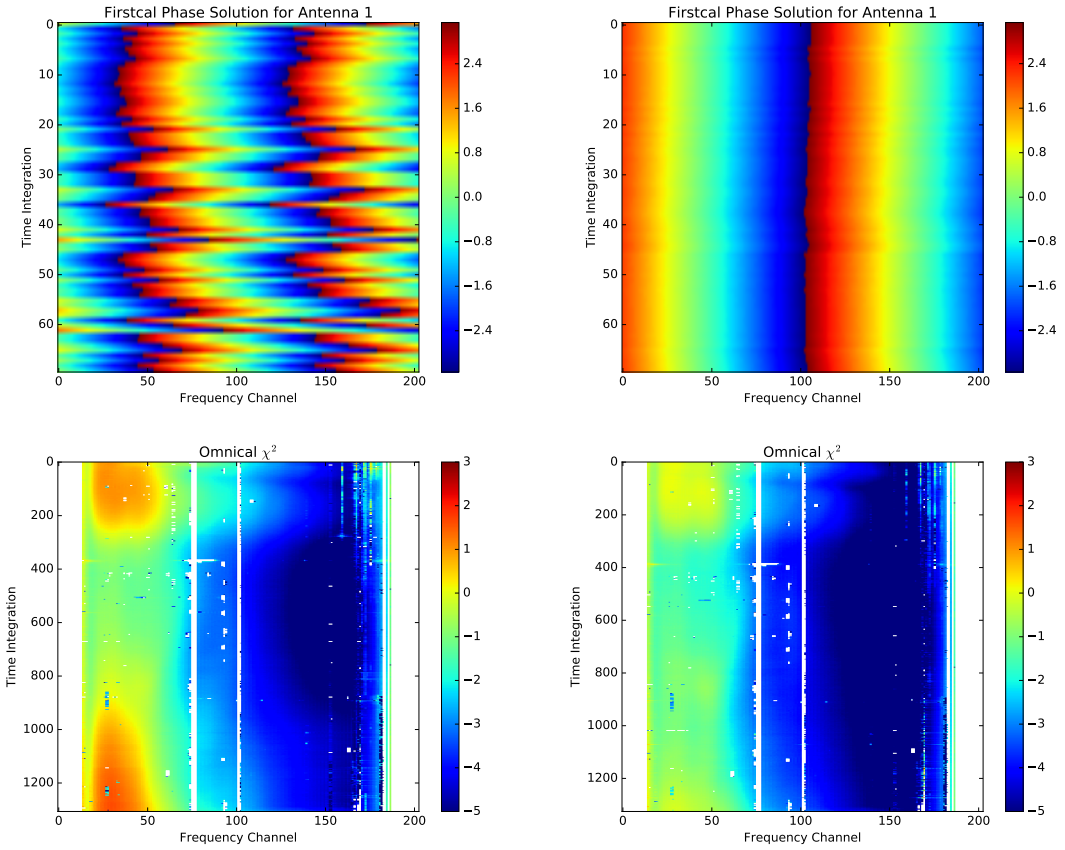

Figure 5.6: `FirstCal` phase solutions for Antenna 1 (Epoch 1, XX polarization) where no antennas are flagged (top left) and some antennas are flagged (via the methods described in Chapter 5.2.2, top right). Omnical χ^2 results are also shown for the two cases (bottom). We do not include any of the 16 dead antennas associated with correlator FX2. It is crucial to flag misbehaving antennas (especially extreme outlier antennas) prior to redundant calibration, motivating the development of automated quality assessment tools prior to the post-processing of data.

we solve for an overall gain correction factor as:

$$f_{gain} = \frac{a}{b} \tag{5.5}$$

and an overall phase factor as:

$$f_{\phi} = \phi_a - \phi_b. \tag{5.6}$$

Combining these two gives a bandpass solution (one number per frequency):

$$f = f_{gain} e^{i f_{\phi}}. \tag{5.7}$$

Using this bandpass, we fit an eighth order polynomial to smooth the solutions and apply this multiplicative factor to the delay-filtered, LST-binned PAPER-64 data. We calibrate each epoch of data separately. We note that this coarse absolute calibration is only a rough calibration, and a more careful, sky-based calibration is recommended for precise results. However, using this simple calibration yields foreground data that agrees with the PAPER-64 data to within $\sim 20\%$.

Finally, the last step of processing is fringe-rate filtering, where an optimal filter is used to combine the data and filter out excess noise. PAPER-128 has an original integration time of 32 s, making the optimal filter length to be ~ 3910 s, slightly longer than that of PAPER-64. While we follow the original PAPER-64 pipeline (pre-power spectrum estimation) closely for our analysis of PAPER-128, it would be interesting to use a simplified pipeline as in Chapter 4 for comparison. However, because of the relatively small quantity and poor quality of PAPER-128 data, our goal for this work is to provide rough power spectrum results as a benchmark comparison rather than precision results.

5.4 Power Spectrum Results

We form power spectrum estimates for both epochs of data and two frequency channel ranges: 139-149 MHz (z=8.9) and 154-164 MHz (z=7.9). Both bands consist of 21 channels (each 0.5 MHz) and are relatively absent of RFI. We focus only on one baseline-separation type, namely all 30 m East/West baselines (of which there are 51 in the first epoch and 79 in the second), and we use eight hours of LST for both (the exact ranges differ though, as there are only five hours of LST overlap between the two epochs). Our power spectrum formalism follows the analysis outlined in Chapter 3 and uses all updated methods regarding bootstrapping and noise sensitivity estimation. Because empirical inverse covariance weighting is shown to be extremely lossy for fringe-rate filtered PAPER data (Chapter 3.2.1) and regularization techniques do not significantly improve power spectrum sensitivity (Chapter 3.2.3), we form unweighted limits ($\mathbf{C} \equiv \mathbf{I}$) in an effort to present straightforward, non-lossy results.

