Draft version March 30, 2018
Preprint typeset using LaTeX style AASTeX6 v. 1.0

CHARACTERIZING SIGNAL LOSS, ERROR, AND BIAS IN THE 21CM REIONIZATION POWER SPECTRUM: A REVISED STUDY OF PAPER-64

Carina Cheng[1,◇], Aaron R. Parsons[1,2], Matthew Kolopanis [3], Daniel C. Jacobs[3], Adrian Liu[1,4,†], Saul A. Kohn[5], Jonathan C. Pober[6], James E. Aguirre[5], Zaki S. Ali[1], Gianni Bernardi[7,8,9], Richard F. Bradley[10,11,12], Chris L. Carilli[13,14], David R. DeBoer[2], Matthew R. Dexter[2], Joshua S. Dillon[1,*], Pat Klima[11], David H. E. MacMahon[2], David F. Moore[5], Chuneeta D. Nunhokee[8], William P. Walbrugh[7], Andre Walker[7]

[1]Astronomy Dept., U. California, Berkeley, CA
[2]Radio Astronomy Lab., U. California, Berkeley CA
[3]School of Earth and Space Exploration, Arizona State U., Tempe AZ
[4]Berkeley Center for Cosmological Physics, Berkeley, CA
[5]Dept. of Physics and Astronomy, U. Penn., Philadelphia PA
[6]Dept. of Physics, Brown University, Providence RI
[7]Square Kilometer Array, S. Africa, Cape Town South Africa
[8]Dept. of Physics and Electronics, Rhodes University, South Africa
[9]INAF-Instituto di Radioastronomia, Bologna Italy
[10]Dept. of Electrical and Computer Engineering, U. Virginia, Charlottesville VA
[11]National Radio Astronomy Obs., Charlottesville VA
[12]Dept. of Astronomy, U. Virginia, Charlottesville VA
[13]National Radio Astronomy Obs., Socorro NM
[14]Cavendish Lab., Cambridge UK

ABSTRACT

The Epoch of Reionization (EoR) is an uncharted era in our Universe's history during which the first stars and galaxies led to the ionization of neutral hydrogen in the intergalactic medium. There are many experiments investigating the EoR by tracing the 21 cm line of neutral hydrogen, a signal which is very faint and difficult to isolate. With a new generation of instruments and a statistical power spectrum detection in our foreseeable future, it has become increasingly important to develop techniques that help maximize sensitivity while validating results. Additionally, it is imperative to understand the trade-offs between different methods and their effects on power spectrum estimates. In this paper, we focus on three major themes — signal loss, power spectrum error estimation, and bias in measurements. We describe techniques that affect these themes using both toy models and data taken by the 64-element configuration of the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER). In particular, we highlight how detailed investigations of these themes have led to a revised, higher 21 cm power spectrum upper limit from PAPER-64. This revised result, presented in a companion paper by Kolopanis et al. (*submitted*), mostly stems from an improved signal loss calculation for loss associated with empirically estimated covariances and supersedes results from previously published PAPER analyses.

1. INTRODUCTION

By about one billion years after the Big Bang ($z \sim 6$), the first stars and galaxies are thought to have ionized all the neutral hydrogen that dominated the baryonic matter content in the Universe. This transition period, during which the first luminous structures formed from gravitational collapse and began to emit intense radiation that ionized the cold neutral gas into a plasma, is known as the Epoch of Reionization (EoR). The EoR is a relatively unexplored era in our *cosmic dawn*, which spans the birth of the first stars to the full reionization of the intergalactic medium (IGM). Its history encodes important information regarding the nature of the first galaxies and the processes of structure formation. Direct measurements of the EoR would unlock powerful characteristics about the IGM, revealing connections between the matter distribution exhibited via cosmic microwave background (CMB) studies and the highly structured web of galaxies we observe today (for a review, see

◇ccheng@berkeley.edu
†Hubble Fellow
*NSF AAPF Fellow

Barkana & Loeb (2001), Furlanetto et al. (2006) and Loeb & Furlanetto (2013)).

One promising technique to probe the EoR is to target the 21 cm wavelength signal that is emitted and absorbed by neutral hydrogen via its spin-flip transition (Furlanetto et al. 2006; Barkana & Loeb 2008; Morales & Wyithe 2010; Pritchard & Loeb 2010; Pritchard & Loeb 2012). This technique is powerful because it can be observed both spatially and as a function of redshift — that is, the wavelength of the signal reaching our telescopes can be directly mapped to a distance from where the emission originated before stretching out as it traveled through expanding space. 21 cm tomography offers a unique window into the evolution of ionization, temperature, and density fluctuations.

In addition to the first tentative detection of the EoR signal made by the Experiment to Detect the Global EoR Signature (EDGES; Bowman et al. 2018; Bowman & Rogers 2010), there are several radio telescope experiments that have succeeded in using the 21 cm signal from hydrogen to place constraints on the brightness of the signal. Examples of experiments investigating the mean brightness temperature of the EoR relative to the CMB are 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 background RAdio Spectrum (SARAS; Patra et al. 2015). Radio interferometers which seek to measure statistical power spectra include the Giant Metre-wave Radio Telescope (GMRT; Paciga et al. 2013a), the LOw Frequency ARray (LOFAR; van Haarlem et al. 2013), the Murchison Widefield Array (MWA; Tingay et al. 2013), the 21 Centimeter Array (21CMA; Peterson 2004; Wu 2009), the Square Kilometre Array (SKA; Koopmans et al. 2015), and PAPER (Parsons et al. 2010). The Hydrogen Epoch of Reionization Array (HERA), which is currently being built, is a next-generation instrument that aims to combine lessons learned from previous experiments and is forecast to be able to make a high-significance power spectrum detection with an eventual 350 elements using current analysis techniques (Pober et al. 2014; Liu & Parsons 2016; Dillon & Parsons 2016; DeBoer et al. 2017).

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 paper, 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, and then perform a detailed case study using data from the 64-element configuration of PAPER, motivating a revised PAPER-64 power spectrum from the lessons learned. A companion paper, Kolopanis et al. (*submitted*), builds off of the methods in this paper to present PAPER-64 results for multiple redshifts and baseline types.

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

This paper is organized into two main halves. In Section 2 we introduce the three themes of our focus, using a toy model to develop intuition for each one. In Section 3 we present a case study into each theme using data from the PAPER-64 array, highlighting key changes from the previously published result in Ali et al. (2015), henceforth known as A15, which have led to a revised PAPER-64 power spectrum result (Kolopanis et al. (*submitted*)). We conclude in Section 4.

2. POWER SPECTRUM THEMES AND TECHNIQUES

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 paper, 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.

Power Spectrum Themes

A deep understanding of the following three themes is essential for the accuracy and interpretation of a 21 cm power spectrum result. Stemming from a re-analysis of PAPER-64 data, we believe these themes serve as an important checklist for a rigorous power spectrum analysis.

- **Signal Loss** (Section 2.1): 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. Computing signal loss correctly has subtle challenges but is necessary to ensure the accuracy of any result.

- **Error Estimation** (Section 2.2): 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** (Section 2.3): 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 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 paper we will highlight some sources of bias, discuss ways to mitigate their effects, and describe tests that a true EoR detection must pass.

Power Spectrum Techniques

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 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 Section 2.1.2.

- **Weighting:** A dataset 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. 2015). 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 has large impacts on signal loss. We investigate the impact of different types of weighting on signal loss in Section 2.1.

- **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 dataset 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 dataset. We explore this potential pitfall of bootstrapping in Section 2.2.

- **Jackknife testing:** A resampling technique useful for estimating bias, jackknives can be taken along different dimensions of a dataset 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 Section 2.3.2.

In the next three subsections, we study each theme in depth, focusing on how power spectrum technique trade-offs affect each. We use a toy data model 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.

2.1. *Signal Loss*

Signal loss can arise in a variety of ways in the 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 applying a weighting matrix to data, a loss that can be significant depending on the choice of weighting and one that was previously underestimated in the A15 analysis.

Driven by the need to mitigate foreground bias, PAPER's previous analyses use a weighting method that aims to down-weight foregrounds. This weighting is applied to data, which is then propagated into a final estimator using the power spectrum estimation technique of quadratic estimators (QE) as done 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. (2015), Switzer et al. (2015), and Trott et al. (2016). Before showing how signal loss can arise when using different weighting matrices, we first summarize QE as performed in the PAPER analysis.

We begin with our data vector, $\mathbf{x}$, which contains our measured visibilities for a single baseline in Jy. It has length $N_t N_f$, but in practice we manipulate it as an array with dimensions (N_f, N_t), where N_t is the number of time integrations and N_f is the number of frequency channels. 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 do this when forming the unnormalized power spectrum estimate $\widehat{q}_\alpha$:

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

Here, $\mathbf{Q}$ is a family of matrices that takes our frequency-domain visibilities and Fourier transforms them, while also converting from Jy to Kelvin and taking into account cosmological scalings. It is formally evaluated as $\mathbf{Q}^\alpha \equiv \frac{\partial \mathbf{C}}{\partial p_\alpha}$ (which has dimensions (N_f,N_f)), or the derivative of the covariance $\mathbf{C} \equiv \langle \mathbf{x}\mathbf{x}^\dagger \rangle$ with respect to the true bandpower p_α, 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. 2012b), which is a good approximation for the short baselines that PAPER analyzes). $\mathbf{R}$ is a weighting matrix with dimensions (N_f,N_f) — as an 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.

We normalize our power spectrum estimates using the matrix $\mathbf{M}$:

$$\widehat{\mathbf{P}} = \mathbf{M}\widehat{\mathbf{q}}, \tag{2}$$

where $\widehat{\mathbf{P}}$ is the estimate of the true power spectrum $\mathbf{P}$. The data analyst has a choice for $\mathbf{M}$. For simplicity in this section we choose $\mathbf{M}$ to be diagonal, although we explore other cases for the analysis of PAPER-64 data as explained in Section 3.3, subject to the constraint that each element of $\widehat{\mathbf{P}}$ is a weighted sum of the elements of $\mathbf{P}$ with weights that sum to unity.

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 Section 3.1.

2.1.1. *Toy Model: Inverse Covariance Weighting*

One choice for the weighting matrix $\mathbf{R}$ used in power spectrum analysis is an inverse covariance matrix. This type of weighting is attractive for power spectrum analyses because it yields the smallest possible error bars on a measurement. Said differently, it gives the minimum variance estimate of the power spectrum (Tegmark 1997; Bond et al. 1998). In Liu & Tegmark (2011), it was shown that inverse covariance weighting also serves as a way to down-weight some portion of foregrounds (namely, those which do not share the same covariance structure as the cosmological signal), motivating our use of the weighting for previous PAPER analyses.

One important feature of weighting data by its true inverse covariance is that, while it can suppress some foregrounds, it cannot suppress the EoR signal (nor foreground modes that masquerade as signal-like modes). By construction, inverse covariance weighting in the quadratic estimator does not lead to signal loss. Therefore, in an ideal world with perfect foreground, instrumental, and EoR models, we could form $\mathbf{C}$ in a way that accurately describes our measured data and use inverse covariance weighting to produce minimum variance power spectrum estimates without destroying the signal. In other words, the optimal quadratic estimator is by construction an unbiased estimator of the power spectrum.

In practice, we do not have perfect models for $\mathbf{C}$, and it is the use of an estimated covariance $\widehat{\mathbf{C}}$ instead of the true covariance $\mathbf{C}$ that can lead to loss. We will now build intuition into how empirically estimated inverse covariance weighting can give rise to signal loss through the use of a toy model.

We construct a simple dataset that contains visibility data with 100 time integrations and 20 frequency channels. This model represents realistic dimensions of

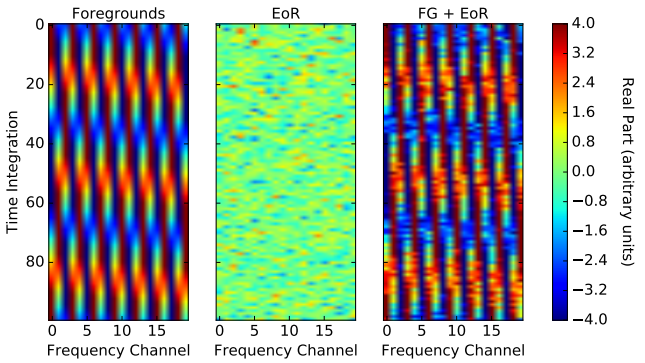

Figure 1. Our toy model dataset to which we apply different weighting schemes to in order to investigate signal loss. We model a mock foreground-only visibility with a sinusoid signal that varies smoothly in time and frequency. We model a mock visibility of an EoR signal as a random Gaussian signal. We add the two together to form $\mathbf{x} = \mathbf{x}_{\rm FG} + \mathbf{x}_{\rm EoR}$. Real parts are shown here.

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 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 matrix is then $\mathbf{x} = \mathbf{x}_{\rm FG} + \mathbf{x}_{\rm EoR}$, to which we apply different weighting schemes throughout Section 2.1. The three data components are shown in Figure 1.

Without a perfect model for the covariance matrix $\mathbf{C}$ of our data, one attractive way to estimate it 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 such that:

$$\widehat{\mathbf{C}} \equiv \langle \mathbf{x}\mathbf{x}^\dagger \rangle_t, \tag{3}$$

assuming $\langle \mathbf{x} \rangle_t = 0$ (a reasonable assumption since fringes average to 0 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 $\mathbf{R} \equiv \widehat{\mathbf{C}}^{-1}$.

First, we compute the power spectrum of our toy model dataset $\mathbf{x}$ using QE formalism and $\mathbf{R} \equiv \widehat{\mathbf{C}}^{-1}$. The result is shown in green in the left plot of Figure 2. 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. 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. This is evidence of signal loss. 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. For a toy model mathematical derivation of signal loss arising from a data-estimated covariance matrix, see Appendix A. Here we will describe the origin of this signal loss intuitively.

To begin to understand the lossy behavior of this result, we can closely study our estimated covariance eigenspectrum, shown in Figure 3. Since it is estimated from our data, its eigenspectrum differs from the eigenspectrum of the true covariance $\mathbf{C}$, and this difference has important consequences on our result.

An eigenspectrum ranks the eigenvalues of a matrix from highest to lowest and can be thought of as a spectrum of weights that are given to each spectral mode in the data. In other words, the eigenvalues encode the strength of different shapes in the dataset. The eigenspectrum of the identity matrix $\mathbf{I}$ is flat (all 1's) because it gives equal weighting to all modes. In our application, covariance matrices tend to have sloped eigenspectra, meaning that modes are given different weights in QE power spectrum estimation. The modes with the highest eigenvalues are down-weighted the most.

