Draft version November 29, 2017
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], Saul A. Kohn[4], Adrian Liu[1,5], James E. Aguirre[4], Zaki S. Ali[1], Richard F. Bradley[6,7,8], Gianni Bernardi[9,10,11], Chris L. Carilli[12,13], David R. DeBoer[2], Matthew R. Dexter[2], Joshua S. Dillon[1], Pat Klima[7], David H. E. MacMahon[2], David F. Moore[4], Chuneeta D. Nunhokee[10], Jonathan C. Pober[14], William P. Walbrugh[9], Andre Walker[9]

[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]Dept. of Physics and Astronomy, U. Penn., Philadelphia PA
[5]Berkeley Center for Cosmological Physics, Berkeley, CA
[6]Dept. of Electrical and Computer Engineering, U. Virginia, Charlottesville VA
[7]National Radio Astronomy Obs., Charlottesville VA
[8]Dept. of Astronomy, U. Virginia, Charlottesville VA
[9]Square Kilometer Array, S. Africa, Cape Town South Africa
[10]Dept. of Physics and Electronics, Rhodes University, South Africa
[11]Harvard-Smithsonian Cen. for Astrophysics, Cambridge MA
[12]National Radio Astronomy Obs., Socorro NM
[13]Cavendish Lab., Cambridge UK
[14]Dept. of Physics, Brown University, Providence RI

ABSTRACT

The Epoch of Reionization (EoR) is an uncharted era in our Universe's history during which the birth of 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 and validate results. Additionally, it is imperative to understand the trade-offs between different methods and their effects on common power spectrum themes. In this paper, we focus on three major themes — signal loss, power spectrum error bar 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 mostly stems from an updated 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 very 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. 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 information about the intergalactic medium, revealing connections between the matter distribution exhibited via cosmic microwave background (CMB) studies and the highly structured web of galaxies we observe today.

One promising technique to probe the EoR is to target the 21 cm wavelength emission that is emitted and absorbed by neutral hydrogen via its spin-flip transition (Furlanetto et al. 2006; Morales & Wyithe 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.

Tracing the 21 cm line as a function of redshift offers a window into the evolution of ionization, temperature, and density fluctuations during this transformative era.

Although a detection of the EoR remains elusive, 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 Experiment to Detect the Global EoR Signature (EDGES; Bowman & Rogers 2010), the Large Aperture Experiment to Detect the Dark Ages (LEDA; Greenhill & Bernardi 2012), 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. 2013), 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), 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 forecasted to be able to make a successful high-significance power spectrum detection with an eventual 350 elements using current analysis techniques (Pober et al. 2014; Liu & 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 (Santos et al. 2005; Ali et al. 2008; de Oliveira-Costa et al. 2008; Jelić et al. 2008; Bernardi et al. 2009; Bernardi et al. 2010; Ghosh et al. 2011). 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 themes that are essential to investigate for a robust 21 cm power spectrum analysis. We also highlight four power spectrum techniques and their tradeoffs, 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 21 cm power spectrum from the lessons learned.

This paper is organized into two main sections. 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 Ali et al. (2015), henceforth known as P64, which have led to a revised PAPER-64 power spectrum result (Kolopanis et al., *in prep*). We conclude in Section 4.

2. POWER SPECTRUM THEMES AND TECHNIQUES

There are many choices a 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 the source of a detection be properly identified? Many common techniques, such as averaging data, weighting, bootstrapping, and jackknife testing, address these issues but harbor additional tradeoffs. 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 maximize sensitivity but fail to suppress foregrounds.

Despite there being 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 suppress 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 check-list 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 not corrected for, 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 the 21 cm power spectrum result determine the difference between a detection and a null result, which have two very different implications.

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 positive 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 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 increases the sensitivity of a dataset and reduces the number of independent samples by an amount dependent on the width of the averaging window. However, it can also affect the presence of foregrounds and systematics. We explain the trade-offs of filtering in more detail in Section 2.1.

- **Weighting:** A dataset can be weighted to emphasize certain features and minimize others. One particular flavor of weighting employed by PAPER is inverse covariance weighting in frequency, which is a generalized version of inverse variance weighting that also takes into account frequency correlations. This weighting has the effect of down-weighting correlated information (i.e. foregrounds) and up-weighting noise-like information (i.e. EoR). However, a challenge of inverse covariance weighting is in accurately defining a covariance matrix that best describes the data.

- **Bootstrapping:** In addition to using theoretical models for covariance matrices and theoretical error estimation methods, bootstrapping is one way to estimate errors. Bootstrapping is a useful method for estimating errors of a dataset from itself. By randomly drawing many samples 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.

- **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 (Keating et al. 2016). An EoR detection must pass jack-knife and null tests.

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 fitting a polynomial during spectral calibration, applying a delay-domain filter, or by weighting data by 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 P64 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 optimal quadratic estimators (OQE) as done in Liu & Tegmark (2011), Dillon et al. (2013), Liu et al. (2014a), Liu et al. (2014b), and Trott et al. (2012). Before showing how signal loss can arise when using different weighting matrices, we first summarize OQE as performed in the PAPER analysis.

We begin with our data matrix, $\mathbf{x}$, which contains our measured visibilities for a single baseline in Jy. It has length (N_t, N_f), where N_t is the number of time integrations and N_f is the number of frequency channels. Visibilities are measurements of the Fourier transform of the sky along 2 spatial dimensions (using the flat-sky approximation), and since we are interested in 3-dimensional Fourier-modes we only need to take one Fourier transform of our data along the line-of-sight dimension. We do this when forming the un- normalized power spectrum estimate $\hat{q}_\alpha$:

$$\hat{q}_\alpha = \frac{1}{2}\mathbf{x}^\dagger \mathbf{R}\mathbf{Q}^\alpha \mathbf{R}\mathbf{x}. \quad (1)$$

$\mathbf{Q}$ is a family of matrices that takes our frequency-domain visibilities and Fourier-transforms them into power spectrum space, while also taking into account physical constants such as the Boltzmann constant used to convert Jy to Kelvin, as well as cosmological scalings. The index α denotes a waveband in $k_\parallel$, where $k_\parallel$ is the

Fourier-dual to frequency under the delay approximation (Parsons et al. 2012b). $\mathbf{R}$ is a weighting matrix — as an example, inverse covariance weighting would set $\mathbf{R} = \mathbf{C}^{-1}$ and an unweighted case would use $\mathbf{R} = \mathbf{I}$, the identity matrix.