Figure 5.7 shows the power spectrum results for PAPER-128. The two epochs are displayed as columns and the two redshifts as rows. Black and gray data points represent

positive and negative power spectrum values, respectively (calculated as the average power spectrum value over all baseline pairs), and they have 2σ error bars that are calculated from bootstrapping over baselines. The solid green curve is our theoretical prediction of our sensitivity, computed analytically using Equation (3.15). This sensitivity prediction is also plotted as gray shaded regions around the data points.

We do not do a deep investigation of systematics, though it is clear that all power spectra are systematics-dominated, especially at low k, albeit to varying degrees. The level of bias that is present is thought to be highly dependent on the quality of the particular days and antennas that contribute to each LST-binned data set. An investigation into the redundancy of PAPER-128 baselines would be useful to help control the systematics, as shown through the re-analysis of PAPER-64 in Chapter 4. Unfortunately, binning together both epochs of data in order to increase our sensitivity is not straightforward, as each individual LST-binned epoch contains different baselines and LST coverages. In practice, one could re-bin individual days, but this would require a more precise absolute calibration method to be implemented on a per-day basis prior to LST-binning. We hence present only rough limits from each epoch individually and note that the varying systematics between epochs and redshifts are an interesting indication of how sensitive the data is to non-redundancies.

Our most sensitive upper limits that are consistent with noise are $(194.8\,\mathrm{mK})^2$ at $k = 0.30\,h$ Mpc^{-1} and $z = 7.9$ (from Epoch 2) and $(130.3\,\mathrm{mK})^2$ at $k = 0.19\,h$ Mpc^{-1} and $z = 8.9$ (from Epoch 1). The limits at these two redshifts are within a factor of $\sim$ 2-3 from those of PAPER-64. Although taken at face value the numbers themselves represent more stringent limits than those of PAPER-64, we note uncertainty in the form of possible loss associated with our rough absolute calibration method (i.e., a 20% calibration error becomes a 40% error on a power spectrum result). These results have not been subjected to the same scrutiny as those of PAPER-64 and hence we regard Figure 4.13 as the most accurate and stringent power spectrum limits from the PAPER experiment. While it's certainly possible that combining multiple epochs of PAPER-128 data together (and using more baseline types) may yield better limits, we do not think this analysis is worthwhile as the second season of data is severely dominated by systematics. Instead, we look forward to exciting future results from HERA, knowing that many of the techniques developed for both PAPER-64 and PAPER-128 are being incorporated and continuously improved upon.