Because $\widehat{\mathbf{C}}$ is estimated from the data, its eigenvectors and eigenvalues are strongly coupled to the particular data realization that was used to compute it. For example, the strongest mode of $\widehat{\mathbf{C}}$ (highest eigenvalue) describes the sinusoid foreground mode in the toy model (the peak in Figure 2). In Figure 4 we show the estimated covariances of our toy model datasets along with $\widehat{\mathbf{C}}^{-1}$ weighted data. The foreground sinusoid is clearly visible in $\widehat{\mathbf{C}}_{\rm FG}$.

In general, the strongest eigenmodes of $\widehat{\mathbf{C}}$ typically describe bright foregrounds — the most prominent shapes in a dataset. For these "strong" modes, where foregrounds outshine the EoR signal, down-weighting is beneficial. In some sense, we desire signal loss in this regime, if by 'signal' we mean 'foregrounds.' In this case it is beneficial for the "strong" eigenmodes to be coupled to the data in a way that produces loss. For our toy model, the successful suppression of the foreground mode is demonstrated in Figure 2 by the missing foreground peak in the weighted power spectrum estimate (green).

The danger of an empirically estimated covariance matrix comes mostly from not being able to describe

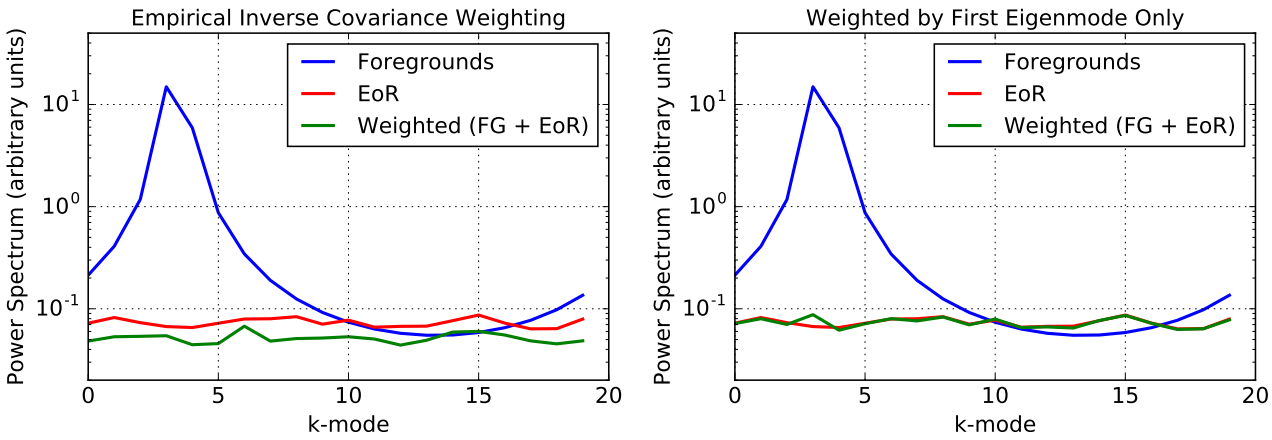

Figure 2. Resulting power spectrum estimates for the toy model simulation described in Section 2.1.1 — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). The power spectrum of foregrounds peaks at a k-mode based on the frequency of the sinusoid used to create the mock FG signal. In the two panels, we compare empirically estimated inverse covariance weighting where $\mathbf{C}$ is derived from the data (left), and projecting out the first eigenmode only (right). In the former case, signal loss arises from the coupling of the eigenmodes of $\widehat{\mathbf{C}}$ to the data. For an empirically estimated $\widehat{\mathbf{C}}$, its eigenvalues differ from those of the true covariance, where the weakest (EoR-dominated) eigenmodes are the most strongly coupled to the data and can lead to the most loss. Hence, there is negligible signal loss when assigning identical weights to all eigenmodes except the first, since we are not using the relative weights of the weaker eigenmodes.

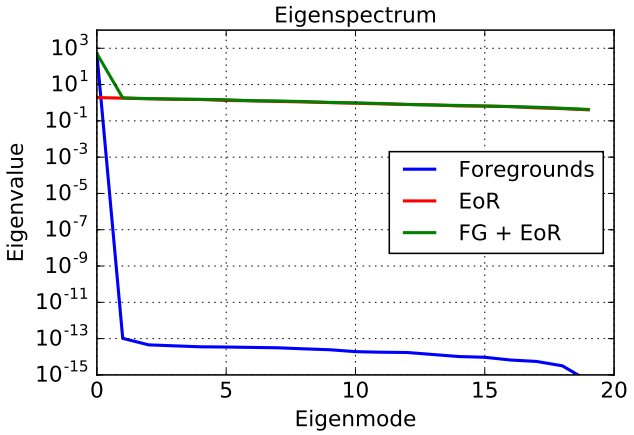

Figure 3. Eigenspectrum of $\widehat{\mathbf{C}}_{\mathrm{FG}}$ (blue), $\widehat{\mathbf{C}}_{\mathrm{EoR}}$ (red), and $\widehat{\mathbf{C}}_{\mathrm{FG+EoR}}$ (green). The eigenspectrum of $\widehat{\mathbf{C}}_{\mathrm{FG}}$ peaks at the zeroth eigenmode, due to the presence of only one sinusoid. These empirically estimated covariance matrices have eigenspectra that are different from that of the true $\mathbf{C}$. Specifically, these eigenmodes have the risk of being down-weighted more significantly than they should be because they are coupled to the data in a way that produces loss.

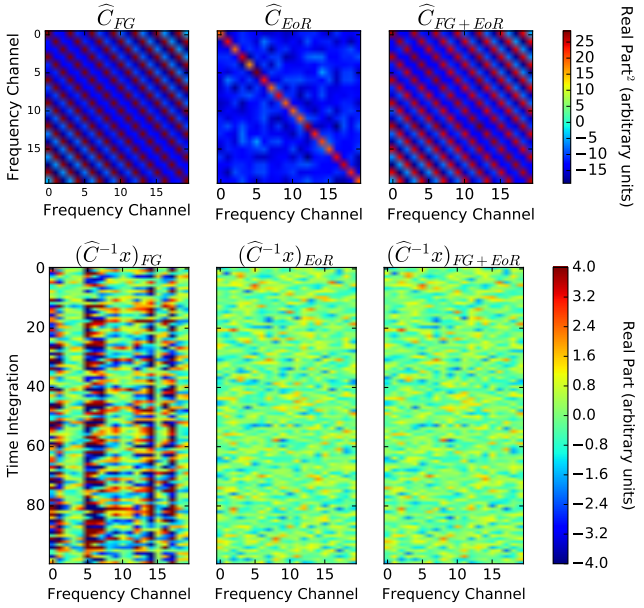

Figure 4. The estimated covariance matrices (top row) and inverse covariance-weighted data (bottom row) for FG only (left), EoR only (middle), and FG + EoR (right). Real parts are shown here.

the "weak" 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 Section 3.1.1), which leads to loss. Mis-estimating $\mathbf{C}$ for EoR-dominated eigenmodes is therefore more harmful than for FG-dominated modes, and since the "weak" modes of an eigenspectrum are typically EoR-dominated, using this part of the spectrum for weighting is most dangerous.

Using what we've learned about the eigenspectrum, we can tweak it in a simple way to suppress foregrounds and yield minimal signal loss. Recall that our toy model foreground is a sinusoid, so it can be perfectly described by a single eigenmode. Using the full dataset's (foreground plus EoR signal) empirical covariance, we can project out the first eigenmode and then flatten the rest of the spectrum to have eigenvalues of 1, thereby down-weighting the foreground-dominated mode more than the rest of the modes. Hence, we are changing the weaker part of the spectrum to be less coupled to the

data, limiting the amount of over-fitting that can happen for those modes (i.e. only allowing over-fitting to occur for the first mode).

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 changes its coupling to the data realization. The resulting power spectrum estimate for this case is shown in the right plot of Figure 2. In this case we recover the EoR signal, demonstrating that if we can disentangle the foreground-dominating modes and EoR-dominating modes, we can down-weight them with negligible signal loss. There are several other ways to regularize $\widehat{\mathbf{C}}$, and we will discuss some in Section 2.1.3.

2.1.2. *Toy Model: Fringe-Rate Filtering*

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

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 which 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. To mimic this filter, we average every four time integrations of our toy model dataset together, yielding 25 independent samples in time (Figure 5). We choose these numbers so that the total number of independent samples is similar to the number of frequency channels — hence our matrices will still be full rank.

The resulting eigenspectrum as compared to the green curve (FG + EoR) in Figure 3 is shown in Figure 6. With fringe-rate filtering resulting in fewer independent modes, it becomes more difficult for the empirical covariance to estimate the true covariance matrix of the fringe-rate filtered data. This can be quantified by evaluating a convergence metric $\varepsilon(\widehat{\mathbf{C}})$ for the empirical covariance, which we define as

$$\varepsilon(\widehat{\mathbf{C}}) \equiv \sqrt{\frac{\sum_{ij}(\widehat{C}_{ij} - C_{ij})^2}{\sum_{ij} C_{ij}^2}}, \tag{4}$$

where $\mathbf{C}$ is the true covariance matrix. In Figure 7 we show this convergence statistic as a function of the number of independent ensemble realizations in one's simulations (horizontal axis) and the number of independent samples in the data following 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 Section 3.1.1 and Appendix A), 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).

Just as important as the eigenvalues are the eigenvectors of our empirical covariances. In general, different eigenvectors converge to their true forms at different rates. This is illustrated by Figure 8, which shows the convergence of eigenvectors in an empirical estimate of a covariance matrix whose true form is a diagonal matrix with eigenvalues spanning four orders of magnitude. We use as a convergence metric $\varepsilon(\widehat{\mathbf{v}})$ for the empirical eigenvectors $\widehat{\mathbf{v}}$, defined as:

$$\varepsilon(\widehat{\mathbf{v}}) \equiv \sqrt{\sum_{i}^{N_f} |\mathbf{v} - \widehat{\mathbf{v}}|_i^2}, \tag{5}$$

where N_f is the number of frequencies (20) in the mock data. The eigenmode convergence curves in Figure 8 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. One sees that the stronger eigenmodes converge to their true eigenvectors more quickly. With only a small number of realizations, these empirically estimated modes already retain little correlation with the specific realizations of data that were used to form the empirical covariance. As we will see in the next section, using only the strongest eigenmodes, which are less coupled to data realizations, minimizes signal loss. In contrast, the weaker eigenmodes retain more memory of the data realizations, which leads to correlations that induce signal loss. Said differently, a steep covariance eigenspectrum can be especially dangerous because it is the "weak" modes that are both EoR-dominated and converge the slowest and are therefore susceptible to the most loss.

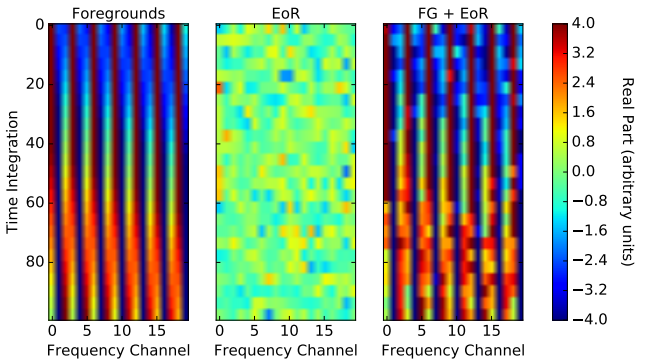

Figure 5. Our 'fringe-rate filtered' (time-averaged) toy model dataset. We average every four samples together, yielding 25 independent samples in time. Real parts are shown here.

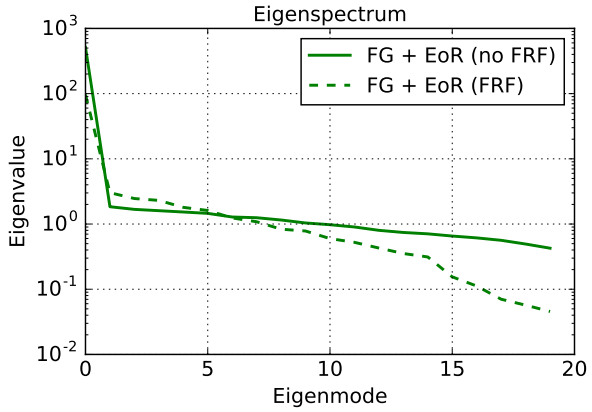

Figure 6. Eigenspectrum of $\widehat{\mathbf{C}}_{\mathrm{FG+EoR}}$, in the case of no fringe-rate filtering (solid green) and with fringe-rate filtering (dashed green). The dashed, steep eigenspectrum has a greater risk of signal loss because its weak eigenmodes are more strongly coupled to the data than those of the solid eigenspectrum.

The power spectrum results for the fringe-rate filtered toy model data are shown in Figure 9. As expected, there is a much larger amount of signal loss for this time-averaged dataset 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 9 fails to trace the shape of the uniform-weighted EoR power spectrum.

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 poorly characterized, steep eigenspectra 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.

2.1.3. *Toy Model: Other Weighting Options*

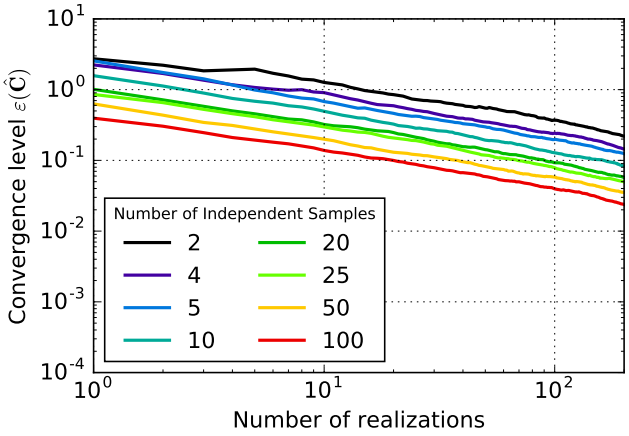

Figure 7. The convergence level, as defined in Equation (4), of empirically estimated covariances of mock EoR signals with different numbers of independent samples. In red, the mock EoR signal is comprised entirely of independent samples. Subsequent colors show time-averaged signals. As the number of realizations increases, we see that the empirical covariances approach the true covariances. With more independent samples, the quicker an empirical covariance converges, and the less signal loss we would expect to result.