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

$$\hat{\mathbf{p}} = \mathbf{M}\hat{\mathbf{q}}, \tag{2}$$

where $\hat{\mathbf{p}}$ is the normalized estimate of the true power spectrum. The data analyst has a choice for $\mathbf{M}$ — for simplicity in this section we set $\mathbf{M} = \mathbf{I}$, although we explore other cases for the analysis of PAPER-64 data as explained in Section 3.3.

In the next three sections, we investigate the effects of weighting matrices on signal loss by experimenting with different matrices $\mathbf{R}$ and examine their impact on the resulting power spectra estimate $\hat{\mathbf{p}}$. Our ultimate 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, for the sake of completeness. The specifics of how signal loss crops up in PAPER's analysis is later described in Section 3.1.

2.1.1. *Toy Model: Inverse Covariance Weighting*

To build our intuition into how a particular choice of weighting, namely inverse covariance weighting, can lead to signal loss, we use a toy model. We construct a simple dataset that contains 2-dimensional data (representing visibility data with 100 time integrations and 20 frequency channels). This model represents realistic dimensions of about an hour of PAPER data which might be used for a power spectrum analysis. For PAPER-64 (both the previous analysis and our new analysis)we use ∼eight hours of data for our analysis (with channel widths of 0.5 MHz and integration times of 43 seconds), but here we scale it down for this toy model 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) oscillate in time and frequency in our measurements. We therefore form a mock bright foreground signal, $\mathbf{x}_{FG}$, 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 EoR signal, $\mathbf{x}_{EoR}$, as a complex, Gaussian random signal. A more realistic EoR signal would have a sloped power spectrum in $\mathbf{p}(k)$ (instead of flat, as in the case of white noise), which could be simulated by introducing frequency correlations into the mock EoR signal. However, in this paper we treat all k's separately, so a simplistic white noise approximation can be used. Our combined data vector is then $\mathbf{x} = \mathbf{x}_{FG} + \mathbf{x}_{EoR}$, to which we apply different weighting schemes. The three data vectors are shown in Figure 1.

One choice for the weighting matrix $\mathbf{R}$ is an inverse covariance matrix. This type of weighting is attractive for power spectrum analyses because it is an aggressive way to down-weight foregrounds and yields the smallest possible error bars on a measurement (Tegmark 1997; Bond et al. 1998). The covariance matrix $\mathbf{C}$, which in our case describes covariances between frequency channels, can be estimated in a variety of ways. Given perfect foreground, instrumental, and EoR models, we could form $\mathbf{C}$ in a way that accurately describes our measured data. However, if our foreground model is flawed, for example, our estimate of $\mathbf{C}$ would not be successful at down-weighting them in the data.

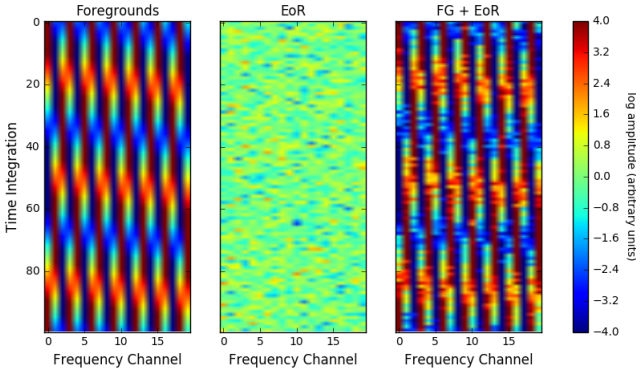

Figure 1. Our toy model dataset. We model a mock foreground-only visibility with a sinusoid signal that varies smoothly in time and frequency. We model a mock EoR signal as a random Gaussian signal. Real parts are shown here.

Therefore, one attractive way to estimate $\mathbf{C}$ is to empirically derive it from the data vector $\mathbf{x}$ itself:

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

assuming $\langle \mathbf{x} \rangle = 0$ (a reasonable assumption since fringes average to 0 over a sufficient amount of time), where $\langle\rangle$ denotes a finite average over time t. The inverse covariance matrix is therefore $\mathbf{R} = \hat{\mathbf{C}}^{-1}$.

First, we compute the power spectrum of $\mathbf{x}$ using OQE formalism and a weighting matrix of $\hat{\mathbf{C}}^{-1}$. The result is shown in green in the left plot of Figure 2 (the right plot shows an alternate weighting approach and will be discussed further below). Also plotted in the figure are the unweighted ($\mathbf{R} = \mathbf{I}$) power spectrum of $\mathbf{x}_{FG}$ (blue) and $\mathbf{x}_{EoR}$ (red).

As shown, our inverse covariance-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 order to understand the behavior of this result, we can closely study our covariance matrix, $\hat{\mathbf{C}}$. shown in Figure 4.

If $\mathbf{C}$ is computed from the data itself, it carries the risk of over-fitting information in the data and introducing a multiplicative bias (per k) to estimates of the signal. For a mathematical derivation of signal loss arising from a data- estimated covariance matrix, see Ap-

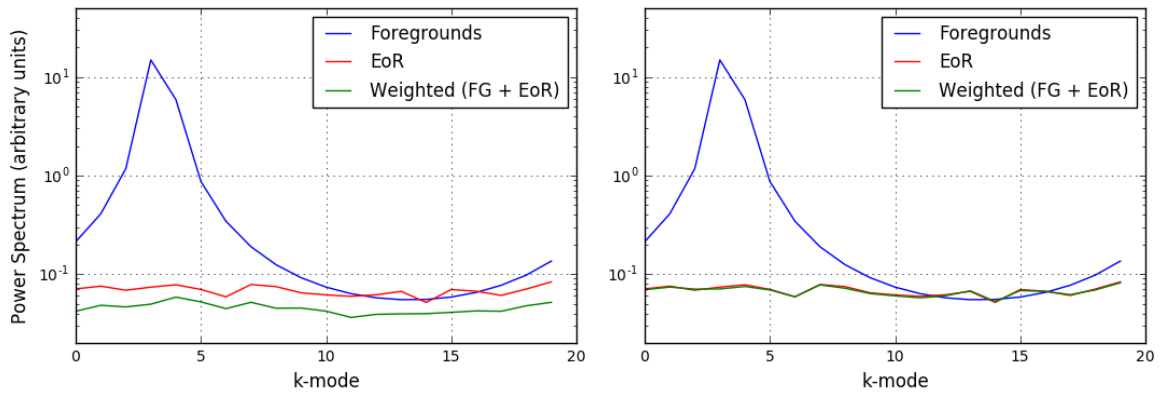

Figure 2. Resulting power spectrum estimates for the toy model simulation described in Section 2.1.1 — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). The foregrounds peak at a k mode based on the frequency of the sinusoid used to create it. We compare inverse covariance weighting where $\mathbf{C}$ is derived from the data (left), with projecting out the first eigenmode only (right). In the former case, signal loss arises from using information from all eigenmodes of $\hat{\mathbf{C}}$. Because $\hat{\mathbf{C}}$ is empirically estimated, its eigenvalues pose the risk of describing random fluctuations that happen to exist in the data realization but may not exist in the true covariance. Consequently, these modes are over-fitted and down-weighted, leading to signal loss. There is no signal loss when using only the zeroth eigenmode, since we are not using information from the weaker eigenmodes which $\hat{\mathbf{C}}$ does not describe accurately.

pendix A. Here we will describe the origin of this signal loss intuitively.