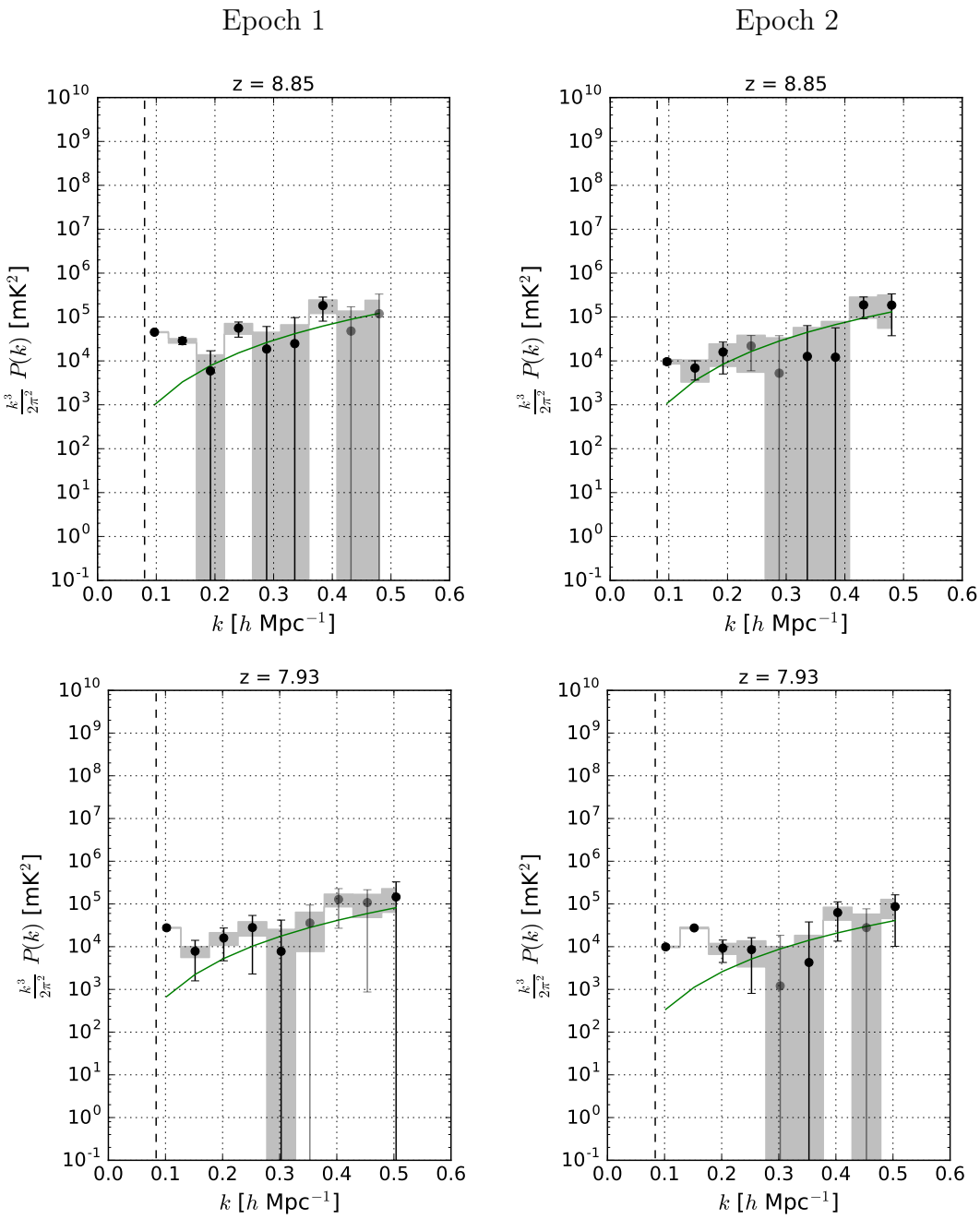

Figure 5.7: Power spectrum results for PAPER-128 season 1 data for two redshifts (rows) and two epochs (columns), using one baseline separation-type only (30 m East/West baselines). Black and gray points represent positive and negative power spectrum values, respectively, with 2σ error bars determined from bootstrapping. The 2σ theoretical noise sensitivity prediction is shown in green. Gray shaded regions correspond to theoretical errors on each data point.

Chapter 6

Future Work

6.1 Eigenspectrum Characterization for Signal Loss

In Chapter 2, we discussed how weighting data by an empirical covariance can lead to overfitting of the cosmological signal, resulting in signal loss in a 21 cm power spectrum. We investigated the relationship between an empirical covariance (namely, its convergence to the true covariance) and the number of independent samples in a data set, and we found, as expected, that convergence rates are fastest when averaging over large numbers of data realizations (Figure 2.4). Additionally, we began to explore the relationship between eigenvector convergence and the shape of an eigenspectrum, finding that steep eigenspectra (with eigenmodes that have eigenvalues that differ greatly from each other) converge the fastest (Figure 2.5).

In this section, we expand our preliminary analysis characterizing covariance eigenspectra and outline future work on this topic. Ideally, we would like to define a metric that relates an eigenspectrum to the amount of potential signal loss it will incur when weighting by it. This would be useful because it would provide a method to estimate, *a priori*, how much signal loss may result given a particular covariance. To define this metric, we must understand how different properties of an eigenspectrum, including its shape and its error bars, affect its convergence.

Let's first begin by looking at the rate of convergence of empirical eigenvectors to their true forms. In Chapter 2.2.3 we described how eigenvectors of an empirical covariance are strongly coupled to the data (which leads to loss), and the rate at which they can effectively "un-couple" and approach their true shapes affects how much signal loss results. A steep eigenspectrum, where each eigenmode has a distinct eigenvalue, can more easily de-tangle its eigenvectors, whereas a flat eigenspectrum contains many degenerate eigenvectors and thus requires many more data realizations to "un-do" the coupling.

To quantify this effect, we simulate a noise-like EoR signal similarly to Chapter 2.2.3, where we construct a covariance whose true form is a diagonal matrix in the Fourier domain with eigenvalues spanning some range. The Fourier-transform of this covariance is the covariance of the EoR in the frequency domain, or $\mathbf{C}_{\mathrm{EoR}}$. For different numbers of realizations,

we draw random EoR signals that are consistent with $\mathbf{C}_{\text{EoR}}$ and compute the combined empirical covariance by averaging over the realizations. On the vertical axis of Figure 6.1, we plot a convergence metric that describes the covariance's empirical eigenvectors $\widehat{\mathbf{v}}$ compared to the true eigenvectors $\mathbf{v}$, namely Equation (2.13), or:

$$\varepsilon(\widehat{\mathbf{v}}) \equiv \sqrt{\sum_{i}^{N_f} |\mathbf{v} - \widehat{\mathbf{v}}|_i^2}, \tag{6.1}$$

where small $\varepsilon(\widehat{\mathbf{v}})$ denotes faster convergence.

We run this simulation over a range of realizations and for true covariances whose eigenvalues range from a span of one order of magnitude to five (i.e., a relatively flat spectrum to steep). Most importantly, the horizontal axis of Figure 6.1 captures two important properties of an eigenspectrum that we would like to relate to convergence: uncertainty (errors) of the spectrum, and relative slope of the spectrum. The reason we choose to look at these two properties is because together they describe how much "overlap" there is between eigenmodes. More overlap (i.e., a flat spectrum and/or large errors) means that it is more difficult for the empirical eigenvectors to de-couple from each other, slowing convergence.

In practice, the eigenvalue errors in our simulation come from the distribution of empirical eigenvalues obtained from running N total simulations, where N is large. The quantity we then form for every i^{th} eigenmode (except the smallest one, when ordered from largest to smallest) is a fractional error (FE):

$$FE = \frac{\sigma_i + \sigma_{i+1}}{2(\lambda_i - \lambda_{i+1})}, \tag{6.2}$$

where σ are eigenvalue errors and λ are eigenvalues. This quantity thus captures both how much overlap there is between eigenvalues (the numerator, or average of two adjacent errors) and how steep the spectrum is (the denominator).

From Figure 6.1, we see that the fractional error quantity we defined is closely related to eigenvector convergence, with eigenspectra with smaller errors converging fastest. Looking at the different colors, we also see that eigenspectra with steeper slopes converge fastest, which is in alignment with our discussion above. We have therefore showed how eigenvector convergence depends on properties of an eigenspectrum, and our next step is to relate convergence to signal loss.