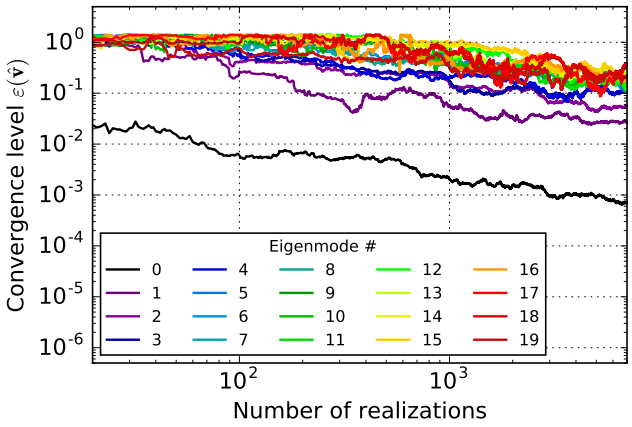

Figure 8. The convergence level, as defined in Equation (5), of empirically estimated eigenvectors for different numbers of mock data realizations. The colors span from the 0th eigenmode (has the highest eigenvalue) to the 19th eigenmode (has the lowest eigenvalue), where they are ordered by eigenvalue in descending order. This figure shows that "strong" eigenmodes (those with the highest eigenvalues) converge more quickly than "weak" eigenmodes, implying that weighting by empirically estimated "weak" modes poses the most risk for signal loss.

In Section 2.1.1 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 and eigenspectra for the different cases in Figures 10 and 11. These examples are not meant to be taken as suggested analysis methods

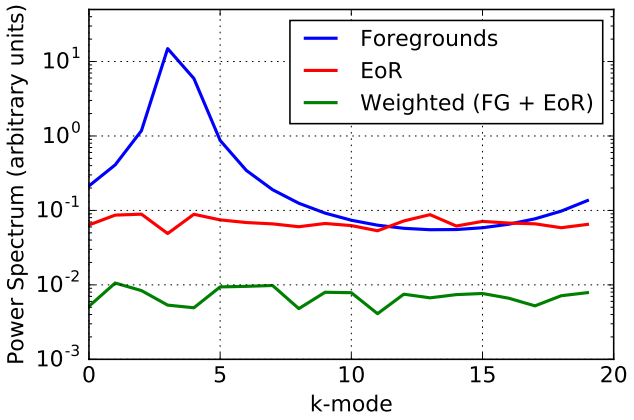

Figure 9. Resulting power spectrum estimate for the 'fringe-rate filtered' (time-averaged) toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). We use empirically estimated inverse covariance weighting where $\mathbf{C}$ is computed from the data. 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.

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 $\mathbf{C}_{\rm EoR}$ as such (i.e. the identity), instead of computing it based on $\mathbf{x}_{\rm EoR}$ itself. Next, we add $\mathbf{C}_{\rm EoR}+\widehat{\mathbf{C}}_{\rm FG}$ (where $\widehat{\mathbf{C}}_{\rm FG}=\langle\mathbf{x}_{\rm FG}\mathbf{x}_{\rm FG}^{\dagger}\rangle_t$) to obtain a final $\widehat{\mathbf{C}}$ to use in weighting. In Figure 10 (upper left), we see that there is negligible signal loss. This is because by modeling $\mathbf{C}_{\rm EoR}$, we avoid over-fitting EoR fluctuations in the data that our model doesn't know about (but an empirically derived $\widehat{\mathbf{C}}_{\rm EoR}$ would). This is also shown by comparing the (steeper) green and (flatter) red curves in Figure 11. In practice such a weighting option is not feasible, as it is difficult to model $\mathbf{C}_{\rm EoR}$, and $\widehat{\mathbf{C}}_{\rm FG}$ is unknown because we don't know how to separate out the foregrounds from the EoR in our data.

The second panel (top right) in Figure 10 uses a regularization method of setting $\widehat{\mathbf{C}}\equiv\widehat{\mathbf{C}}+\gamma\mathbf{I}$, 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. Although there is negligible signal loss using this regularization, the small green peak at the third k-mode represents residual foregrounds that still exist since the shapes encoded in the off-diagonal frequency correlations of the covariance matrix were deemed not as prominent as the diagonal elements using this weighting scheme.

The third panel (bottom left) in Figure 10 minimizes signal loss a different way — by only using the first three eigenmodes of the estimated covariance. Recalling that our toy model foregrounds can be described entirely by the first eigenmode, this method intentionally projects out the strongest modes only by replacing all but the three highest weights in the eigenspectrum with 1's (equal weights). Again, avoiding the over-fitting of "weak" modes which are coupled to the data results in negligible signal loss. However, we do not perfectly recover the shape of the EoR power spectrum because we lost information when ignoring the relative weights of most modes. While this case is illuminating for the toy model, in practice it is not obvious which eigenmodes are FG or EoR dominated (and they could be mixed as well), so determining which subset of modes to down-weight is not trivial.

The last regularization scheme we are highlighting here is setting $\widehat{\mathbf{C}}\equiv\widehat{\mathbf{C}}\circ\mathbf{I}$ (element-wise multiplication), or inverse variance weighting (keeping only the diagonal elements of $\widehat{\mathbf{C}}$). In the bottom right panel of Figure 10, we see that this method does not down-weight the foregrounds — this regularization altered $\widehat{\mathbf{C}}$ in a way where it is no longer coupled to either the "strong" or "weak" eigenmodes. For this toy model, our foregrounds are spread out in frequency and therefore have non-negligible frequency-frequency correlations. Multiplying by the identity, 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 de-coupled 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 than others based on the spectral nature of the components in a dataset.

In summary, we have a choice of how to weight 21 cm data. Ideally, we want to down-weight bright foregrounds without removing the underlying cosmological signal. However, there are trade-offs between the weighting method used, its foreground-removal effectiveness, the number of independent samples in a dataset, and the amount of resulting signal loss.

2.2. *Error Estimation*

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

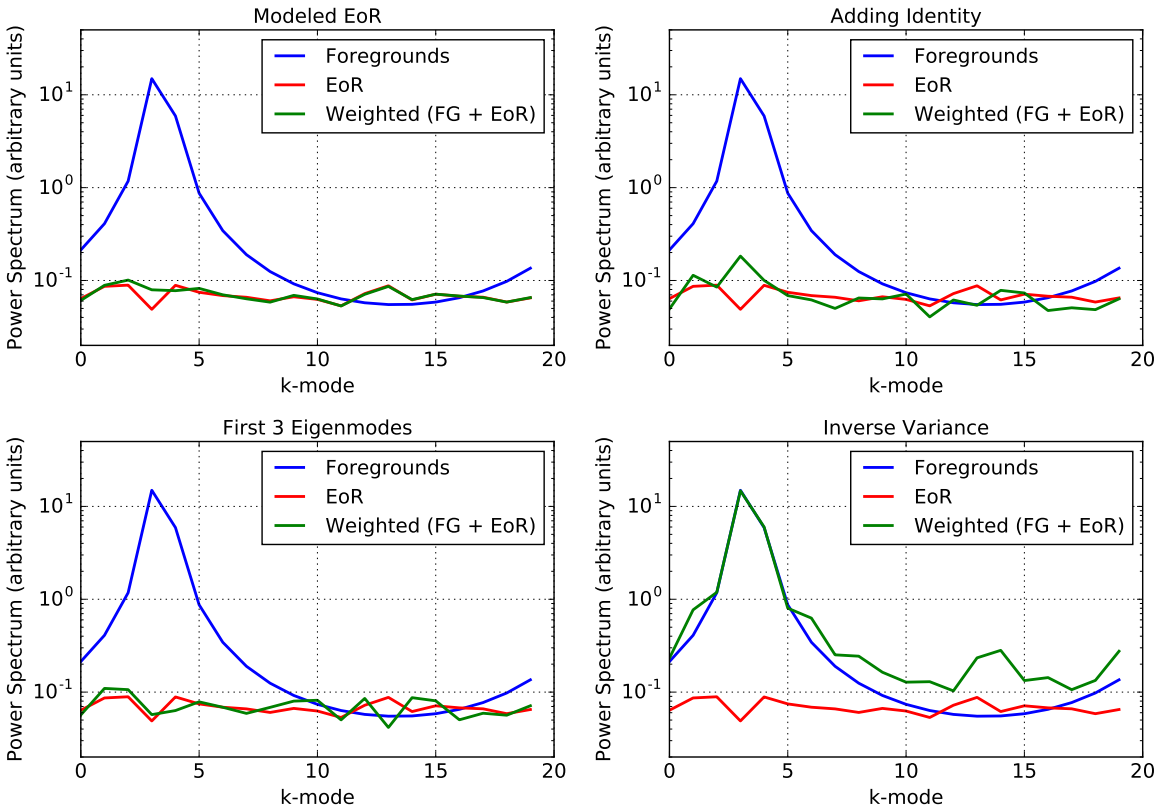

Figure 10. Resulting power spectra estimates for our 'fringe-rate filtered' (time-averaged) toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). We show four alternate weighting options that each minimize signal loss, including modeling the covariance 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 $\mathbf{C}_{\rm FG}$ and $\mathbf{C}_{\rm EoR}$ like we do in the toy model.

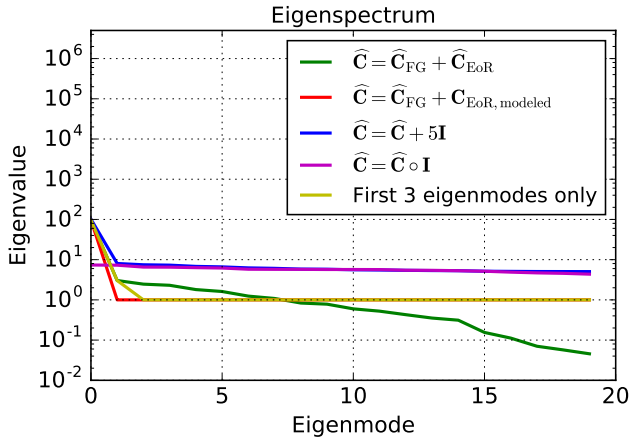

Figure 11. We compare the eigenspectrum of an empirically calculated $\widehat{\mathbf{C}}$ (green) to that of four alternate weighting options, including modeling the covariance matrix of EoR (red), regularizing $\widehat{\mathbf{C}}$ by adding an identity matrix to it (blue), using only the first three eigenmodes of $\widehat{\mathbf{C}}$ (yellow), and multiplying an identity matrix with $\widehat{\mathbf{C}}$ (magenta). All eigenspectra (except the green) are relatively flat and don't exhibit signal loss. All were computed for the 'fringe-rate filtered' (time-averaged) toy model case presented in Section 2.1.2.

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 resampling and averaging along different axes of a dataset, 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 1994; 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 dataset of length $N = 1000$ and unity variance (zero mean). This could represent 1000 power spectrum measurements, 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 ($N_{\rm boot} = 500$). 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 dataset. This error is indicated by the gray star in Figure 12 and matches our theoretical prediction (green).

One major caveat of bootstrapping arises when working with correlated data. If, for example, a dataset 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, under-estimates 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), under-estimates the error by a factor of ~ 3 (black points in Figure 12, 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 12, black points), and we can both under-estimate the error by drawing too many or over-estimate 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 12).

This example highlights the importance of understanding how analysis techniques (e.g. 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 Section 3.2.2, we explain how we under-estimated 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 dataset. However, we have illustrated a situation in which bootstrapping can lead to under-estimated errors and therefore under-estimated 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 cer-

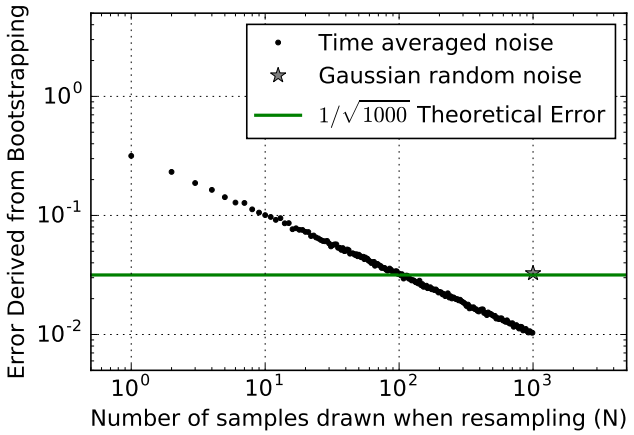

Figure 12. Error estimation from bootstrapping as a function of the number of elements drawn per bootstrap when sampling with replacement. The star represents the standard deviation of $N_{\rm boot} = 500$ bootstraps, each created by drawing 1000 elements (with replacement) from a length 1000 array of 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 under-estimated if drawing more elements than there are independent samples in the data. The estimated errors match up with the theoretical prediction only at $N = 100$.

tain analysis choices have compromised the method and whether a variation or an avoidance of traditional resampling could be preferred instead.

2.3. *Bias*

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.

2.3.1. *Foreground and Noise Bias*

In Section 2.1, 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. 2012a; Parsons et al. 2012b; 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. 2015; Pober et al. 2016; 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. 2012b; 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. (2015), 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 Section 3.3).

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 (1). 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 appears as a constant phase bias in time in visibilities, and it varies slowly compared to 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.

2.3.2. *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 Section 3.3.

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 (e.g., 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 dataset, 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 dataset.

Suppose we have two measurements (for example, from two baselines), $\mathbf{x}_a$ and $\mathbf{x}_b$. The measurements have dimensions of 200 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 in part of the measurement, we add a toy sinusoid foreground, similar to the one used in Section 2.1.1, to the first 100 time integrations of both measurements. This represents a foreground signal present in, for example, the first half of the LST range used for analysis, but not the second half. Mathematically, if $\mathbf{n}$ is noise, $\mathbf{e}$ is the EoR signal, and $\mathbf{f}$ is the foreground signal, the two measurements (which are cross-multiplied to form power spectra) can be written as:

$$\mathbf{x}_a = \mathbf{n}_a + \mathbf{e} + \mathbf{f} \tag{6}$$

and

$$\mathbf{x}_b = \mathbf{n}_b + \mathbf{e} + \mathbf{f}. \tag{7}$$

The two jackknife samples are $\mathbf{x}_1$ and $\mathbf{x}_2$, representing jackknives in time. These can be written (for both measurements a and b) as:

$$\mathbf{x}_1 = \mathbf{n}_1 + \mathbf{e}_1 + \mathbf{f} \tag{8}$$

$$\mathbf{x}_2 = \mathbf{n}_2 + \mathbf{e}_2 \tag{9}$$

For example, $\mathbf{x}_{1a}$ represents a jackknife sample (first half of the data) for the first measurement. Similarly, $\mathbf{x}_{2b}$ represents a jackknife sample (second half of the data) for the second measurement. We note that $\mathbf{e}$ itself will be different in two jackknives taken along LST, but could be identical in other cases (such as a jackknife along even/odd Julian days).

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.