It turns out that because we estimated $\hat{\mathbf{C}}$ from our data, its eigenspectrum differs from the eigenspectrum of our true $\mathbf{C}$, and this difference has 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. This is usually not the case for a covariance matrix, for which a sloped eigenspectrum means that modes are given different weights. The modes with the highest eigenvalues are down-weighted the most.

If the true covariance matrix $\mathbf{C}$ of our data was known, then every single eigenvalue and eigenvector of $\mathbf{C}$ would be representative of real fluctuations in the data. However, when using an estimated $\hat{\mathbf{C}}$ that is derived from one particular data realization, its eigenvalues and eigevectors may differ from the truth. Said differently, shapes that may not exist (or have a weaker existence) in a true covariance may appear stronger in the estimated covariance. Hence, they will be down-weighted more than they should be.

In general, the strongest modes of $\hat{\mathbf{C}}$ (highest eigenvalues) are expected to be dominated by bright foregrounds — the most prominent shapes in the data. For these 'strong' modes, where foregrounds outshine the EoR signal, $\hat{\mathbf{C}} \sim \mathbf{C}$ and down-weighting these modes is beneficial. This is demonstrated in the toy model by the successful suppression of the foreground mode, where our estimated covariance matrix identifies the sinusoid shape and assigns it the highest eigenvalue (the peak in Figure 3). In Figure 4 we show the covariances

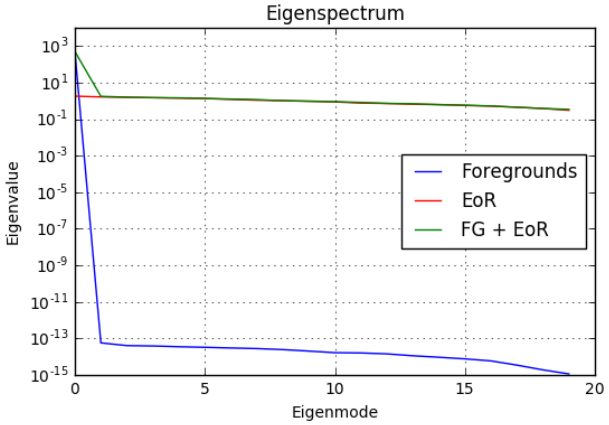

Figure 3. Eigenspectrum of $\hat{\mathbf{C}}_{FG}$ (blue), $\hat{\mathbf{C}}_{EoR}$ (red), and $\hat{\mathbf{C}}_{FG+EoR}$ (green). The eigenspectrum of $\hat{\mathbf{C}}_{FG}$ peaks at the zeroth eigenmode, due to the presence of only one sinusoid. These empirically estimated covariance matrices have eigenspectra that are different from that of a true $\mathbf{C}$, meaning that some eigenmodes (shapes in the data) may be down-weighted more significantly than they should be, producing signal loss. Specifically, down-weighting the weak eigenmodes, where the EoR signal is greater than the FG signal, leads to loss.

of our toy model datasets along with inverse covariance weighted data. The foreground sinusoid is clearly visible in $\hat{\mathbf{C}}_{FG}$ but effectively absent from $\hat{\mathbf{C}}_{FG}^{-1}\mathbf{x}_{FG}$.

The danger of an empirically estimated covariance matrix comes from not being able to describe the 'weak' eigenmodes of $\mathbf{C}$ accurately, for which the EoR signal is brighter than foregrounds. If these weak eigenmodes of $\hat{\mathbf{C}}$ characterize EoR modes, the EoR signal will be down-weighted (not as much as the strong foreground modes, but still down- weighted). This leads to the 'overfitting' of noise (or 'overfitting' of signal), which is signal loss.

Using what we've learned about the eigenspectrum, we can tweak it in a simple way to suppress foregrounds

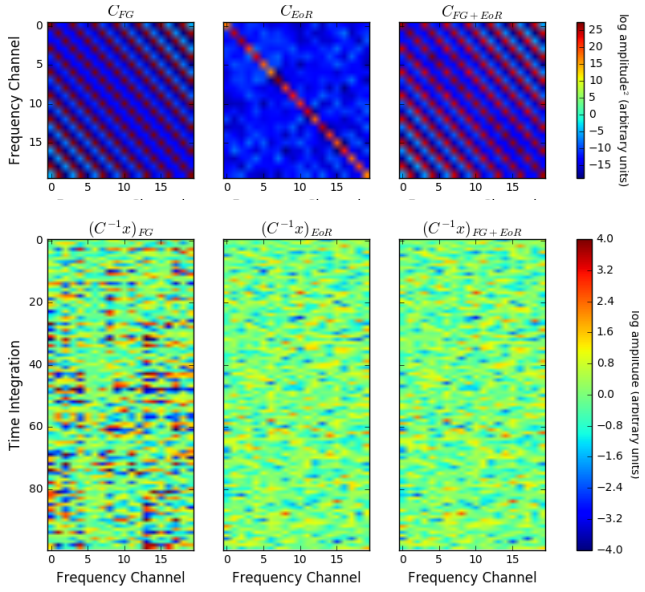

Figure 4. The covariance matrices (top row) and inverse covariance-weighted data (bottom row) for FG only (left), EoR only (middle), and FG + EoR (right). Real parts are shown here.

and yield zero signal loss. Recall that our toy model foreground is a sinusoid that can be perfectly described by a single eigenmode. We can project out the zeroth eigenmode (i.e. zero out all eigenmodes except the first one), thereby down-weighting the foreground-dominating mode only. Hence, we are not using information from all the weaker eigenmodes which carry the risk of describing EoR fluctuations. Altering $\hat{\mathbf{C}}$ as such is one example of a regularization method, in which we are changing $\hat{\mathbf{C}}$ in a way that flattens its eigenspectrum, making it more similar to that of $\mathbf{I}$. The resulting power spectrum estimate for this case is shown in the right plot of Figure 2. In this case we recover EoR, demonstrating that if we can disentangle the foreground-dominating modes from EoR-dominating modes, we can down-weight them without signal loss. There are several other ways to regularize $\hat{\mathbf{C}}$, 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 inaccurately characterizing weak eigenmodes (EoR dominated modes) with a data-estimated covariance. We will next show how this effect is exaggerated 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, a time domain filter can be designed to up-weight temporal modes corresponding to 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, left). We choose these numbers so that the total number of independent samples is similar to the number of frequency channels — therefore, our matrices will be full rank.