Up until now, we have implied that a converged covariance (and more specifically, converged eigenvectors) is synonymous with minimal/no loss (i.e., what we would like to achieve). This would mean that given an eigenspectrum, all we would have to do is calculate the fractional error metric as defined above in order to determine how well-converged it is, and therefore how much signal loss there is. However, this relationship is complicated by the fact that *signal loss can still result even if eigenvectors are converged*. For example, we can zoom in to look at the resulting power spectrum for the simulation with the steepest slope (slope of five orders of magnitude, or the black points in Figure 6.1). Figure 6.2 shows this power spectrum for different numbers of realizations (dashed colors) compared to the true

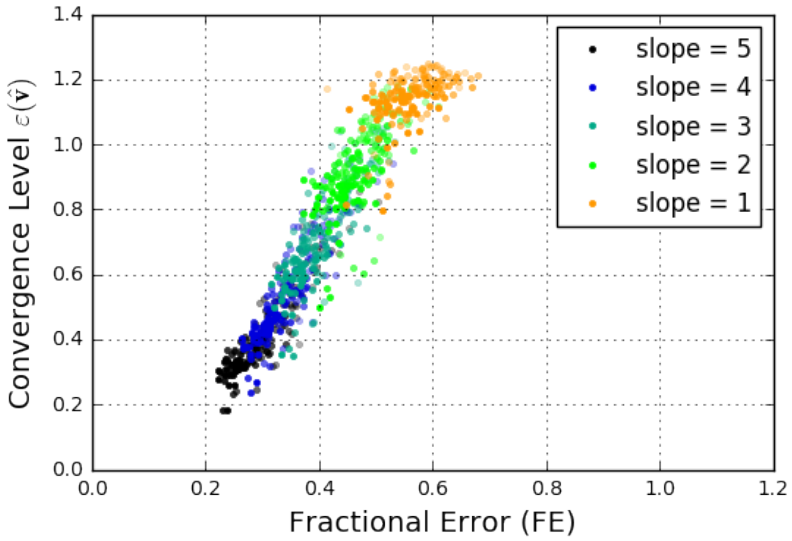

Figure 6.1: The convergence level (y-axis), as defined by Equation (6.1), of empirically estimated eigenvectors compared a fractional error metric of the eigenspectrum (x-axis) which takes into account both how well-defined and how steep the spectrum is. The different colors denote the number of orders of magnitude that the true eigenspectrum spans. It appears that the defined fractional error quantity is closely related to eigenvector convergence, with smaller errors and steeper spectrum slopes converging fastest.

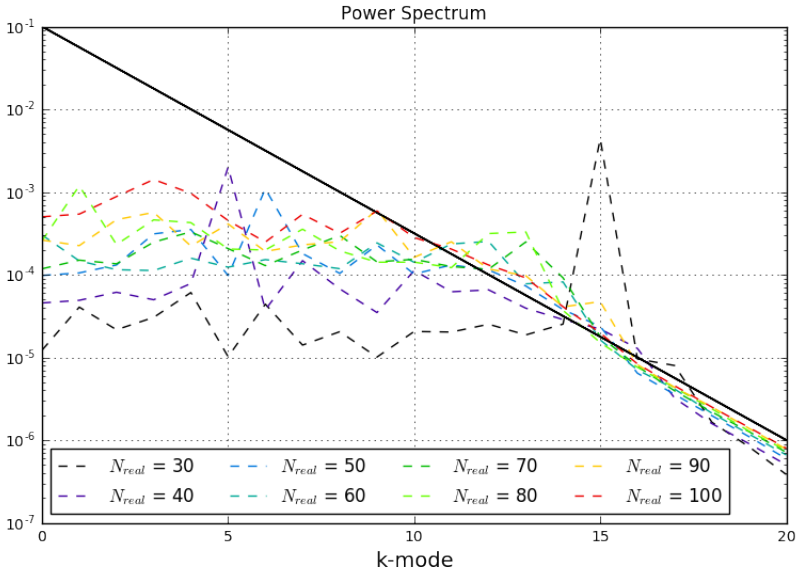

Figure 6.2: Resulting power spectra for different numbers of realizations (dashed colors) compared to the true power spectrum (solid black), which spans five orders of magnitude. Although increasing the number of realizations decreases loss as expected, there is still obvious signal loss at high-amplitude k-modes.

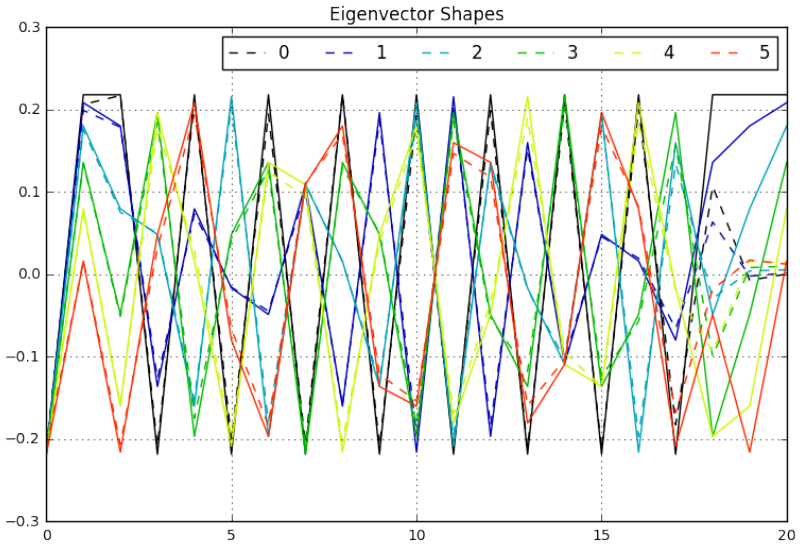

Figure 6.3: Eigenvector shapes for a few of the first modes (different colors) for the simulation corresponding to $N_{real} = 100$ in Figure 6.2. The empirical eigenvectors (dashed) are in general converged to their true forms (solid), implying that there should be minimal signal loss. However, we see significant signal loss in Figure 6.2.