We form three different power spectrum estimates shown in Figure 13. The first is a null test where we

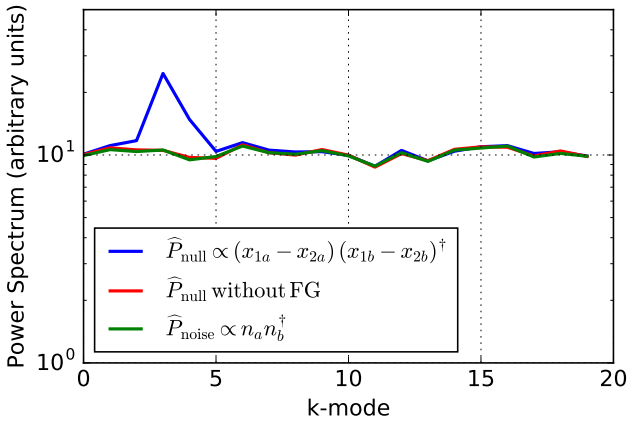

Figure 13. Power spectrum estimates for a null jackknife test with the presence of a foreground systematic (blue), without the foreground systematic (red), and noise alone (green). Because the first null test is not consistent with noise, it suggests the presence of a systematic in either $\mathbf{x}_1$ or $\mathbf{x}_2$. Null tests of clean measurements should be consistent with thermal noise.

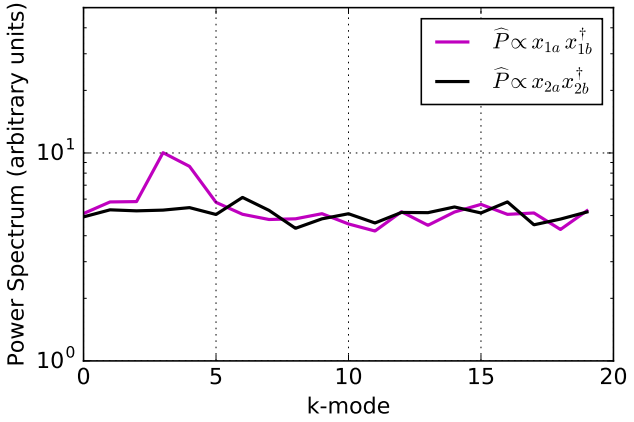

Figure 14. Power spectrum estimates for $\mathbf{x}_1$ and $\mathbf{x}_2$, two jackknives of the toy model. They suggest the presence of a systematic in $\mathbf{x}_1$ only (which is exactly what was put in), illustrating how jackknives can be used to tease out excesses. Clean measurements should remain consistent despite the jackknife taken.

subtract $\mathbf{x}_2$ from $\mathbf{x}_1$ for both measurements a and b. This is equivalent to splitting up a full dataset along an axis (in this case, time) and subtracting the two to remove sky signal that should ideally be present in both. We cross-multiply measurements a and b to form an unbiased (thermal noise-wise) estimate (blue curve). The second estimate, shown in red, is the same null test with the foreground systematic removed (eliminate $\mathbf{f}$ in Equations 6 and 7). Finally, we also show the noise power spectrum (green).

From this test we see a clear difference between the null test with the presence of the foregrounds, 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 dataset, 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 $\mathbf{x}_1$ (Figure 14). If foreground or noise biases exist in a dataset, 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 Section 3.3 we perform some examples of these tests on PAPER-64 data in order to prove that our excesses are not EoR and to identify their likely cause.

3. DEMONSTRATION IN PAPER-64 DATA

In the previous sections 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 data from the PAPER experiment to make a new analysis of the PAPER-64 dataset originally presented in A15 and obtain a revised power spectrum estimate.

As a brief review, 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. For our case study, we focus solely on Stokes I estimated data (Moore et al. 2013) from PAPER's 30 m East/West baselines (Figure 15). All data is compressed, calibrated (using self-calibration and redundant calibration), delay-filtered (to remove foregrounds inside the wedge), LST-binned, and fringe-rate filtered. For detailed information about the backend system of PAPER-64, its observations, and data reduction pipeline, we refer the reader to Parsons et al. (2010) and A15. We note that all data processing steps are identical to those in A15 until after the LST-binning step in Figure 3 of A15.

The previously best published 21 cm upper limit result from A15 uses 124 nights of data to place a 2σ upper limit on $\Delta^2(k)$, defined as

$$\Delta^2(k) = \frac{k^3}{2\pi^2}\,\hat{P}(k), \tag{10}$$

of $(22.4\ \mathrm{mK})^2$ in the range $0.15 < k < 0.5\,h\ \mathrm{Mpc}^{-1}$

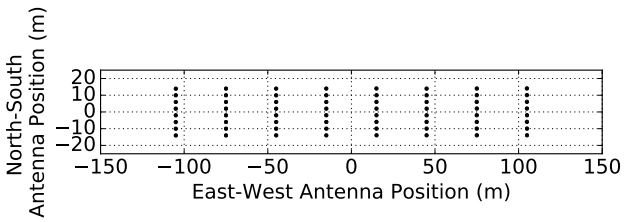

Figure 15. The PAPER-64 antenna layout. We use only the 30 m East/West baselines for the revised analysis in this paper (i.e. the shortest horizontal spacings).

at $z = 8.4$. The revision of this limit (Kolopanis et al. (*submitted*)) stems from previously under-estimated signal loss and under-estimated error bars, both of which we address in the following sections.

For our revised analysis, we use 8.1 hours of LST (RA 0.5-8.6 hours) and 51 total baselines (A15 uses a slightly longer RA range of 0-8.6 hours; we found that some early LSTs were more severely foreground contaminated). 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 paper, we only use one baseline type instead of the three as in A15, but Kolopanis et al. (*submitted*) uses all baselines presented in A15.

The most significant changes from A15 occur in our revised power spectrum analysis, which is explained in the rest of this paper, 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 by widening it in fringe-rate space. This was chosen in order to increase the number of independent modes and reduce signal loss, though as we will explain in the next section, this signal loss was still under-estimated. 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 16.

3.1. *PAPER-64: Signal Loss*

In Section 2.1, we showed how signal loss arises when weighting data using information from the data itself. Here we describe a methodology for estimating the amount of signal loss caused by a particular power spectrum estimator when applied to a particular dataset. The exact amount of signal loss will depend on the specific realizations of the signals present in the data and is not something we can directly compute. In this work, 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

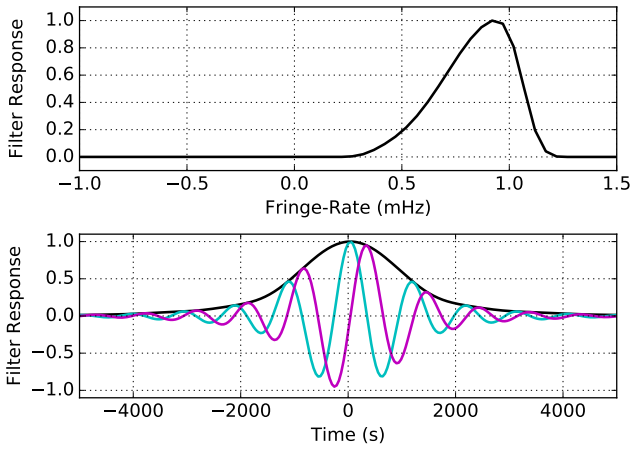

Figure 16. Top: the normalized optimal power-spectrum sensitivity weighting in fringe-rate space for our fiducial baseline and Stokes I polarization beam. Bottom: the time-domain convolution kernel corresponding to the top panel. Real and imaginary components are illustrated in cyan and magenta, respectively, 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 dataset.

significant complicating factors which were previously not taken into account.

We present our PAPER-64 signal loss investigation in three parts: we first give an overview of our injection framework and illustrate how different power spectrum components affect loss. We next describe our methodology in practice and detail how we map our simulations into a posterior for the EoR signal. Finally, we build off of Section 2.1 by experimenting with different regularization schemes on PAPER data in order to minimize loss. Throughout each section, we also highlight major differences from the signal loss computation used in A15, which previously under-estimated losses.

3.1.1. *Signal Loss Methodology*

In short, our method consists of adding an EoR-like signal into the 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 data.[1]

Suppose that $\mathbf{e}$ is the mock injected EoR signal (at

[1] One specific change from A15 is that we add this simulated signal into the analysis pipeline before the final fringe-rate filter is applied to the data. Previously, the addition was done after that final fringe-rate filter step. This change results in an increased estimate of signal loss, likely due to the use of the fringe-rate filter

some amplitude level), and $\mathbf{x}$ is our data. We define $\mathbf{r}$ to be the data plus the EoR signal:

$$\mathbf{r} = \mathbf{x} + \mathbf{e}. \tag{11}$$

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: $P_{\rm in}$ and $P_{\rm out}$, defined as

$$P_{{\rm in},\alpha} \equiv \mathrm{M}^{\alpha}_{\rm in}\mathbf{e}^{\dagger}\mathbf{I}\mathbf{Q}^{\alpha}\mathbf{I}\mathbf{e} \tag{12}$$

and

$$\begin{aligned} P_{{\rm out},\alpha} &\equiv \widehat{\mathbf{P}}_{r,\alpha} \\ &= \mathrm{M}^{\alpha}_{r}\mathbf{r}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{r}, \end{aligned} \tag{13}$$

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 non-diagonal matrix normalization using $\mathbf{M}$ as in Equation (2).

The quantity $P_{\rm in}$, defined by Equation (12), is a uniformly weighted estimator of the power spectrum of $\mathbf{e}$. Since there is no noise contribution to $\mathbf{e}$, 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 $P_{\rm in}$ 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. This is then to be compared to $P_{\rm out}$, which approximates the (lossy) power spectrum estimate that is output by our analysis pipeline prior to any signal loss adjustments. In principle, one could compute $P_{\rm out}$ using end-to-end simulations of the instrument and data analysis pipeline. In practice, however, such simulations may not accurately reflect real-life systematics and foregrounds. To overcome this obstacle, one can make the assumption that since the EoR signal is expected to be small, the data vector $\mathbf{x}$ itself is our best model of these contaminants. Making this assumption, the injected EoR signal $\mathbf{e}$ takes on the role of the true EoR signal, and the sum of $\mathbf{x}$ and $\mathbf{e}$ (i.e., Equation (11)) replaces $\mathbf{x}$ as our model of the measured data. Therefore, $P_{\rm out}$ can be directly compared to the measurement that we make.

Under this injection framework, we can begin to see explicitly why there can be large signal loss. Expanding out Equation (13), $P_{\rm out}$ becomes:

$$\begin{aligned} P_{{\rm out},\alpha} &= \mathrm{M}^{\alpha}_{r}(\mathbf{x}+\mathbf{e})^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}(\mathbf{x}+\mathbf{e}) \\ &= \mathrm{M}^{\alpha}_{a}\mathbf{x}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{x} + \mathrm{M}^{\alpha}_{b}\mathbf{e}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{e} \\ &\quad + \mathrm{M}^{\alpha}_{c}\mathbf{x}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{e} + \mathrm{M}^{\alpha}_{d}\mathbf{e}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{x} \end{aligned} \tag{14}$$

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:

$$\begin{aligned} P_{{\rm out},\alpha} &= \mathrm{M}^{\alpha}_{a}\mathbf{x}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{x} + \mathrm{M}^{\alpha}_{b}\mathbf{e}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{e} \\ &\quad + 2\mathrm{M}^{\alpha}_{c}\mathbf{x}^{\dagger}\mathbf{R}_{r}\mathbf{Q}^{\alpha}\mathbf{R}_{r}\mathbf{e} \end{aligned} \tag{15}$$

In order to investigate the effect of each of these terms on signal loss, all three components are plotted in Figure 17 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 examine the behavior of this equation in three different regimes of the injected signal - very weak (left ends of the $P_{\rm in}$ axes in Figure 17), 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 $\sqrt{P_e}\sqrt{P_x}$. We also note that in this regime, $\widehat{\mathbf{P}}_r$ (black) approaches the data level (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 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 17, 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 estimated 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 $\sqrt{P_e}\sqrt{P_x}$.

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

as a simulator. However, this pipeline difference, while significant, is not the dominant reason why signal loss was under-estimated in A15 (the dominant reason is explained in the main text in Section 3.1.1).

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 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 $P_{\rm out}$ estimator (black). We note that in A15, signal loss was computed by only looking at the second term in Equation (15) (green), which incorrectly implies no loss at the data level. Ignoring the effect of the negative power from the cross-terms is the main reason for under-estimating 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 (15) 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 $\mathbf{r}=\mathbf{x}+\mathbf{e}$:

$$\begin{aligned}\widehat{\mathbf{C}}_r &= \langle \mathbf{r}\mathbf{r}^\dagger\rangle_t \\ &= \langle \mathbf{x}\mathbf{x}^\dagger\rangle_t + \langle \mathbf{x}\mathbf{e}^\dagger\rangle_t + \langle \mathbf{e}\mathbf{x}^\dagger\rangle_t + \langle \mathbf{e}\mathbf{e}^\dagger\rangle_t, \end{aligned} \tag{16}$$

where we can neglect the first term because $\mathbf{x}$ is small. 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 (which is also discussed in Appendix A) 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 (1) 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.

3.1.2. *Signal Loss in Practice*

We now shift our attention towards computing upper limits on the EoR signal for the fringe-rate filtered PAPER-64 dataset 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. With this weighting, our expressions for $P_{\rm in}$ and $P_{\rm out}$ become:

$$P_{{\rm in},\alpha} = {\rm M}_{\rm in}^\alpha \mathbf{e}^\dagger \mathbf{I}\mathbf{Q}^\alpha \mathbf{I}\mathbf{e} \tag{17}$$

$$P_{{\rm out},\alpha} = {\rm M}_r^\alpha \mathbf{r}^\dagger \widehat{\mathbf{C}}_r^{-1}\mathbf{Q}^\alpha \widehat{\mathbf{C}}_r^{-1}\mathbf{r}. \tag{18}$$

One issue to address is how one incorporates the randomness of $P_{\rm out}$ 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 two ways — there is a variation driven by the random seed (i.e. cosmic variance) and 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).