The resulting eigenspectrum as compared to the green curve (FG + EoR) in Figure 3 is shown in Figure 5 (right plot). The spectrum, in the case of fringe-rate filtering (dashed line), falls more steeply, especially for the last few eigenmodes. This is because we have fewer independent modes and therefore less information to characterize the eigenspectrum. Because we do a worse job characterizing, we have more signal loss. **[CC: need help filling out the rest here... and also citing galaxy lit]**.

The power spectrum results for inverse covariance weighted fringe-rate filtered data is shown in Figure 6. As expected, there is a much larger amount of signal loss for this time-averaged dataset. Additionally, because of the steepened eigenspectrum, the last few eigenmodes receive more weight relative to the others. Therefore, most of our power spectrum information is coming from only these last few modes and as a result, it is a noisier estimate. This is evident by noticing that the green curve in Figure 6 fails to trace the shape of the unweighted 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*

In Section 2.1.1 we showed one example (projecting the zeroth eigenmode) of how altering $\hat{\mathbf{C}}$ can make the difference between zero and some signal loss, if we can distinguish between real eigenmodes in a true covariance matrix from those in an estimated one. We will now use our toy model to describe several other ways to tailor $\hat{\mathbf{C}}$ 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 7 and 8.

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 $\mathbf{C}_{EoR}$ as such, instead of computing it based on $\mathbf{x}_{EoR}$ itself. Next, we add $\mathbf{C}_{EoR} + \hat{\mathbf{C}}_{FG}$ (where $\hat{\mathbf{C}}_{FG} = \langle \mathbf{x}_{FG}\mathbf{x}_{FG}^{\dagger} \rangle$) to obtain a final $\hat{\mathbf{C}}$ to use in

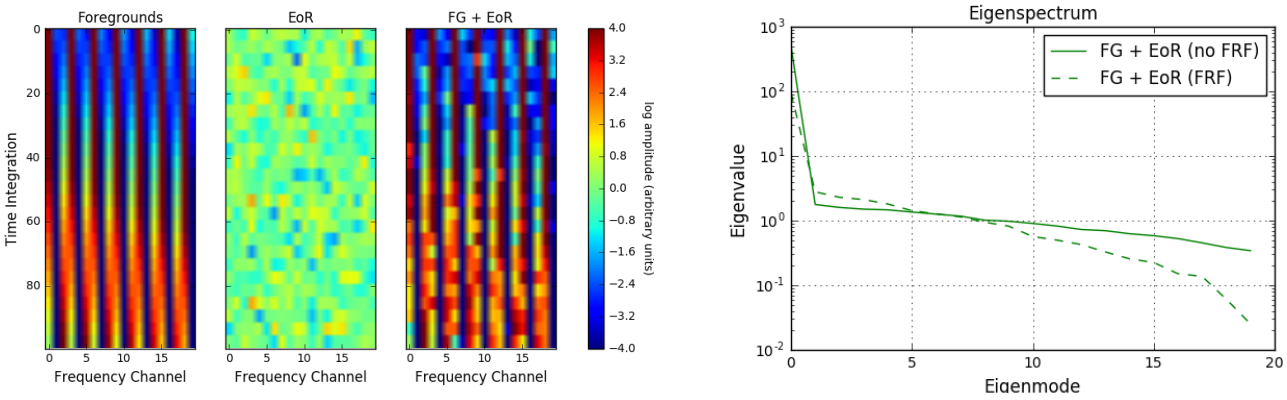

Figure 5. Left: Our 'fringe-rate filtered' (time-averaged) toy model dataset. We average every four samples together, yielding 25 independent samples in time. Real parts are shown here. Right: Eigenspectrum of $\hat{\mathbf{C}}_{FG+EoR}$, in the case of no fringe-rate filtering (solid green) and with fringe-rate filtering (dashed green). The dashed spectrum is steeper, resulting in more dramatic differences in weighting between eigenmodes. Consequently, if all eigenmodes are used in weighting it is possible to down-weight EoR modes more severely than for the non-averaged case.

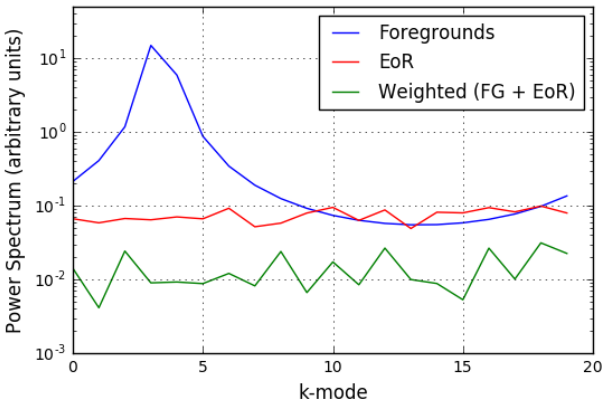

Figure 6. Resulting power spectrum estimate for the fringe-rate filtered toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). We use inverse covariance weighting where $\mathbf{C}$ is computed from the data. There is a larger amount of signal loss than for the non-averaged data, a consequence of a steepened, weakly characterized eigenspectrum from using fewer independent modes in the data. The power spectrum estimate is also noisier, since most of the information used in constructing the estimate came from only a few eigenmodes (the weakest, EoR-dominated modes, which are down-weighted the least).

weighting. In Figure 7 (upper left), we see that there is negligible signal loss. This is because by modeling $\mathbf{C}_{EoR}$, we avoid over-fitting fluctuations in the data that our model doesn't know about (but an empirically derived $\hat{\mathbf{C}}$ would). This is evident when comparing the green and red curves in Figure 8. We don't see perfect EoR recovery though, as the green and red curves vary slightly in Figure 7. This deviation is a consequence of the difference between the true and modeled covariance matrices. In practice such a weighting option is not feasible, as we do not know $\mathbf{C}_{EoR}$.

The second panel (top right) in Figure 7 uses a regularization method of setting $\hat{\mathbf{C}} = \hat{\mathbf{C}} + \gamma\mathbf{I}$, where i=j=20 (number of frequencies) and $\gamma = 5$ (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 matrix more heavily than those off-diagonal, thereby flattening out its eigenspectrum. If all modes are given similar weights, we avoid down-weighting EoR modes more than others.