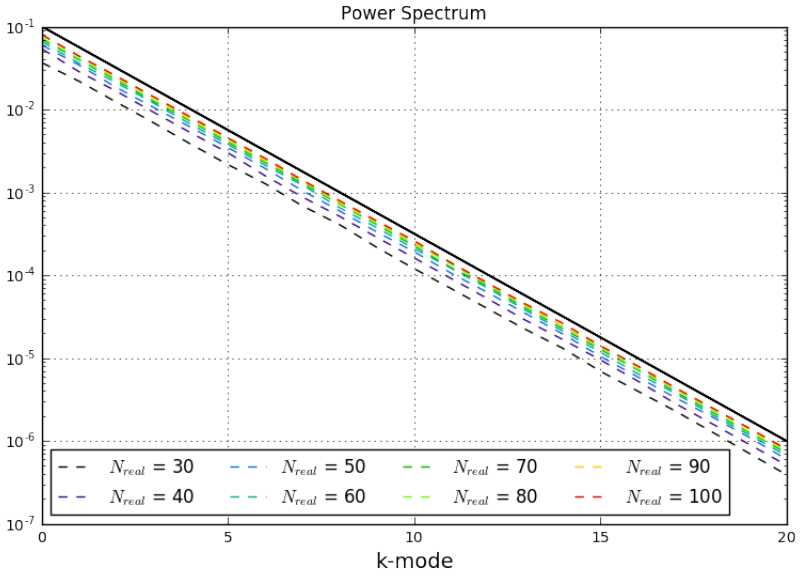

Figure 6.4: Resulting power spectra for different numbers of realizations (dashed colors) compared to the true power spectrum (solid black), which spans five orders of magnitude. This plot differs from Figure 6.2 only in window function shapes; here our window functions are set to estimate power spectrum modes independently from each other. We find that by de-tangling window function modes, we avoid power spectrum deviations such as in Figure 6.2 due to information from high k-modes dragging down low k-modes.

power spectrum (solid black). Although increasing the number of realizations decreases loss as expected, there is still obvious signal loss at low k-modes. What is unusual about this, is that when we take a closer look at some of the eigenvector shapes in Figure 6.3 (looking at just $N_{real} = 100$, or the highest number of realizations), we see that the empirical modes (dashed colors) *are* in general converged to their true forms (solid colors). Taken at face value, this implies that there should be minimal signal loss, but this is curiously not the case.

Deeper investigations reveal that the origin of the loss has to do not with eigenvector convergence, but rather the shape of the window functions that are used in power spectrum estimation. Recall that the window function relates our estimated power spectrum $\widehat{\mathbf{P}}$ to the true power spectrum $\mathbf{P}$ (Chapter 2.2.1):

$$\widehat{\mathbf{P}} = \mathbf{W}\mathbf{P}. \tag{6.3}$$

Window functions describe how different k-modes are related to each other — peaky window functions draw power spectrum information from independent k-modes, while long-tailed window functions mean that power spectrum modes draw information from multiple modes. Hence, window functions have the potential to entangle modes in a similar way that flat (or large-error) eigenspectra can. Indeed in our previous simulation we chose a normalization matrix such that our window functions for the low k-modes were heavily influenced by information from high k-modes, essentially dragging down the power spectrum at low k's.

As a useful check, we can force our window functions to be $\mathbf{W} = \mathbf{I}$, or the identity matrix, so that each k-mode is estimated independent from each other. In practice, this is done by choosing our normalization matrix $\mathbf{M}$, where $\mathbf{W} = \mathbf{MG}$ (Chapter 3.4.1), in such a way so that $\mathbf{W} = \mathbf{I}$. In doing so, our resulting power spectrum estimates deviate much less from the true power spectrum (Figure 6.4). From this analysis, it is clear that power spectrum deviations are dependent on multiple factors, with window functions playing an unexpected role. In the future, it will be necessary to balance the advantages of wide window functions (which are typically chosen to minimize vertical power spectrum errors) with the potential power spectrum deviation that can result (the amount of which is dependent in part on the steepness of the eigenspectrum).

In this section we have seen how the convergence of empirical eigenvectors is related to both the slope/shape of an eigenspectrum and how well defined the spectrum is (which is dependent on number of realizations). We have also shown how power spectrum estimates are affected by the choice of window function used in power spectrum estimation. As a summary, the following general relationships have been identified:

- Steeper eigenspectrum = more signal loss
- Steeper eigenspectrum = faster eigenvector convergence
- More data realizations = less signal loss
- More data realizations = faster eigenvector convergence

- Smaller eigenvalue errors = faster eigenvector convergence
- Wider window functions = more signal loss

Future work is needed in order to tie these individual relationships together to form one easily computable metric that can be used to map eigenspectrum properties to signal loss. The goal for this characterization is to intimately understand the interplay between all the effects of signal loss in order to be able to bridge data properties into a theoretical estimate for loss. For example, one hopes to be able to look at a HERA eigenspectrum that's computed solely from data, place some error bars on the spectrum based on a theoretical estimate of noise and how many data samples are used, and then characterize the spectrum and use this information to inform decisions about power spectrum weighting. Developing an intuition and quantitative approach for estimating signal loss associated with a particular covariance and eigenspectrum will be extremely powerful for assessing power spectrum results and will save computational time by providing a way to estimate loss without having to perform expensive simulations that quantify loss after-the-fact.

As the 21 cm community moves towards more rigorous analysis techniques and EoR sensitivities, the work in this section, as well as future work deepening our understanding of empirical covariances and signal loss, complement a general desire for new techniques that help shape and motivate power spectrum analysis choices.

6.2 HERA

The full HERA array is nearing its completion in construction, and preliminary HERA analyses are ongoing. While this thesis focuses on power spectrum methods as applied to PAPER, the lessons we have learned are already influencing HERA analysis. More specifically, we showed how aggressive fringe-rate filtering of PAPER data leads to lossy, inaccurate power spectra; consequently HERA's initial power spectrum results will use data that is not fringe-rate filtered. With PAPER we illustrated how empirical inverse covariance weighting is coupled with substantial loss and how regularization of those covariances does not seem to carry a significant advantage over uniform weighting; hence HERA analysis is currently focused on forming unweighted power spectra.