Phrased in the context of Bayes' rule, we wish to find the posterior probability distribution $p(P_{\rm in}|P_{\rm out})$, which is the probability of $P_{\rm in}$ given the uncorrected/measured power spectrum estimate $P_{\rm out}$. Bayes' rule relates the posterior, which we don't know, to the likelihood, which we can forward model. In other words,

$$p(P_{\rm in}|P_{\rm out}) \propto \mathcal{L}(P_{\rm out}|P_{\rm in})\, p(P_{\rm in}), \tag{19}$$

where $\mathcal{L}$ is the likelihood function defined as the distribution of data plus injection ($P_{\rm out}$) 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 $P_{\rm out}$ 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 with the goal of simultaneously capturing variance due to noise 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. $P_{\rm out}$ 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 $P_{\rm in}$

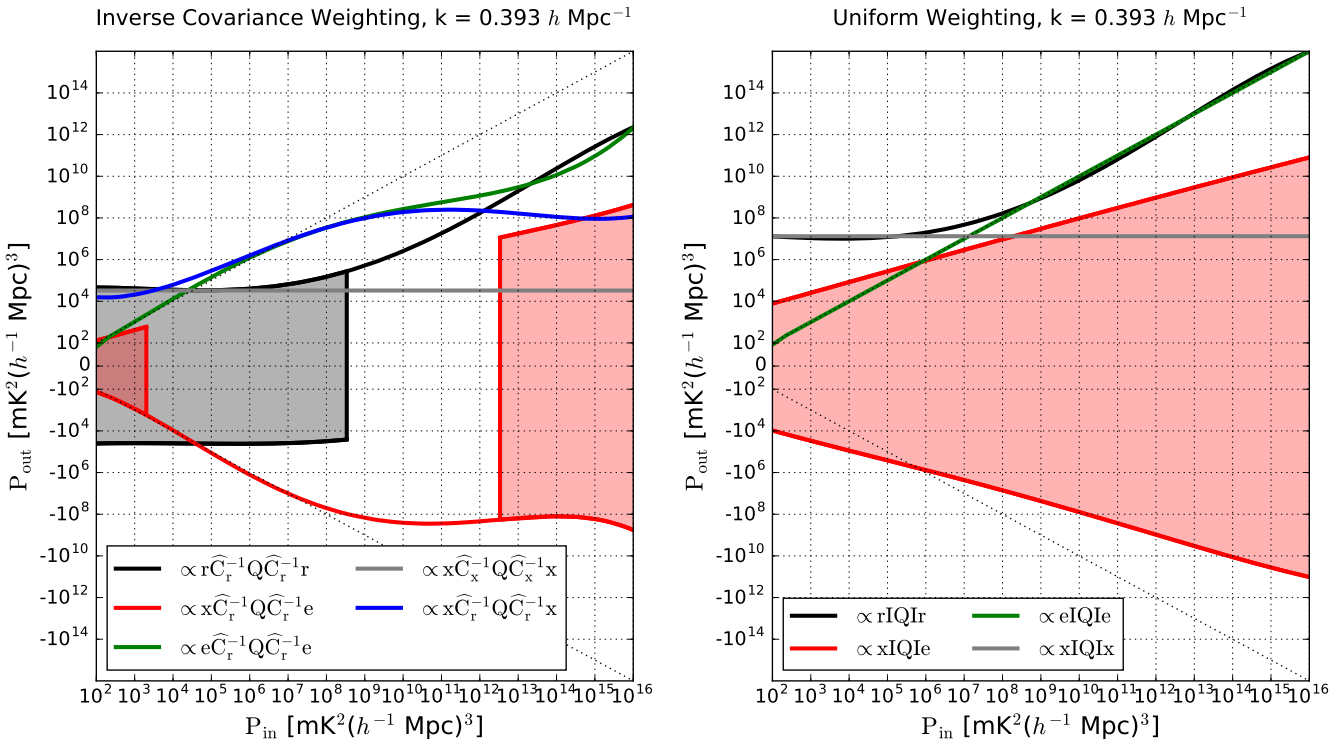

Figure 17. Illustration of the power spectrum amplitude of various power spectrum terms as a function of injected EoR power level summed into the data. The net output ($P_{\rm out} = \widehat{\mathbf{P}}_r$) is shown in black, the individual terms, as defined by Equation (15), are shown in blue, red, green; the dotted black line indicates 1:1 input to output mapping. The empirically estimated inverse covariance weighted case used in A15 is on the left and the uniform-weighted case is on the right. The details of the simulation used to generate the figure is explained in Section 3.1.2; here we fit smooth polynomials to our data points to make an illustrative example. The gray horizontal line is the power spectrum value of data alone (doesn't depend on injected power). The green signal-signal component is the term used in A15 to estimate signal loss. It is significantly higher than the net output because the cross-terms (red) are large and negative (black = green + red + blue) . In the regime where cross-correlations between injection and data are not dominant, the cross-terms have a noise-like term with width $\sqrt{P_e}\sqrt{P_x}$. However, at power levels comparable to the data (the middle region), the cross-terms can produce large, negative signal due to the couplings between $\mathbf{x}$ and $\mathbf{e}$ which affect $\widehat{\mathbf{C}}_r$. This causes the difference between the green curve (which exhibits negligible loss at the data level) and the black curve (which exhibits ~3 orders of magnitude of loss at the data level for the weighted case).

axis, and smooth the distribution of $P_{\rm out}$ values using kernel density estimators (Scott 2008) for each bin (and normalizing them). Stitching all of them together results in a 2-dimensional smoothed transfer function, shown in Figure 18 for both the weighted (left) and unweighted (right) cases. Although we only show figures for one k-value, we note that the shape of the transfer curve is similar for all k's. This transfer function is the likelihood function in Bayes' rule, namely $\mathcal{L}(P_{\rm out}|P_{\rm in})$. We then invoke Bayes' interpretation and re-interpret it as the posterior $p(P_{\rm in}|P_{\rm out})$ where we recall that $P_{\rm out}$ is a model of our data. To do this we make a horizontal cut across at the data value $\widehat{\mathbf{P}}_x$ (setting $P_{\rm out} = \widehat{\mathbf{P}}_x$) to yield a posterior distribution for the signal. We compute the 95% confidence interval (an upper limit on EoR) from these distributions.

By-eye inspection of the transfer function in Figure 18 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 10^4 mK^2 (h^{-1} Mpc)^3, for example, would map approximately to an upper limit of $\sim 10^8$ mK^2 (h^{-1} Mpc)^3, implying a signal loss factor of $\sim 10^4$. For the uniform-weighted case (right plot), we see no loss at a data value of $\sim 10^7$ mK^2 (h^{-1} Mpc)^3.

The loss-corrected power spectrum result for empirically estimated inverse covariance weighted PAPER-64 data is shown in Figure 19 (black dashed), which we can compare to the original lossy result (black and gray points). Post-signal loss estimation, the power spectrum limits are higher than both the theoretical noise level (green) and uniform-weighted power spectrum (blue dashed). 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.

3.1.3. *Minimizing Signal Loss*

With a signal loss formalism established, we now have the capability of experimenting with different weighting

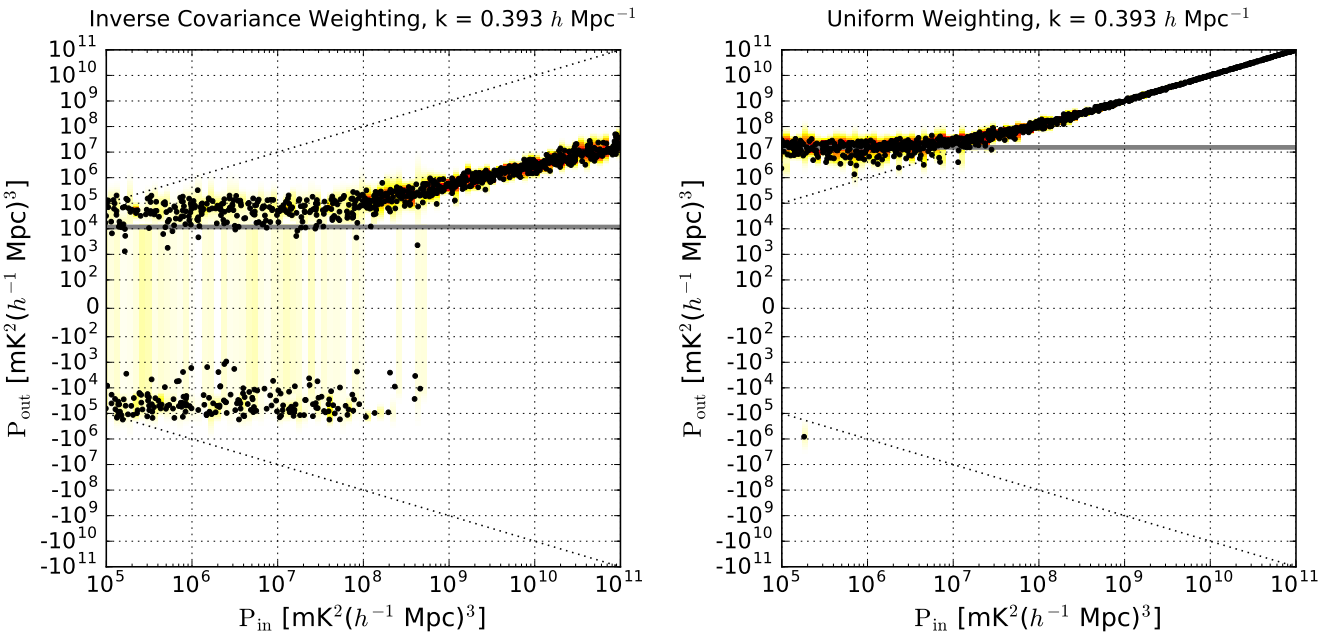

Figure 18. Signal loss transfer functions showing the relationship of $P_{\rm in}$ and $P_{\rm out}$, as defined by Equations (12) and (13). Power spectra values (black points) are generated for 20 realizations of $\mathbf{e}$ per signal injection level. Kernel density estimations of the power spectrum transfer functions are shown as colored heat-maps on top of the points for the cases of 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 level, 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 17 for comparison.

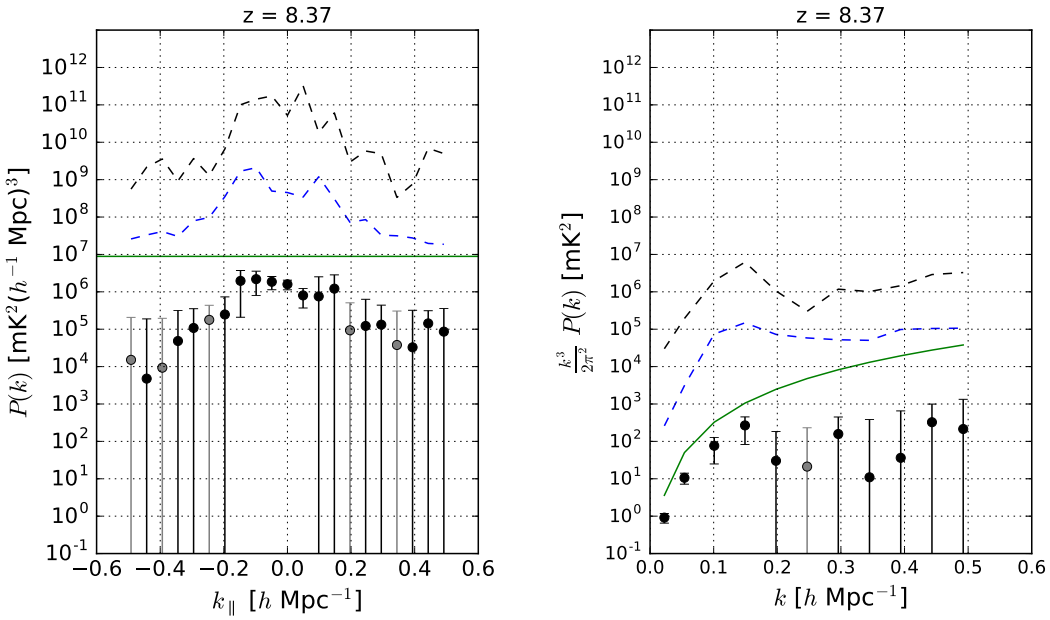

Figure 19. The power spectrum of PAPER-64 data, weighted by an inverse covariance estimator. The power spectra values pre-signal loss estimation are shown as black and gray points (positive and negative values, respectively, with 2σ error bars). The dashed black curve is the 2σ upper limit on the EoR signal post-signal loss estimation. The dashed blue line is the uniform-weighted power spectrum (2σ upper limit). The solid green line is the theoretical 2σ noise level prediction based on observational parameters, whose calculation is detailed in Section 3.2.1.

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 Section 2.1.3 how limiting the number of down-weighted eigenmodes (i.e. flattening out part of the eigenspectrum and effectively de-coupling the "weak" eigenmodes 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 for one k-value, both before and after signal loss estimation, are shown in the top panel in Figure 20. We see that the amount of signal loss increases as weighting becomes more aggressive (gray curve). In other words, more "weak" (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 (black curve). 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 not weighting (blue dashed line). 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 20 that this occurs when only the strongest ~ 2 eigenmodes are down-weighted and the rest are given equal weights. For a similar discussion on projecting out modes (zero-ing out eigenmodes, rather than just ignoring their relative weightings as we do in this study), see Switzer et al. (2013).

We also saw in Section 2.1.3 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 20. The gray and black 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 the 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 black and gray 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 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 20 in both panels and represents a third regularization technique. This line (red) 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 (black solid, black dashed, red) perform similarly at their best (i.e. when ~ 2 eigenmodes are down-weighted in the case of the black solid curve). However, for the remainder of this paper, we choose to use the weighting option of $\widehat{\mathbf{C}} \equiv \widehat{\mathbf{C}} + 0.035\,\mathrm{Tr}(\widehat{\mathbf{C}})\mathbf{I}$, or $\gamma = 0.035$, which we will denote as $\widehat{\mathbf{C}}_{\mathrm{eff}}$. We choose this weighting scheme merely as a simple example of regularizing PAPER-64 covariances, though we note that in Kolopanis et al. (*submitted*), we choose the most optimal regularization parameter for our final limits for every redshift bin independently. We also note that all of the regularizations described here perform similarly to the uniformly weighted case. However, we find that the regularization of empirical covariances does outperform the uniform weighted case for redshift bins in which there are stronger foregrounds, as presented in Kolopanis et al. (*submitted*).

A revised PAPER-64 power spectrum (using only one baseline separation type and $\widehat{\mathbf{C}}_{\mathrm{eff}}$) is shown in Figure 21. Again, the dashed black curve represents our upper limit on EoR, post-signal loss estimation. Black and gray points correspond to positive and negative power spectrum values respectively (pre-signal loss estimation), with 2σ errors bars. Also plotted are the uniform-weighted power spectrum upper limit (dashed blue) and theoretical prediction of noise (solid green). In Kolopanis et al. (*submitted*), a larger parameter space is explored for regularization and multiple baselines are used in the analysis to produce a best 2σ power spectrum upper limit of $(240.0\ \mathrm{mK})^2$ at $k = 0.3\ h\ \mathrm{Mpc}^{-1}$ and $z = 8.37$, a higher limit than A15 by a factor of $\sim$10 in mK.