The third panel (bottom left) in Figure 7 flattens out the eigenspectrum of $\hat{\mathbf{C}}$ a different way - by only using the first three eigenmodes. Recalling that the foregrounds can be described entirely by the first eigenmode, this method intentionally projects out modes that are EoR-dominated by replacing all but the three highest weights in the eigenspectrum with 1's (equal weights). Again, flattening the eigenspectrum results in negligible signal loss. However, we do not perfectly recover the shape of EoR because we lost information when projecting out certain modes.

The last regularization scheme we are highlighting here is setting $\hat{\mathbf{C}} = \hat{\mathbf{C}} \circ \mathbf{I}$ (element- wise multiplication), or inverse variance weighting (keeping only the diagonal elements of $\hat{\mathbf{C}}$). In the bottom right panel of Figure 7, we see that this method does a poor job down-weighting foregrounds. 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 are only left with correlation information between the same two frequencies. Because we disregard information from all other frequency combinations in this case, we do a poor job suppressing the foreground. But because we flattened the eigenspectrum, we also avoid signal loss.

Although the fourth method did not successfully recover EoR for this particular simulation, it is important that we show that there are many options for estimating a covariance matrix, and some may be more effective than others based on the spectral nature of the components in a dataset. One may imagine a situation where a particular systematic is contained to an isolated frequency (such as radio frequency interference or

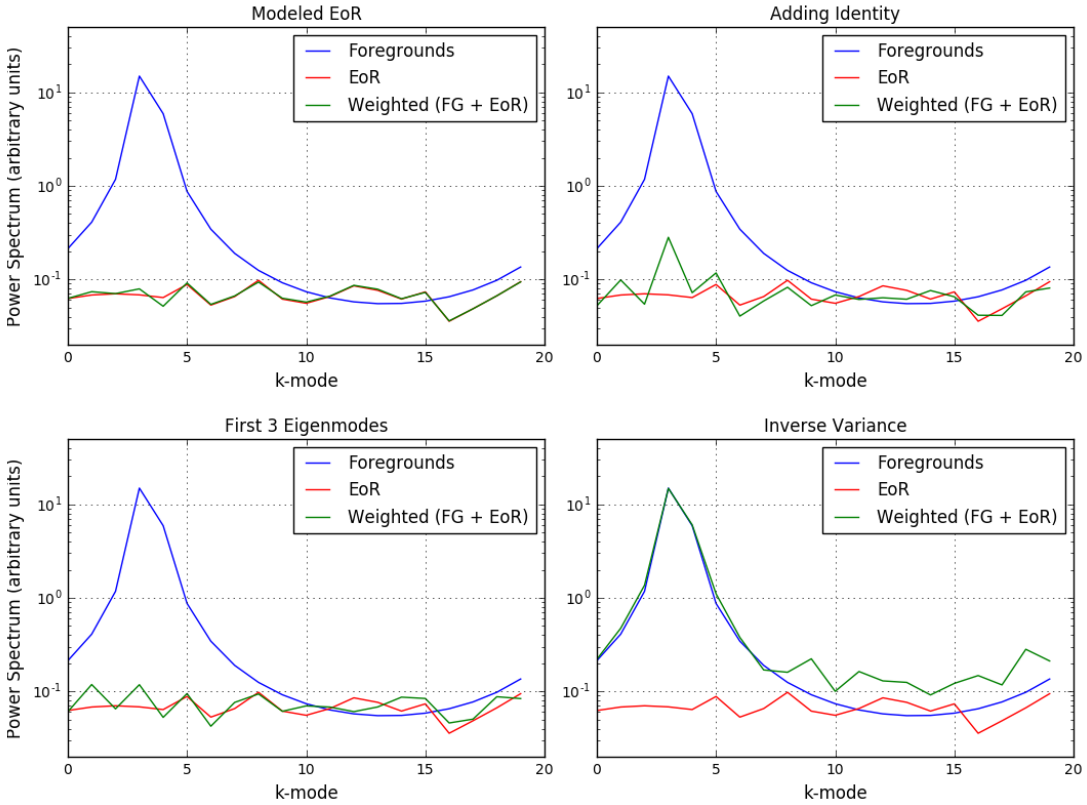

Figure 7. Resulting power spectra estimates for our fringe-rate filtered toy model simulation — foregrounds only (blue), EoR only (red), and the weighted FG + EoR dataset (green). We show four alternate weighting options that each avoid signal loss, including modeling the covariance matrix of EoR (upper left), regularizing $\hat{\mathbf{C}}$ by adding an identity matrix to it (upper right), using only the first three eigenmodes of $\hat{\mathbf{C}}$ (lower left), and keeping only the diagonal elements of $\hat{\mathbf{C}}$ (lower right).

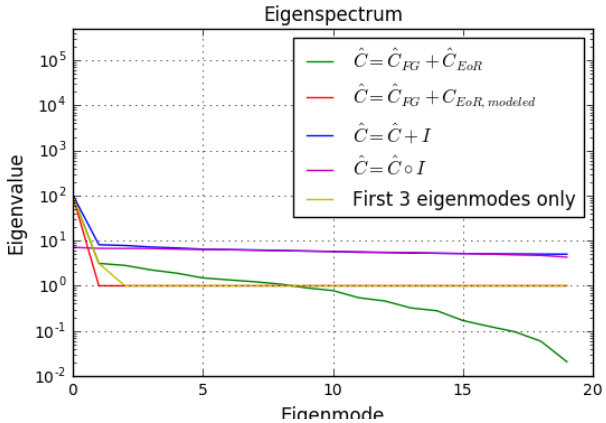

Figure 8. We compare the eigenspectrum of an empirically calculated $\hat{\mathbf{C}}$ (green) to that of four alternate weighting options, including modeling the covariance matrix of EoR (red), regularizing $\hat{\mathbf{C}}$ by adding an identity matrix to it (blue), using only the first three eigenmodes of $\hat{\mathbf{C}}$ (yellow), and multiplying an identity matrix with $\hat{\mathbf{C}}$ (magenta). All eigenspectra (except the green) are relatively flat and don't exhibit signal loss. All were computed for the fringe-rate filtered toy model case.

crosstalk). In such a case, preserving only the diagonal elements of $\hat{\mathbf{C}}$ would be an effective way of removing this contamination.

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. As we investigated however, there are tradeoffs between the weighting method used, its foreground-removal effectiveness, 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 empirically calculating errors on a power spectrum measurement are calculating the variance of a dataset, 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 don't always agree due to a number of factors, including time, frequency, and antenna-dependent noise and non-uniform weightings.