Similarly, investigations of bootstrapping and error estimation in PAPER have influenced a HERA power spectrum pipeline that now computes power spectrum errors in a variety of ways, including via bootstrapping, propagation, histogramming, noise realizations, and simulations. Pure noise simulations, which we first introduced as a way to help verify PAPER's sensitivity, are now routinely processed as part of HERA's validation. PAPER discoveries regarding the effect of correlated data samples on error estimation have inspired deeper investigations into other subtleties surrounding correlated noise and bootstrapping. And discussions of systematics and the implementation of jackknife tests for PAPER have contributed in setting a new standard for HERA in terms of establishing useful, routine tests.

In general, the work in this thesis has helped to raise the bar for thorough HERA analysis. It has inspired more detailed documentation, more well-defined tests, more eyes on data, and a more urgent desire for nuanced understandings, cross-checks, and validation. With this strong foundation, HERA is positioned well to produce robust, believable results and tackle challenges that lay ahead.

Chapter 7

Conclusion

The work presented in this thesis can be summarized by the following:

- The original 21 cm power spectrum results from the 64-element configuration of PAPER, first presented in A15, have been found to suffer from cosmological signal loss associated with the use of empirical covariances. Qualitative and quantitative investigations of the origin of this loss have shaped a deeper understanding of why signal loss arises, how to mitigate it, and how data analysis choices affect loss and power spectrum sensitivities. From this work emerges a new method to quantify signal loss which sets a standard across the field as a whole in terms of performing detailed calculations and assessments of loss.

- Numerous other power spectrum analysis errors that affect both Parsons et al. (2014) and A15 have been discovered and revised. These errors mostly concern error estimation techniques used by PAPER. While both the empirical and theoretical errors were underestimated in the original analyses, updated methods are used to produce the power spectrum results throughout this work. At a broader level, this work has influenced the development of additional error estimation techniques for both PAPER and HERA analyses in a push for more robust tests and validation.

- Using a simplified pipeline and incorporating all updated methods, PAPER-64 places new 21 cm upper limits across a range of redshifts that are competitive with results from other experiments. Most notably, novel components of this analysis include investigations of PAPER's redundancy, comparisons with sky simulations, and analyses of jackknife tests.

- The 128-element configuration of PAPER has been found to present unique challenges in terms of data quality. Algorithms have been developed in order to locate and remove contaminated data, which are now standard quality-check measures that have been incorporated into HERA's real-time processing system. However, because the quality of PAPER-128 data is poor compared to that of PAPER-64, we only present rough

power spectrum results for two epochs of data, which are found to be competitive with the revised PAPER-64 results.

- Initial HERA analyses are ongoing and signal loss characterization continues in an effort to build intuition behind our power spectrum analysis choices. Preliminary work suggests that the shape and accuracy of an eigenspectrum of a covariance is closely related to the amount of signal loss that is incurred when weighting data by the covariance. Future work in characterizing eigenspectra will help in understanding the interplay between convergences, weightings, and signal loss.

Taken as individual parts, the work in this thesis represents a collection of subtle lessons concerning 21 cm power spectrum estimation. But looked at as a whole, it tells a bigger story that serves as a reminder of the uncertainty, challenges, and unpredictability of science in general. What began as a promising future for PAPER-128 (following a field-leading result from PAPER-64) led to the discovery of errors in PAPER-64 and subsequently an unanticipated retraction and revision. However, it is important to remember that good science — regardless of the final outcome — is science that is honest, careful, repeatable, and communicated. The broader story told in this thesis is, in some ways, one that is very common (but not often talked about) in science. It is clear that the issues we have found are not unique to any one experiment and, although unexpected, are pushing scientific progress in a positive way. As we close the chapter on PAPER, the lessons we have learned will continue to influence the path ahead, as unpredictable as it may be. What is predictable, though, is that we are now equipped with better tools and deeper understandings, and the field of 21 cm cosmology has a lot to look forward to.

Bibliography

Ade, P., Bock, J., Bowden, M., et al. 2008, ApJ, 674, 22
Ade, P. A. R., Ahmed, Z., Aikin, R. W., et al. 2017, Phys. Rev. D, 96, 102003
Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2018, ApJ, 863, 201
Andrae, R. 2010, ArXiv e-prints, arXiv:1009.2755 [astro-ph.IM]
Araujo, D., Bischoff, C., Brizius, A., et al. 2012, ApJ, 760, 145
Barkana, R. 2009, MNRAS, 397, 1454
—. 2018, Nature, 555, 71
Barkana, R., & Loeb, A. 2008, MNRAS, 384, 1069
Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. 2016, ApJ, 833, 102
Becker, R. H., Fan, X., White, R. L., et al. 2001, AJ, 122, 2850
Bernardi, G., de Bruyn, A. G., Brentjens, M. A., et al. 2009, A&A, 500, 965
Bernardi, G., de Bruyn, A. G., Harker, G., et al. 2010, A&A, 522, A67
Bernardi, G., Greenhill, L. J., Mitchell, D. A., et al. 2013, ApJ, 771, 105
Bernardi, G., Zwart, J. T. L., Price, D., et al. 2016, MNRAS, 461, 2847
BICEP2 Collaboration, Keck Array Collaboration, Ade, P. A. R., et al. 2016, ApJ, 833, 228
Bischoff, C., Brizius, A., Buder, I., et al. 2011, ApJ, 741, 111
Bond, J. R., Jaffe, A. H., & Knox, L. 1998, Phys. Rev. D, 57, 2117
Bowman, J. D., Morales, M. F., & Hewitt, J. N. 2009, ApJ, 695, 183
Bowman, J. D., & Rogers, A. E. E. 2010, Nature, 468, 796
Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, Nature, 555, 67
Burns, J. O., Lazio, J., Bale, S., et al. 2012, Advances in Space Research, 49, 433
Chang, T.-C., Pen, U.-L., Bandura, K., & Peterson, J. B. 2010, Nature, 466, 463
Chapman, E., Zaroubi, S., Abdalla, F. B., et al. 2016, MNRAS, 458, 2928
Chapman, E., Abdalla, F. B., Harker, G., et al. 2012, MNRAS, 423, 2518
Chiang, H. C., Ade, P. A. R., Barkats, D., et al. 2010, ApJ, 711, 1123
Crites, A. T., Henning, J. W., Ade, P. A. R., et al. 2015, ApJ, 805, 36
Das, S., Sherwin, B. D., Aguirre, P., et al. 2011a, Physical Review Letters, 107, 021301
Das, S., Marriage, T. A., Ade, P. A. R., et al. 2011b, ApJ, 729, 62
Datta, A., Bowman, J. D., & Carilli, C. L. 2010, ApJ, 724, 526