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 could 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). Alternately, covariances

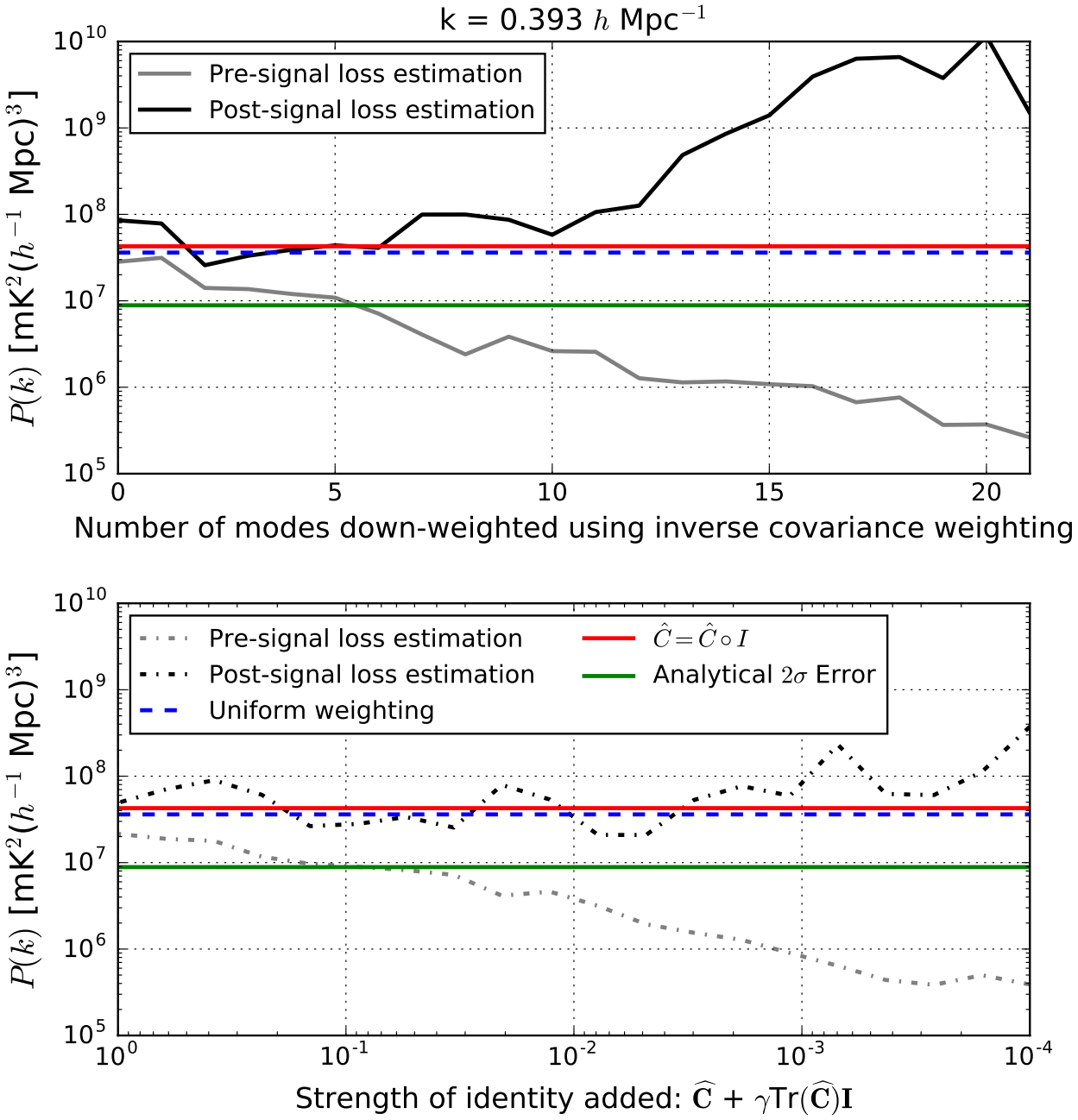

Figure 20. Power spectra 2σ upper limits for $k = 0.393\,h$ Mpc^{-1} for fringe-rate filtered PAPER-64 data. Top: Values are shown before (gray) and after (black) signal loss estimation as a function of number of eigenmodes of $\widehat{\mathbf{C}}$ that are down-weighted. This regularization knob is tuned from 0 modes on the left (i.e. unweighted) to 21 modes on the right (i.e. the full inverse covariance estimator). Over ~ 3 orders of magnitude of signal loss results when using empirically estimated inverse covariance weighting. Bottom: Power spectrum upper limits before (gray) and after (black) 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 (dashed blue) and inverse variance weighted case (red); both are after signal loss estimation. Finally, a theoretical prediction for noise (2σ error) is plotted as solid green. In the revised PAPER-64 analysis in this paper, we choose to use a regularization scheme of $\widehat{\mathbf{C}}_{\rm eff} \equiv 0.035\,{\rm Tr}(\widehat{\mathbf{C}})\mathbf{I} + \widehat{\mathbf{C}}$ ($\gamma = 0.035$) as a simple example of regularization that minimizes loss.

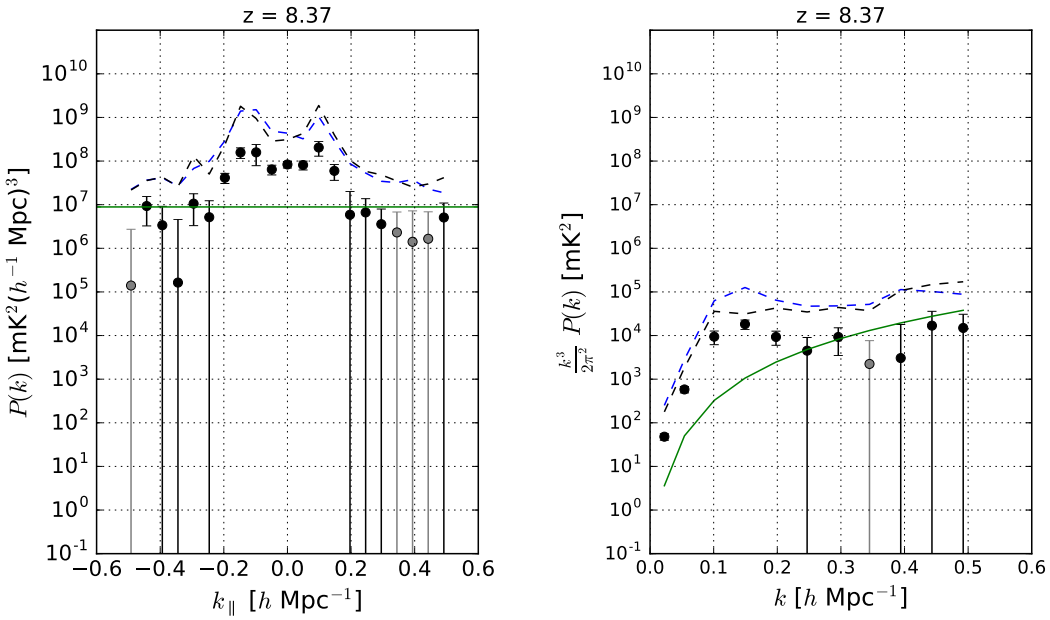

Figure 21. Power spectrum of PAPER-64 using $\widehat{\mathbf{C}}_{\rm eff}$. The weighted power spectrum, pre-signal loss estimation, is shown by the black and gray points (corresponding to positive and negative power spectrum values, respectively, with 2σ error bars). The dashed black curve is the 2σ upper limit on the EoR signal post-signal loss estimation. There is a small amount of loss using this regularization. The dashed blue line is the uniform-weighted power spectrum (2σ upper limit). The solid green line is the theoretical 2σ noise level prediction based on observational parameters. This power spectrum result differs from A15 in that it only uses data from one type of baseline (30 m East/West baselines) instead of three. Major differences from previously published results stem from revisions regarding signal loss, bootstrapping, and the theoretical error computation. In Kolopanis et al. (*submitted*), a larger parameter space is explored for regularization and multiple baselines are used in the analysis to produce a best 2σ power spectrum upper limit of $(240.0\ \mathrm{mK})^2$ at $k = 0.3\ h\ \mathrm{Mpc}^{-1}$ and $z = 8.37$, a higher limit than A15 by a factor of $\sim$10 in mK.

could be estimated from all other 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. (2015)) 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 Section 3.2, so we also regard those results as suspect and superseded by those of Kolopanis et al. (*submitted*)).

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

In Section 3.1, 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 paper, we will transition to other new aspects of our analysis, framed within the context of error estimation and (non-EoR) bias in PAPER-64.

3.2. *PAPER-64: Error Estimation*

In this section we discuss the ways in which we estimate errors for PAPER-64 power spectra. We first walk through an expression for a theoretical error estimation (of thermal noise) based on observational parameters. Although a theoretical model often differs from true errors as explained in Section 2.2, it is helpful to understand the ideal case and the factors that affect its sensitivity. Additionally, we build on the lessons learned about bootstrapping in Section 2.2 to revise our bootstrapping method as applied to PAPER-64 data in order to compute accurate errors from the data itself.

In particular, we highlight major changes in both our sensitivity calculation and bootstrapping method that differ from the A15 analysis of PAPER-64. While we do not discuss the changes within the context of PAPER-32, it is worth noting that the power spectrum results in Parsons et al. (2014) are affected by the same issues.

3.2.1. *Theoretical Error Estimation*

Re-analysis of the PAPER-64 data included a detailed study using several independently generated noise simulations. What we found was that these simulations all agreed but were discrepant with the previous analytic sensitivity calculations. The analytic calculation is only an approximation; however, the differences were large enough (factors of 10 in some cases) to warrant a careful investigation. The analytic calculation attempts to combine a large number of pieces of information in an approximate way, and when re-considering some of the approximations, we have found there to be large effects. What follows here is an accounting of the differences which have been discovered. Our revised theoretical error estimation, which is often plotted as the solid green curve in many of the previous power spectrum plots, is computed with these changes accounted for.

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

$$N(k) = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{\sqrt{2 N_{\rm lst} N_{\rm seps}}\, t_{\rm int} N_{\rm days} N_{\rm bls} N_{\rm pols}}. \tag{20}$$

We will now explain each factor in Equation (20) and highlight key differences from the numbers used in A15.

- 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:

$$T_{\rm sys} = 180\left(\frac{\nu}{0.18}\right)^{-2.55} + T_{\rm rcvr}, \tag{21}$$

  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 mis-calculation 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

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

independent "even" and "odd" visibility measurements (this cross-multiplication is done principally to avoid a noise bias). 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 paper, we only use one type of baseline (PAPER's 30 m East/West baselines), hence $N_{\rm seps} = 1$. The updated limits in Kolopanis et al. (*submitted*) use three separation types.

- $t_{\rm int}$: 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:

$$t_{\rm int} = t_i \frac{\int 1\, df}{\int w^2(f)\, df}, \tag{22}$$

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.

- $N_{\rm days}$: 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" datasets), this number should be reduced by a factor of 2. Additionally, because our LST coverage is not 100% complete (it doesn't overlap for every single day), we compute a more realistic value of $N_{\rm days}$ as:

$$\frac{1}{N_{\rm days}} = \sqrt{\left\langle \frac{1}{N_i^2} \right\rangle_i}, \tag{23}$$

where i indexes LST (Jacobs et al. 2015). For PAPER-64, our revised estimate of $N_{\rm days}$ is ~ 34 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 ($N_{\rm bls\_total}$) neglects the fact that we average baselines into groups for computational speed-up when cross-multiplying data. Our revised estimate for the parameter is:

$$N_{\rm bls} = \frac{N_{\rm bls\_total}}{N_{\rm gps}} \sqrt{N_{\rm gps}^2 - N_{\rm gps}}, \tag{24}$$

where, as with 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 ($\sqrt{N_{\rm gps}^2 - N_{\rm gps}}$). For the revised PAPER-64 analysis with only one baseline separation type, this becomes $N_{\rm bls} \sim 46$ instead of 51.

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

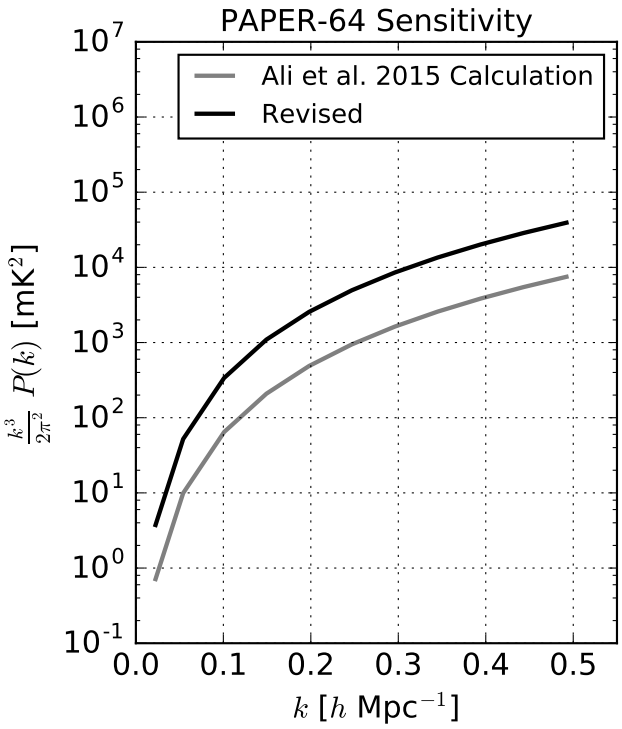

Figure 22. An updated prediction for the thermal noise level of PAPER-64 data (black) is shown in comparison to previously published sensitivity limits (gray). Both sensitivity analyses plotted assume only one baseline type (an additional factor of $\sqrt{3}$ for three baseline types is needed to match A15 exactly). Major factors that contribute to the discrepancy are $\Omega_{\rm eff}$, $N_{\rm days}$ and $N_{\rm bls}$, as in Equation (20) and described in Section 3.2.1, which when combined decreases our sensitivity (higher noise floor) by a factor of ~ 5 in mK^2.

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

Our revised sensitivity estimate for PAPER-64 is shown in comparison with that of A15 in Figure 22. Together, the revised parameters yield a decrease in sensitivity (higher noise floor) by a factor of ~ 5 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 (21)) but accounting for the true $N_{\rm days}$ as determined by LST 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:

$$T_{\rm rms} = \frac{T_{\rm sys}}{\sqrt{\Delta\nu \Delta t N_{\rm days} N_{\rm pols}}}, \quad (25)$$

where $\Delta\nu$ is channel spacing, Δt is 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 23, where the black and gray points represent pre-signal loss estimated positive and negative power spectrum values, respectively (with 2σ error bars and weighting matrix $\widehat{\mathbf{C}}_{\rm eff}$), the dashed black line represents the post-signal loss 2σ upper limit on the EoR signal, the dashed blue line represents the 2σ uniform-weighted power spectrum limit, and the solid green line denotes our 2σ theoretical noise prediction as calculated by Equation (20). All three show good agreement, validating our analytical thermal noise calculation.