For PAPER's analysis, we choose to compute error bars on our measurements that have been derived from its inherent variance. A common technique used to do this is bootstrapping. For pedagogical purposes, we first define the technique of bootstrapping and then illustrate

some of its pitfalls through toy models.

Bootstrapping uses sampling with replacement to estimate a posterior distribution. For example, measurements (like power spectra) can be made from different samples of data. Averaging each sample set gives us a realization of the (power spectrum) answer. Realizations are correlated with each other to a degree set by the fraction of sampled points that are held in common between them. Hence, through the process of resampling and averaging along different axes, we can estimate error bars for our results which represent the underlying distribution of values that are allowed by our measurements.

Suppose we have N different measurements targeting the same quantity (N power spectrum measurements). Bootstrapping means that we form N_{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 (every Nth slot 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_{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_{boot} is typically large, and the standard deviation across bootstraps (the error we are computing) converges for large N_{boot}. It is the value of N that we must be careful in setting, as described in the rest of this section.

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 ($N_{boot} = 500$). 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 9 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. 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 9, black points), and we can both under-estimate the error by drawing too many or over-estimate it by drawing too few. However, if we know that we have 100 independent samples, the error associated with drawing 100 samples with replacement does match the theoretical prediction as expected.

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.

We will now discuss a second subtle feature of bootstrapping that fails to maximize the sensitivity of a dataset. Suppose we have 5 independent measurements of the sky (from 5 different baselines, for example). A bootstrap then consists of 5 measurements that are drawn randomly with replacement from the original set. If all 5 spaces are filled randomly, there is a high probability that some measurements will be repeated in the bootstrap because they are drawn more than once. The bootstrap may therefore consist of only 3 or 4 independent measurements — a number smaller than the total number of samples. In fact, the probability of drawing 5 completely independent measurements is less than 4%.

In order to maximize sensitivity, we desire as many independent samples as possible. However, solely using all 5 independent measurements does not allow variation between bootstraps.

Therefore, we use a slightly modified bootstrapping

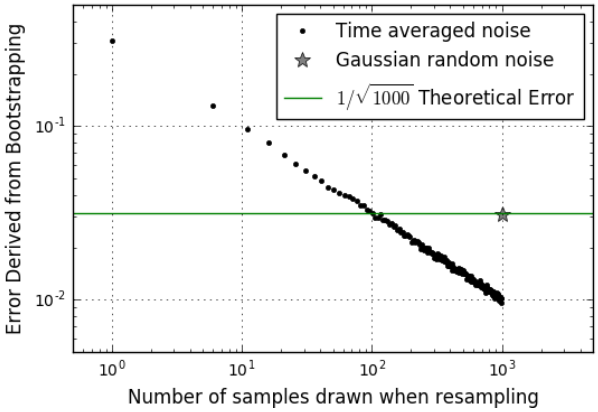

Figure 9. Error estimation produced by bootstrapping as a function of the number of elements drawn per bootstrap when sampling with replacement. The star represents the standard deviation of 500 bootstraps, each created by drawing 1000 elements (with replacement) from a length 1000 array of a Gaussian random signal. The black points correspond to time- averaged data (correlated data) which has 100 independent samples. They illustrate how errors can be underestimated if drawing more elements than there are independent samples in the data. The estimated errors match up with the theoretical prediction only at $N = 100$.

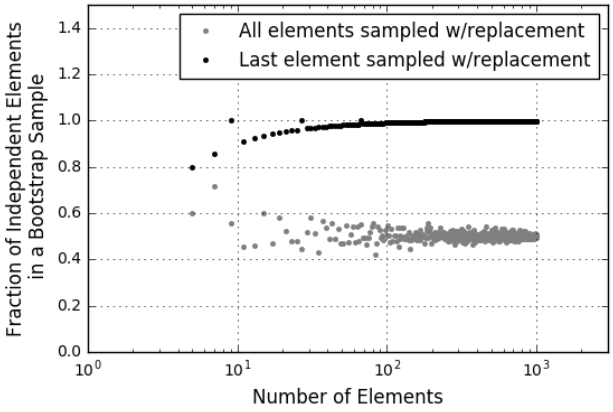

Figure 10. The fraction of independent elements in one bootstrap sample as a function of the total number of elements in a dataset (i.e. if 4 out of 5 elements are independent, this would yield a fraction of 0.8). A fraction of 1.0 means that all elements in the bootstrap sample are independent, and therefore sensitivity is maximized (alternately, this axis can be thought of as power spectrum sensitivity, where sensitivity increases moving upwards along the y-axis). Two bootstrapping methods are shown here — sampling all elements with replacement (gray) and sampling only the last element with replacement (black). The former fails to maximize sensitivity since some elements end up being repeated due to random sampling.

method. For each bootstrap, we draw $N - 1 = 4$ samples without replacement and draw the last slot randomly with replacement. In doing so there is a small chance a bootstrap will consist of 5 independent measurements, but even with 4 our sensitivity is nearly maximized. change is still a legitimate way of error estimating since random sampling only occurs for one value in a bootstrap. However, as long as the number of possible variations is greater than the number of bootstraps we perform, it is a valid way to uncover the inherent variability in a dataset.

In Figure 10 we compare the two methods of bootstrapping: sampling all elements randomly (black points) versus sampling just the final element randomly (gray points). The two converge for datasets with small numbers of elements, when filling a few spots randomly is nearly the same as filling only the last spot randomly. However, there is also increased scatter in this regime due to small number statistics.

The difference between the two methods is most pronounced for datasets with large numbers of elements. As a dataset grows in size, it becomes increasingly rare to draw entirely independent samples, and thus the benefit of ensuring, say, 999 independent samples out of 1000 is most noticeable. The more elements in our dataset, the more sensitivity we can gain by mandating that most of our measurements are independent ones. An analog for the y-axis in Figure 10 is therefore the sensitivity of a final power spectrum measurement.

In summary, bootstrapping can be an effective and straightforward way to estimate errors of a dataset. However, we have illustrated two situations in which bootstrapping can lead to under-estimated errors and sensitivities. First, 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. We have also shown a situation in which a bootstrapping method does not provide the most sensitive measurement possible. While bootstrapping is convenient because it provides a way to estimate errors from the data itself, one must assess whether certain analysis choices have compromised the method and whether a variation of traditional re-sampling could be preferred instead.