Datta, K. K., Jensen, H., Majumdar, S., et al. 2014, MNRAS, 442, 1491
de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247
DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001
Dillon, J. S., Liu, A., & Tegmark, M. 2013, Phys. Rev. D, 87, 043005
Dillon, J. S., & Parsons, A. R. 2016, ApJ, 826, 181
Dillon, J. S., Liu, A., Williams, C. L., et al. 2014, Phys. Rev. D, 89, 023002
Dillon, J. S., Liu, A., Williams, C. L., et al. 2014, Phys. Rev. D, 89, 023002
Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015a, Phys. Rev. D, 91, 123011
Dillon, J. S., Tegmark, M., Liu, A., et al. 2015b, Phys. Rev. D, 91, 023002
Dodelson, S. 2003, Modern Cosmology (San Diego, CA: Academic Press)
Dodelson, S., & Schneider, M. D. 2013, Phys. Rev. D, 88, 063537
Efron, B., & Tibshirani, R. J. 1993, An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability No. 57 (Boca Raton, Florida: Chapman & Hall)
Ewall-Wice, A., Dillon, J. S., Liu, A., & Hewitt, J. 2017, MNRAS, 470, 1849
Fialkov, A., Barkana, R., Pinhas, A., & Visbal, E. 2014, MNRAS: Letters, 437, L36
Fialkov, A., Barkana, R., & Visbal, E. 2014, Nature, 506, 197
Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, Phys. Rep., 433, 181
Ghosh, A., Bharadwaj, S., Ali, S. S., & Chengalur, J. N. 2011, MNRAS, 418, 2584
Greig, B., Mesinger, A., & Pober, J. C. 2015, ArXiv e-prints, arXiv:1509.02158
Haiman, Z., & Knox, L. 1999, ASP Conf. Ser., 181, 227
Harker, G., Zaroubi, S., Bernardi, G., et al. 2009, MNRAS, 397, 1138
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Hazelton, B. J., Morales, M. F., & Sullivan, I. S. 2013, ApJ, 770, 156
Hickish, J., Abdurashidova, Z., Ali, Z., et al. 2016, Journal of Astronomical Instrumentation, 05
Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19
Högbom, J. A. 1974, A&AS, 15, 417
Jacobs, D. C., Pober, J. C., Parsons, A. R., et al. 2015, ApJ, 801, 51
Jacobs, D. C., Hazelton, B. J., Trott, C. M., et al. 2016, ApJ, 825, 114
Jaynes, E. 1968, IEEE Transactions on Systems Science and Cybernetics, 4, 227
Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2008, MNRAS, 389, 1319
Joachimi, B. 2017, MNRAS, 466, L83
Joseph, R. C., Trott, C. M., & Wayth, R. B. 2018, AJ, 156, 285
Keating, G. K., Marrone, D. P., Bower, G. C., et al. 2016, ApJ, 830, 34
Kerrigan, J., Pober, J., Ali, Z., et al. 2018, ApJ, 864
Kohn, S. A., Aguirre, J. E., Nunhokee, C. D., et al. 2016, ApJ, 823, 88
Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Proceedings of Science, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 1
Lenc, E., Gaensler, B. M., Sun, X. H., et al. 2016, ApJ, 830, 38
lglewicz, B., & Hoaglin, D. C. 1993, Volume 16: How to Detect and Handle Outliers (The ASQC Basic References in Quality Control: Statistical Techniques)
Liu, A., & Parsons, A. R. 2016, MNRAS, 457, 1864