3.2.2. *Bootstrapping*

We bootstrap PAPER-64 power spectra in order to determine confidence intervals for our results. In this section, we highlight one major change in the way we estimate errors since A15, using the lesson we have learned about bootstrapping independent samples.

As discussed in Section 2.2, bootstrapping is only a valid way of estimating errors if a dataset is comprised of independent samples, or the number of independent samples is well known. 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. A total of 400 bootstraps were created in this step ($N_{\rm boot} = 400$), each comprised of randomly selected values sampled with replacement along the time axis. More specifically, 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, under-estimating errors. A random draw of 700 measurements from this dataset has many repeated values, and the variance between hundreds ($N_{\rm boot}$) 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 baselines. Power spectrum 2σ errors with this bootstrapping change for fringe-rate filtered noise are shown in Figure 24. 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 errors (black) are shown to better agree with the analytic prediction for noise. We note that they do not match perfectly, but we find good agreement between them to within a factor of two, which we believe is reasonable given the approximations we make in computing the analytic expression for thermal noise. While Figure 24 implies that errors are under-estimated by a factor of $\sim$7 in mK^2 for the noise simulation, in practice this factor is slightly lower for the case of real data (a factor of $\sim$5 in mK^2 instead), possibly due to the data being less correlated in time than the fringe-rate filtered noise in the simulation.

Finally, one last change from the A15 method is that our power spectrum points (previously 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 over the baseline axis (all baselines are used), and this estimate is propagated through our signal loss computation. 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 maximum likelihood estimate of our data.

3.3. *PAPER-64: Bias*

In Section 2.3 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 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.

3.3.1. *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 Section 2.1, we have a choice for the normalization matrix $\mathbf{M}$ in Equation (2). For the analysis of PAPER-64

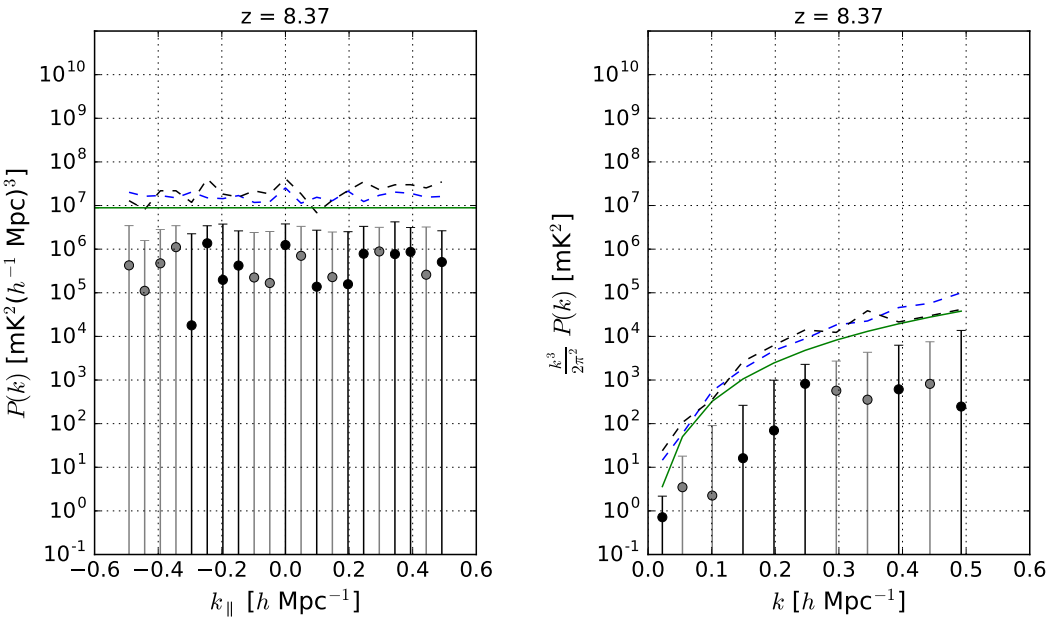

Figure 23. Power spectrum estimate for a noise simulation that mimics the noise level of PAPER-64 data. The original weighted power spectrum points and their 2σ errors are show in black and gray (positive and negative values), where we use $\widehat{\mathbf{C}}_{\rm eff}$ to minimize signal loss. The dashed black line is the 2σ upper limit on EoR post-signal loss estimation. The dashed blue line is the uniform-weighted power spectrum 2σ upper limit. The solid green line is the theoretical 2σ noise level prediction as calculated by Equation (20). All three estimates show good agreement.

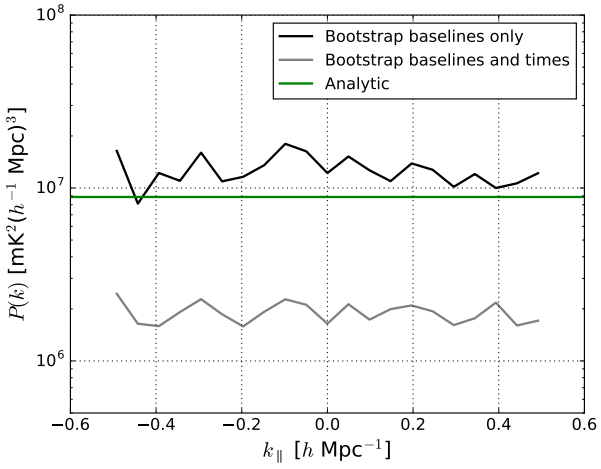

Figure 24. 2σ power spectrum errors for a noise simulation (computed via Equation (25) using PAPER-64 observing parameters) using two different bootstrapping methods. The noise is fringe-rate filtered and a weighting matrix of $\mathbf{I}$ (uniform-weighted) is used in order to disentangle the effects of bootstrapping from signal loss. The bootstrapping method used in A15 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 dataset (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. While this does not match perfectly with the 2σ analytic prediction for noise (green), we find good agreement between them to within a factor of two (which we believe is reasonable given the approximations we make in computing the analytic).

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

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{tr}[\mathbf{R}\mathbf{Q}^\alpha\mathbf{R}\mathbf{Q}^\beta] \tag{26}$$

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

$$\mathbf{F} = \mathbf{L}\mathbf{L}^\dagger. \tag{27}$$

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

$$\mathbf{M} = \mathbf{D}\mathbf{L}^{-1} \tag{28}$$

where $\mathbf{D}$ is a diagonal matrix. In doing so, our window function, defined as $\mathbf{W} = \mathbf{M}\mathbf{F}$, becomes:

$$\mathbf{W} = \mathbf{D}\mathbf{L}^\dagger. \tag{29}$$

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 $\mathbf{F}$ 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 paper, 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 (1). Namely, the PAPER-64 dataset is divided into two halves: even Julian dates and odd Julian dates. Our data vectors are then $\mathbf{x}_{even,1}$ for the "even" dataset and baseline group 1, $\mathbf{x}_{odd,1}$ for the "odd" dataset and baseline group 1, etc. We only form $\widehat{q}$ when the two copies of $\mathbf{x}$ come from different groups and baselines, never multiplying "even" with "even", for example, in order to prevent the squaring of the same thermal noise. We recognize that this method results in a hit to our sensitivity. Sensitivity can be gained, for example, by only dividing up the dataset once (either along time or baselines), or by creating more groups (more cross-multiplications), but in this paper we only attempt to revise the A15 analysis (which uses the same groupings), not produce the most sensitive limits of PAPER-64.

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 16 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, detections of bias exist at all k's without crosstalk removal and these are detections that, depending on the power spectrum level, could be mistaken for EoR.

3.3.2. *Jackknife/Null Tests*

As shown in Figure 21, our re-analysis of the PAPER-64 power spectrum shows biases above the predicted noise level, particularly at low-k values. As discussed in Section 2.3.1, this bias is most likely attributable to foreground leakage. The cause for biases at higher k-values is more difficult to pinpoint.

Here we perform three hybrid jackknife/null tests on PAPER-64 data that aim to isolate systematics in the data and verify that they are not attributable to EoR. Similar to in Section 2.3.2, we take jackknives along different axes of the dataset to produce multiple power spectra. In addition, we difference them to tease out excess variances, which can be thought of as a type of null test.

The three results are shown in Figure 25. Each test displays the differenced power spectrum between two halves of a jackknife. We take jackknives along three different axes:

- Baselines: We split our dataset into two halves, where each contains half of the total baselines used in the analysis. No baselines are repeated between the two datasets.
- Sidereal Hour: We split our dataset into two halves based on LST, namely the first half (LSTs 0.5-4.6 hours) and second half (LSTs 4.6-8.6 hours).
- Day: We split our dataset 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" datasets, we expect it to cancel out.

In investigating Figure 25, we focus on two main aspects of each test — whether the error bars are consistent with zero, and whether they are consistent with thermal noise. The former indicates a 'passing' null test where there is no bias along the jackknife axis. The latter reveals whether there is an additional noise bias present that lies above the thermal noise.

We find that the baselines test is fully consistent with noise, as all error bars pass through zero. Therefore, we conclude that particular baselines are not to be blamed for the bias in our dataset. For the LST test, 90% of all the error bars pass through zero, which is slightly lower than the expected 95%. This indicates a low level bias that is LST-dependent, and likely caused by residual foregrounds that vary in LST. In other words, the two jackknife halves differ by an amount greater than thermal noise and do not completely cancel out the sky signal when differencing. The even/odd test shows 81% of all error bars being consistent with zero, again indicating the presence of a low level bias. The cause of this particular systematic is not obvious, and requires further investigation that we do not carry out in this paper. Most importantly, a clean detection of EoR would have passed all three tests, which is clearly not what we see.

Finally, we note that all three null tests exhibit error bars that are strongly k-dependent. The errors are largest just outside the horizon ($k \sim \pm 0.1\,h$ Mpc^{-1}) which is expected since foregrounds are brightest in this region post-delay filtering. We also note that all three tests, whether biased or unbiased, result in power spectrum errors that are slightly larger than the estimated thermal noise (gray band) at high-k. This indicates an additive noise component that is increasing our errors (while not affecting whether they pass through zero or not).

There are a couple possible explanations for this. 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, 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. The cross-terms, however, can introduce a noise bias that would vary with foreground strength, therefore be-

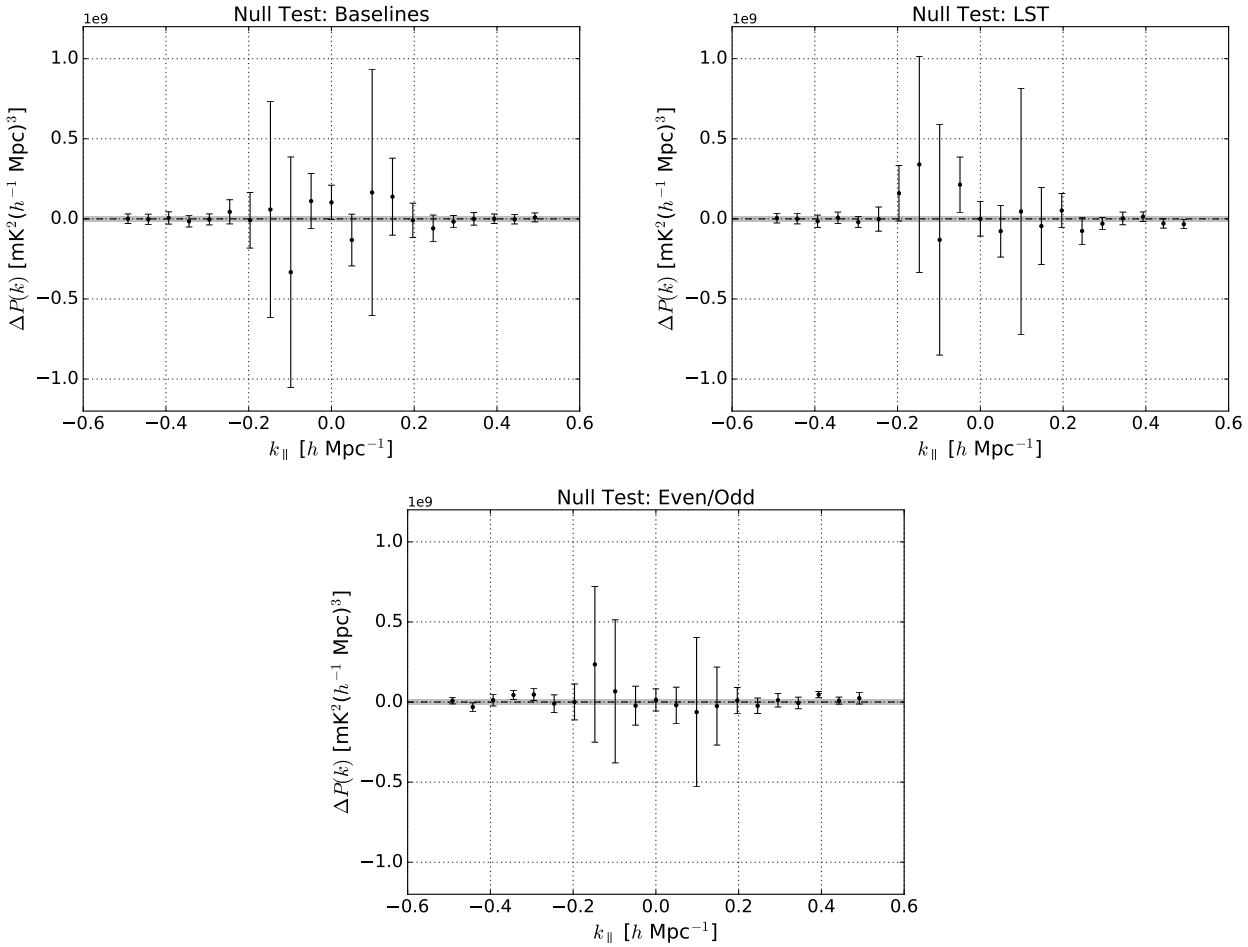

Figure 25. Differenced power spectrum results (with 2σ errors) for three null tests, where a jackknife is taken along the baseline axis (top left), LST axis (top right), and even/odd Julian date axis (bottom). The results shown are unweighted (no signal loss), where the power spectrum values plotted 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 that differ between the tests, though there are low level biases at high-k values in both the LST test and even/odd test. All three tests, whether biased or not, have an extra noise bias that is strongly k-dependent. Residual foregrounds dominate at k-values near the horizon ($k \sim \pm 0.1\, h$ Mpc^{-1}), which is not surprising given that foregrounds are brightest in this region post-delay filtering.

ing consistent with the shape of our results. That being said, an alternate explanation for the large error bars seen at high k's could simply be that our estimate of the thermal noise is too low, as all three tests seem to imply a larger, but consistent, estimate.