2.3. *Bias*

In a 21 cm power spectrum, detections could be the EoR signal, but they could also (and unfortunately more likely) 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. 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*

Foreground bias is perhaps one of the main factors limiting 21 cm results. 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; Parsons et al. 2012a; Parsons et al. 2012b; Dillon et al. 2013; Wang et al. 2013; Parsons et al. 2014; Liu et al. 2014a; Liu

et al. 2014b; Dillon et al. 2015; Pober et al. 2016; Trott et al. 2016), most experiments remain ultimately limited by residuals rather than noise.

Because foreground spectra are smooth in frequency, they preferentially show up at low delay, or k modes. For a particular baseline length, there is a maximum delay imposed on foregrounds, 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" (Parsons et al. 2012b; Liu et al. 2014a; Liu et al. 2014b; Vedantham et al. 2012; Thyagarajan et al. 2013; Pober et al. 2013; Datta et al. 2010). The wedge describes a region in k- space contaminated by 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 remove foregrounds, as done by P64, Parsons et al. (2014), and Jacobs et al. (2015).

However, this isolation is not perfect. In deep measurements, power spectrum measurements at $k_\parallel$ values beyond the length of a baseline are often still contaminated at a low level. This leakage, especially 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 and the antenna response, both of which could smear out the foregrounds and cause leakage outside the wedge.

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 can be used to minimize the covariance of a particular k value with other ones (Liu et al. 2014b). In other words, one can construct an estimator using OQE that forces a window function to have a minimum response to low k values. The window function used in P64 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). Minimizing foreground leakage in this way however, comes with the trade-off of compromising power spectrum sensitivity, since narrow window functions increases errors for each k-mode (Liu et al. 2014b).

Confirming foreground detections at higher k's is more difficult. In the next section (Section 2.3.2), 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 occurs 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.

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 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. 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. Several null tests on data are described in Keating et al. (2016). 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 EoR. 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 more broad 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 $\mathbf{fg}$ is the foreground signal, the two measurements (which are cross- multiplied to form power spectra) can be written as:

$$\mathbf{x}_a = \mathbf{n}_a + \mathbf{e} + \mathbf{fg} \tag{4}$$

and

$$\mathbf{x}_b = \mathbf{n}_b + \mathbf{e} + \mathbf{fg}. \tag{5}$$

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:

$$\mathbf{x}_1 = \mathbf{n} + \mathbf{e} + \mathbf{fg} \tag{6}$$
$$\mathbf{x}_2 = \mathbf{n} + \mathbf{e} \tag{7}$$

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 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 3 different power spectrum estimates shown in Figure 11. The first is a null test where we 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 un-biased (thermal noise-wise) estimate (blue curve). The second estimate, shown in red, is the same null test with the foreground systematic removed (eliminate $\mathbf{fg}$ in Equations 4 and 5). Finally, we also show the noise power spectrum (green).

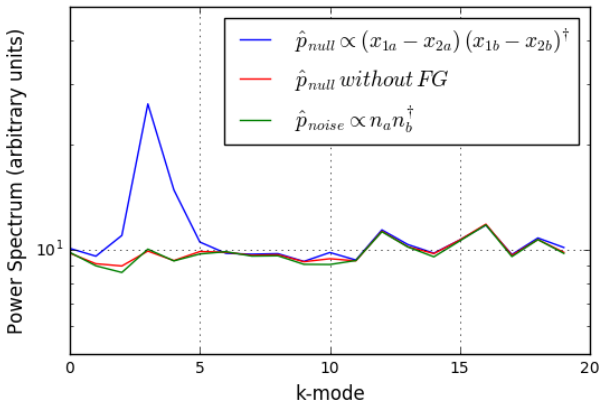

Figure 11. Power spectrum estimates for a null jackknife test with the presence of a foreground systematic (blue), without the foreground systematic (red), and noise alone (green). Because the first null test is not consistent with noise, it suggests the presence of a systematic in either $\mathbf{x}_1$ or $\mathbf{x}_2$. Null tests of clean measurements should be consistent with thermal noise.

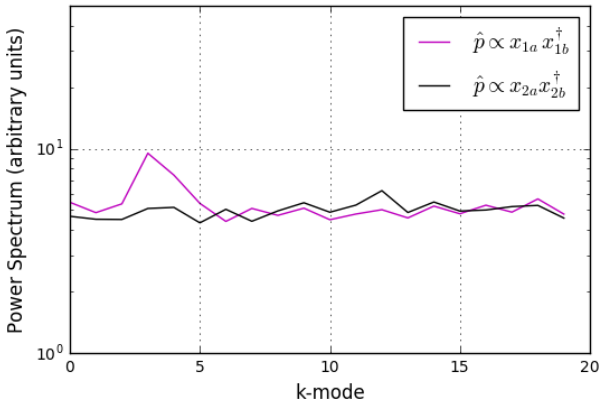

Figure 12. Power spectrum estimates for $\mathbf{x}_1$ and $\mathbf{x}_2$, two jackknives of the toy model. They suggest the presence of a systematic in $\mathbf{x}_1$ only, illustrating how jackknives can be used to tease out excesses. Clean measurements should remain consistent despite the jackknife taken.

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 12). 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 cross-talk noise bias is successfully suppressed with the filter, or if similar-shaped detections in both

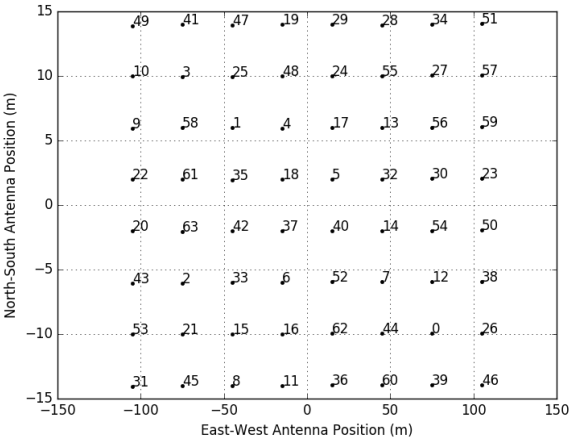

Figure 13. The PAPER-64 antenna layout. We use only the 30 m East/West baselines for the revised analysis in this paper (such as baseline 49_41).