Liu, A., Parsons, A. R., & Trott, C. M. 2014a, Phys. Rev. D, 90, 023018
—. 2014b, Phys. Rev. D, 90, 023019
Liu, A., & Tegmark, M. 2011, Phys. Rev. D, 83, 103006
Liu, A., Tegmark, M., Bowman, J., Hewitt, J., & Zaldarriaga, M. 2009, MNRAS, 398, 401
Liu, A., Tegmark, M., Morrison, S., Lutomirski, A., & Zaldarriaga, M. 2010, MNRAS, 408, 1029
Loeb, A., & Furlanetto, S. 2013, The First Galaxies in the Universe (Princeton University Press)
Lonsdale, C., J. Cappallo, R., F. Morales, M., et al. 2009, Proceedings of the IEEE, 97, 1497
Masui, K. W., Switzer, E. R., Banavar, N., et al. 2013, ApJ, 763, L20
Mellema, G., Koopmans, L. V. E., Abdalla, F. A., et al. 2013, Experimental Astronomy, 36, 235
Mesinger, A., Furlanetto, S., & Cen, R. 2011, MNRAS, 411, 955
Miralda-Escude, J., Haehnelt, M., & Rees, M. J. 2000, The Astrophysical Journal, 530, 1
Monsalve, R. A., Rogers, A. E. E., Bowman, J. D., & Mozdzen, T. J. 2017, The Astrophysical Journal, 847, 64
Moore, D. F., Aguirre, J. E., Parsons, A. R., Jacobs, D. C., & Pober, J. C. 2013, ApJ, 769, 154
Moore, D. F., Aguirre, J. E., Kohn, S. A., et al. 2017, ApJ, 836, 154
Morales, M. F., Bowman, J. D., & Hewitt, J. N. 2006, ApJ, 648, 767
Morales, M. F., & Wyithe, J. S. B. 2010, ARA&A, 48, 127
Neben, A. R., Bradley, R. F., Hewitt, J. N., et al. 2016, ApJ, 826, 199
Nunhokee, C. D., Bernardi, G., Kohn, S. A., et al. 2017, ApJ, 848, 47
Offringa, A. R., de Bruyn, A. G., & Zaroubi, S. 2012, MNRAS, 422, 563
Paciga, G., Albert, J. G., Bandura, K., et al. 2013, MNRAS, 433, 639
Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. 2016, MNRAS, 460, 1567
Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012, ApJ, 753, 81
Parsons, A., Backer, D., Chang, C., et al. 2006, in 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, 2031
Parsons, A. R., & Backer, D. C. 2009, The Astronomical Journal, 138, 219
Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51
Parsons, A. R., Pober, J. C., Aguirre, J. E., et al. 2012, ApJ, 756, 165
Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, AJ, 139, 1468
Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106
Patil, A. H., Yatawatta, S., Zaroubi, S., et al. 2016, MNRAS, 463, 4317
Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65
Patra, N., Subrahmanyan, R., Sethi, S., Udaya Shankar, N., & Raghunathan, A. 2015, ApJ, 801, 138
Patra, N., Parsons, A. R., DeBoer, D. R., et al. 2018, Experimental Astronomy, 45, 177
Paz, D. J., & Sánchez, A. G. 2015, MNRAS, 454, 4326
Pearson, D. W., & Samushia, L. 2016, MNRAS, 457, 993
Pen, U. L., Wu, X. P., & Peterson, J. 2004, ArXiv e-prints, astro-ph/0404083

Petrovic, N., & Oh, S. P. 2011, MNRAS, 413, 2103
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Pober, J. C., Greig, B., & Mesinger, A. 2016a, MNRAS, 463, L56
Pober, J. C., Parsons, A. R., Jacobs, D. C., et al. 2012, AJ, 143, 53
Pober, J. C., Parsons, A. R., Aguirre, J. E., et al. 2013, ApJLetters, 768, L36
Pober, J. C., Parsons, A. R., DeBoer, D. R., et al. 2013, AJ, 145, 65
Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, ApJ, 782, 66
Pober, J. C., Ali, Z. S., Parsons, A. R., et al. 2015, ApJ, 809, 62
Pober, J. C., Hazelton, B. J., Beardsley, A. P., et al. 2016b, ApJ, 819, 8
Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766
Pritchard, J. R., & Loeb, A. 2010, Phys. Rev. D, 82, 023006
—. 2012, Reports on Progress in Physics, 75, 086901
Quenouille, M. H. 1949, Ann. Math. Statist., 20, 355
Robertson, B. E., Ellis, R. S., Furlanetto, S. R., & Dunlop, J. S. 2015, ApJ, 802, L19
Roshi, D. A., & Perley, R. A. 2003, in Astronomical Society of the Pacific Conference Series, Vol. 306, New technologies in VLBI, ed. Y. C. Minh, 109
Santos, M. G., Cooray, A., & Knox, L. 2005, ApJ, 625, 575
Sellentin, E., & Heavens, A. F. 2016, MNRAS, 456, L132
Sherwin, B. D., van Engelen, A., Sehgal, N., et al. 2017, Phys. Rev. D, 95, 123529
Slatyer, T. R., & Wu, C.-L. 2018, Phys. Rev. D, 98, 023013
Sokolowski, M., Tremblay, S. E., Wayth, R. B., et al. 2015, PASA, 32, e004
Sullivan, I. S., Morales, M. F., Hazelton, B. J., et al. 2012, ApJ, 759, 17
Switzer, E. R., Chang, T.-C., Masui, K. W., Pen, U.-L., & Voytek, T. C. 2015, ApJ, 815, 51
Switzer, E. R., Masui, K. W., Bandura, K., et al. 2013, MNRAS, 434, L46
Taylor, A., & Joachimi, B. 2014, MNRAS, 442, 2728
Tegmark, M. 1997, Phys. Rev. D, 55, 5895
Thompson, A. R., Moran, J. M., & Swenson, Jr., G. W. 2001, Interferometry and Synthesis in Radio Astronomy, 2nd Edition (New York: Wiley)
Thyagarajan, N., Udaya Shankar, N., Subrahmanyan, R., et al. 2013, ApJ, 776, 6
Thyagarajan, N., Jacobs, D. C., Bowman, J. D., et al. 2015a, ApJ, 807, L28
—. 2015b, ApJ, 804, 14
Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, 7
Trott, C. M., & Wayth, R. B. 2017, Publications of the Astronomical Society of Australia, 34, e061
Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101
Trott, C. M., Pindor, B., Procopio, P., et al. 2016, ApJ, 818, 139
Tukey. 1958, Ann. Math. Statist., 29, 614
van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
Vedantham, H., Shankar, N. U., & Subrahmanyan, R. 2012, ApJ, 745, 176
Vedantham, H. K., Koopmans, L. V. E., de Bruyn, A. G., et al. 2014, MNRAS, 437, 1056
Voytek, T. C., Natarajan, A., Jáuregui García, J. M., Peterson, J. B., & López-Cruz, O. 2014, ApJ, 782, L9

Wang, J., Xu, H., An, T., et al. 2013, ApJ, 763, 90
Wang, X., Tegmark, M., Santos, M. G., & Knox, L. 2006, ApJ, 650, 529
Weisz, D. R., & Boylan-Kolchin, M. 2017, MNRAS, 469, L83
Wolz, L., Abdalla, F. B., Blake, C., et al. 2014, MNRAS, 441, 3271
Wu, X. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, 474
Zheng, H., Tegmark, M., Buza, V., et al. 2014, MNRAS, 445, 1084