4. CONCLUSION

Although current 21 cm published power spectrum upper limits lie several orders of magnitude above predicted EoR levels, ongoing analyses of deeper sensitivity datasets 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, be able to accurately and robustly calculate both power spectrum and theoretical errors, and consistently undergo a comprehensive set of jackknife and null tests.

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

- Substantial signal loss can result when weighting data using empirically estimated covariances (Section 2.1). Loss of the 21 cm signal is especially significant the fewer number of independent modes that exist in the data. Hence, there exists a tradeoff 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 (Section 3.1). However, a signal-only simulation (i.e. comparing a uniformly weighted vs. weighted power spectrum of EoR only) can under-estimate loss by failing to

account for correlations between the data and signal (Section 3.1.1).

- Errors that are estimated via bootstrapping can be under-estimated if samples in the dataset are significantly correlated (Section 2.2). However, if the number of independent samples in a dataset is well-determined, bootstrapping is a simple and accurate way of estimating errors.

- Meaningful null tests are vital to validate an EoR detection (Section 2.3.2). Similarly, performing jackknife tests along multiple axes of a dataset is necessary for confidence in an EoR detection and can also be used to tease out systematics.

As a consequence of our investigations, we have also revised power spectrum results from PAPER-64. In Kolopanis et al. (*submitted*), we quote a best 2σ power spectrum upper limit of $(240.0\,\mathrm{mK})^2$ at $k = 0.3\,h\,\mathrm{Mpc}^{-1}$ and $z = 8.37$, a higher limit than A15 by a factor of $\sim$10 in mK. The reasons for a previously under-estimated limit and ways in which our new analysis differs can be summarized by the following:

- Signal loss, previously found to be $< 2\%$ in A15, was under-estimated by a factor of >1000 for empirically estimated inverse covariance weighting. For our new analysis, we use a regularized covariance weighting method to minimize loss (Section 3.1.3). However, because our revised weighting method is not as aggressive as the former, our results are still a factor of ~ 10 in mK higher than previous limits. Under-estimated signal loss therefore represents the bulk of our revision. This revision is similar to the re-analysis of results from the GMRT (Paciga et al. 2013a) which were also revised from new signal loss calculations associated with their singular value decomposition foreground filter.

- Power spectrum errors, originally computed by bootstrapping, were under-estimated 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 (Section 3.2.2).

- 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 ~ 2 in mK (Section 3.2.1). We note that our sensitivity prediction is revised by a factor less than our 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, precise error calculations, and comprehensive jackknife tests 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.

5. ACKNOWLEDGEMENTS

CC would like to acknowledge the UC Berkeley Chancellor's Fellowship and National Science Foundation Graduate Research Fellowship (Division of Graduate Education award 1106400). She would also like to thank Eric Switzer, Miguel Morales, and Bryna Hazelton for helpful discussions. PAPER and HERA are supported by grants from the National Science Foundation (awards 1440343, and 1636646). ARP, DCJ, and JEA would also like to acknowledge NSF support (awards 1352519, 1401708, and 1455151, respectively). SAK is supported by the University of Pennsylvania School of Arts and Sciences Dissertation Completion Fellowship. JSD acknowledges NSF AAPF award 1701536. AL acknowledges support for this work by NASA through Hubble Fellowship grant #HST-HF2-51363.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. We graciously thank SKA-SA for site infrastructure and observing support.

APPENDIX

A. A TOY MODEL FOR SIGNAL LOSS

In this Appendix, 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 paper, 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

$$\widehat{P}_\alpha = \frac{1}{2\mathbf{F}_{\alpha\alpha}} \mathbf{x}^t \mathbf{C}^{-1} \mathbf{Q}^\alpha \mathbf{C}^{-1} \mathbf{x}, \tag{A1}$$

where

$$F_{\alpha\alpha} \equiv \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Q}^{\alpha}\mathbf{C}^{-1}\mathbf{Q}^{\alpha}\right) \tag{A2}$$

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

$$\widehat{P}_{\alpha}^{\mathrm{loss}} = \frac{1}{2\widehat{\mathbf{F}}_{\alpha\alpha}}\mathbf{x}^{t}\widehat{\mathbf{C}}^{-1}\mathbf{Q}^{\alpha}\widehat{\mathbf{C}}^{-1}\mathbf{x}, \tag{A3}$$

where

$$\widehat{F}_{\alpha\alpha} \equiv \frac{1}{2}\mathrm{tr}\left(\widehat{\mathbf{C}}^{-1}\mathbf{Q}^{\alpha}\widehat{\mathbf{C}}^{-1}\mathbf{Q}^{\alpha}\right), \tag{A4}$$

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

$$\widehat{\mathbf{C}} = (1-\eta)\mathbf{C} + \eta\mathbf{x}\mathbf{x}^{t}, \tag{A5}$$

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 (A5) into Equation (A3) 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

$$\widehat{\mathbf{C}}^{-1} = \frac{\mathbf{C}^{-1}}{1-\eta}\left[\mathbf{I} - \frac{\eta\mathbf{x}\mathbf{x}^{t}\mathbf{C}^{-1}}{1+\eta(g-1)}\right], \tag{A6}$$

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

$$\widehat{F}_{\alpha\alpha} = \frac{F_{\alpha\alpha}}{(1-\eta)^2}\left[1 - \frac{\eta}{1+\eta(g-1)}\frac{h_{\alpha\alpha}}{F_{\alpha\alpha}} + \frac{1}{2}\left(\frac{\eta}{1+\eta(g-1)}\right)^2\frac{h_{\alpha}^2}{F_{\alpha\alpha}}\right], \tag{A7}$$

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

$$\widehat{P}_{\alpha}^{\mathrm{loss}} = \frac{1}{2}\frac{h_{\alpha}}{F_{\alpha\alpha}}\left[1 - \frac{\eta g}{1+\eta(g-1)}\right]^2\left[1 - \frac{\eta}{1+\eta(g-1)}\frac{h_{\alpha\alpha}}{F_{\alpha\alpha}} + \frac{1}{2}\left(\frac{\eta}{1+\eta(g-1)}\right)^2\frac{h_{\alpha}^2}{F_{\alpha\alpha}}\right]^{-1}. \tag{A8}$$

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

$$\widehat{P}_{\alpha}^{\mathrm{loss}} \approx \frac{1}{2}\frac{h_{\alpha}}{F_{\alpha\alpha}}\left[1 - \eta\left(g - \frac{h_{\alpha\alpha}}{F_{\alpha\alpha}}\right)\right]. \tag{A9}$$

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

$$\langle\widehat{P}_{\alpha}^{\mathrm{loss}}\rangle \approx (1-\eta N_f)p_{\alpha} + 4\eta\frac{\mathrm{tr}(\mathbf{C}^{-1}\mathbf{Q}^{\alpha}\mathbf{C}^{-1}\mathbf{Q}^{\alpha}\mathbf{C}^{-1}\mathbf{Q}^{\alpha})}{\left[\mathrm{tr}(\mathbf{C}^{-1}\mathbf{Q}^{\alpha}\mathbf{C}^{-1}\mathbf{Q}^{\alpha})\right]^2} \approx (1-\eta N_f)p_{\alpha}, \tag{A10}$$

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.

REFERENCES

Ade, P., et al. 2008, Astrophysical Journal, 674, 22
Ade, P. A. R., et al. 2017, PhRvD, 96, 102003
Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166
Ali, Z. S., et al. 2015, ApJ, 809, 61
Andrae, R. 2010, ArXiv e-prints
Araujo, D., et al. 2012, The Astrophysical Journal, 760, 145
Barkana, R., & Loeb, A. 2001, PhR, 349, 125
—. 2008, Monthly Notices of the Royal Astronomical Society, 384, 1069
Beardsley, A. P., et al. 2016, The Astrophysical Journal, 833, 102
Bernardi, G., et al. 2009, A&A, 500, 965
—. 2010, A&A, 522, A67+
Bernardi, G., et al. 2013, The Astrophysical Journal, 771, 105
Bernardi, G., et al. 2016, MNRAS, 461, 2847
BICEP2 Collaboration et al. 2016, ApJ, 833, 228
Bischoff, C., et al. 2011, The Astrophysical Journal, 741, 111
Bond, J. R., Jaffe, A. H., & Knox, L. 1998, PhRvD, 57, 2117
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., 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., et al. 2012, Monthly Notices of the Royal Astronomical Society, 423, 2518
Chiang, H. C., et al. 2010, The Astrophysical Journal, 711, 1123
Crites, A. T., et al. 2015, The Astrophysical Journal, 805, 36
Das, S., et al. 2011a, Physical Review Letters, 107, 021301
—. 2011b, ApJ, 729, 62
Datta, A., Bowman, J. D., & Carilli, C. L. 2010, The Astrophysical Journal, 724, 526
de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., Jonas, J., Landecker, T. L., & Reich, P. 2008, MNRAS, 388, 247
DeBoer, D. R., et al. 2017, Publications of the Astronomical Society of the Pacific, 129, 045001
Dillon, J. S., Liu, A., & Tegmark, M. 2013, PhRvD, 87, 043005
Dillon, J. S., & Parsons, A. R. 2016, The Astrophysical Journal, 826, 181
Dillon, J. S., et al. 2014, Phys. Rev. D, 89, 023002
—. 2015, Phys. Rev. D, 91, 123011
Dodelson, S., & Schneider, M. D. 2013, PhRvD, 88, 063537
Efron, B., & Tibshirani, R. 1994, An Introduction to the Bootstrap, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis)
Ewall-Wice, A., Dillon, J. S., Liu, A., & Hewitt, J. 2017, MNRAS, 470, 1849
Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, PhR, 433, 181
Ghosh, A., Bharadwaj, S., Ali, S. S., & Chengalur, J. N. 2011, MNRAS, 418, 2584
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Jacobs, D. C., et al. 2015, ApJ, 801, 51
Jacobs, D. C., et al. 2016, ApJ, 825, 114
Jelić, V., et al. 2008, MNRAS, 389, 1319
Joachimi, B. 2017, MNRAS, 466, L83
Keating, G. K., Marrone, D. P., Bower, G. C., Leitch, E., Carlstrom, J. E., & DeBoer, D. R. 2016, The Astrophysical Journal, 830, 34
Kerrigan, J., et al. 2018, ArXiv e-prints
Kohn, S. A., et al. 2016, ApJ, 823, 88
Koopmans, L., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 1
Liu, A., & Parsons, A. R. 2016, Monthly Notices of the Royal Astronomical Society, 457, 1864
Liu, A., Parsons, A. R., & Trott, C. M. 2014a, PhRvD, 90, 023018
—. 2014b, PhRvD, 90, 023019
Liu, A., & Tegmark, M. 2011, Phys. Rev. D, 83, 103006
Loeb, A., & Furlanetto, S. 2013, The First Galaxies in the Universe (Princeton University Press)
Masui, K. W., et al. 2013, ApJL, 763, L20
Moore, D. F., Aguirre, J. E., Parsons, A. R., Jacobs, D. C., & Pober, J. C. 2013, The Astrophysical Journal, 769, 154
Morales, M. F., & Wyithe, J. S. B. 2010, ARA&A, 48, 127
Paciga, G., et al. 2013a, MNRAS
—. 2013b, 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. 2012a, ApJ, 753, 81
Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51
Parsons, A. R., Pober, J. C., Aguirre, J. E., Carilli, C. L., Jacobs, D. C., & Moore, D. F. 2012b, ApJ, 756, 165
Parsons, A. R., et al. 2010, AJ, 139, 1468
—. 2014, ApJ, 788, 106
Patil, A. H., et al. 2016, MNRAS, 463, 4317
—. 2017, ApJ, 838, 65
Patra, N., Subrahmanyan, R., Sethi, S., Udaya Shankar, N., & Raghunathan, A. 2015, ApJ, 801, 138
Paz, D. J., & Sánchez, A. G. 2015, MNRAS, 454, 4326
Pearson, D. W., & Samushia, L. 2016, MNRAS, 457, 993
Peterson, U.-L. P. X.-P. W. J. 2004, ArXiv Astrophysics e-prints
Petrovic, N., & Oh, S. P. 2011, MNRAS, 413, 2103
Pober, J. C., et al. 2012, AJ, 143, 53
Pober, J. C., et al. 2013, The Astrophysical Journal Letters, 768, L36
Pober, J. C., et al. 2013, AJ, 145, 65
—. 2014, ApJ, 782, 66
—. 2016, ApJ, 819, 8
Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766
Pritchard, J. R., & Loeb, A. 2010, PhRvD, 82, 023006
—. 2012, Reports on Progress in Physics, 75, 086901
Quenouille, M. H. 1949, Ann. Math. Statist., 20, 355
Santos, M. G., Cooray, A., & Knox, L. 2005, ApJ, 625, 575
Scott, D. W. 2008, Multivariate Density Estimation: Theory, Practice, and Visualization, 125
Sellentin, E., & Heavens, A. F. 2016, MNRAS, 456, L132
Sherwin, B. D., et al. 2017, Phys. Rev. D, 95, 123529
Sokolowski, M., et al. 2015, PASA, 32, e004
Switzer, E. R., Chang, T.-C., Masui, K. W., Pen, U.-L., & Voytek, T. C. 2015, ApJ, 815, 51
Switzer, E. R., et al. 2013, MNRAS, 434, L46
Taylor, A., & Joachimi, B. 2014, MNRAS, 442, 2728
Tegmark, M. 1997, PhRvD, 55, 5895
Thompson, A. R., Moran, J. M., & Swenson, Jr., G. W. 2001, Interferometry and Synthesis in Radio Astronomy, 2nd Edition
Thyagarajan, N., et al. 2013, ApJ, 776, 6
Tingay, S. J., et al. 2013, PASA, 30, 7
Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101
Trott, C. M., et al. 2016, The Astrophysical Journal, 818, 139

Tukey. 1958, Ann. Math. Statist., 29, 614
van Haarlem, M. P., et al. 2013, A&A, 556, A2
Vedantham, H., Shankar, N. U., & Subrahmanyan, R. 2012, The Astrophysical Journal, 745, 176
Voytek, T. C., Natarajan, A., Jáuregui García, J. M., Peterson, J. B., & López-Cruz, O. 2014, ApJL, 782, L9
Wang, J., et al. 2013, The Astrophysical Journal, 763, 90
Wolz, L., Abdalla, F. B., Blake, C., Shaw, J. R., Chapman, E., & Rawlings, S. 2014, Monthly Notices of the Royal Astronomical Society, 441, 3271
Wu, X. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, 474