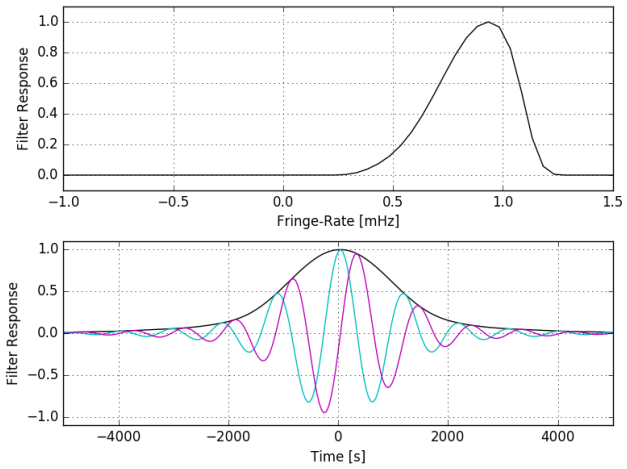

Figure 14. Top: the normalized optimal power-spectrum sensitivity weighting in fringe-rate space for our fiducial baseline and Stokes I polarization beam. Bottom: the time-domain convolution kernel corresponding to the top panel. Real and imaginary components are illustrated in cyan and magenta, respectively, 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.

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 present a case study of these same three themes using data from the PAPER experiment. We use the intuition developed through the toy model simulations in order to make a new analysis of the PAPER-64 dataset originally presented in P64 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 13). For information about the backend system of PAPER-64, its observations, and data reduction pipeline, we refer the reader to Parsons et al. (2010) and P64.

The previously best published 21 cm upper limit result from P64 uses 124 nights of data to place a 2σ upper limit on $\Delta^2(k)$, defined as

$$\mathbf{\Delta^2(k)} = \frac{k^3}{2\pi^2}\hat{\mathbf{p}}(\mathbf{k}), \tag{8}$$

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

For our analysis, we use 8 hours of LST (RA 0.5-8.6 hours) and 51 total baselines (P64 uses a slightly different RA range of 0-8.6 hours). 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 P64. We note that, besides using only 1 baseline type instead of the 3 in P64, the PAPER-64 dataset that we use in this case study differs from that in P64 mainly by the fringe-rate filter. In P64, the applied 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 see 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 14.

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 that simulates the injection and recovery of a cosmological signal in order

to quantify the amount of signal loss accompanying a weighting scheme. In particular, we highlight major differences from the signal loss computation used in P64, which previously underestimated losses.

3.1.1. *Signal Loss Methodology*

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 used in P64, where mock Gaussian noise visibilities (mock EoR signals) are filtered in fringe-rate space to retain only "sky-like" time-modes. This signal is then added to the data. [1]

Suppose that $\mathbf{e}$ is the injected EoR (at some amplitude level), and $\mathbf{x}$ is our data vector. We define $\mathbf{r}$ to be the data plus the EoR signal:

$$\mathbf{r} = \mathbf{x} + \mathbf{e}. \tag{9}$$

We are interested in quantifying how much variance in $\mathbf{e}$ is lost after weighting $\mathbf{r}$ and estimating the power spectrum according to OQE 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 (ignoring normalization factors):

$$P_{{\rm in},\alpha} = \mathbf{e}^\dagger \mathbf{I}\mathbf{Q}^\alpha \mathbf{I}\mathbf{e} \tag{10}$$

and

$$\begin{aligned} P_{{\rm out},\alpha} &= \hat{\mathbf{p}}_{r,\alpha} - \hat{\mathbf{p}}_{x,\alpha} \\ &= \mathbf{r}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{r} - \mathbf{x}^\dagger \mathbf{R}_x \mathbf{Q}^\alpha \mathbf{R}_x \mathbf{x}. \end{aligned} \tag{11}$$

$P_{\rm in}$ represents the unweighted power spectrum of $\mathbf{e}$, our simulated EoR signal, $\hat{\mathbf{p}}_x$ is the power spectrum of the data alone, $\hat{\mathbf{p}}_r$ is the power spectrum of the data plus injection, and $P_{\rm out}$ is the difference between $\hat{\mathbf{p}}_r$ and $\hat{\mathbf{p}}_x$.

In short, we can summarize a signal loss correction factor as the ratio of $P_{\rm out}/P_{\rm in}$, evaluated at the data level $\mathbf{x}$ (our detailed computation is described in the next section). We motivate the fact that we can evaluate the output-to-input power spectrum ratio at $\mathbf{x}$ by the following reasoning.

In a limit with no instrumental noise, the data that we measure, $\mathbf{x}$, is comprised of two signals, such that

$$\mathbf{x} \equiv \mathbf{f} + \mathbf{s}, \tag{12}$$

where $\mathbf{f}$ represents the foregrounds and $\mathbf{s}$ represents the cosmological signal. In general, suppose that our power spectrum algorithm passes the data through some function g, yielding a lossy estimate of the power spectrum $\hat{\mathbf{p}}$. This can be parametrized as

$$\langle \hat{\mathbf{p}} \rangle = \langle g(\mathbf{x}) \rangle = c_{\rm fg}\mathbf{p}_{\rm fg} + c_{\rm eor}\mathbf{p}_{\rm eor}, \tag{13}$$

where $c_{\rm eor}$ and $c_{\rm fg}$ are multiplicative factors accounting for the signal loss in the true EoR power spectrum $\mathbf{p}_{\rm eor}$ and true foreground power spectrum $\mathbf{p}_{\rm fg}$, respectively. It is not *a priori* obvious why this parameterization is suitable; we thus provide a toy model in Appendix A to motivate this, although it should be noted that the result is an approximation.

Given this form, a suitable estimate for $\mathbf{p}_{\rm eor}$ would be $\hat{\mathbf{p}}_x/c_{\rm eor}$, or the uncorrected power spectrum of data divided by the signal loss correction factor. Although such an estimate leaves an additive bias from foregrounds that must be mitigated by other methods (as is the case for any attempt to measure the 21 cm power spectrum), it normalizes $\mathbf{p}_{\rm eor}$ back to its correct level such that there is no multiplicative bias. The most conceptually straightforward way to compute $c_{\rm eor}$ is to model the foregrounds and EoR signal via simulations, and then to form the quantity

$$\hat{c}_{\rm eor} = \frac{g(\mathbf{f}+\mathbf{s}) - g(\mathbf{f})}{\mathbf{p}_{\rm eor}}. \tag{14}$$

However, this approach assumes that we have sufficiently good knowledge of our foreground and signal models, which is certainly not the case — if it were, it would be simpler to construct our covariance matrices from our models, avoiding signal loss altogether! Instead, we can use the data itself as our model of the foregrounds, injecting a new EoR signal $\mathbf{e}$ (with power spectrum $\mathbf{p}^*_{\rm eor}$), computing instead

$$\hat{c}_{\rm eor} = \frac{g(\mathbf{x}+\mathbf{e}) - g(\mathbf{x})}{\mathbf{p}^*_{\rm eor}}, \tag{15}$$

which reduces to the same result because of the linearity of Equation (13). Essentially, one is computing the slope of $\langle \hat{\mathbf{p}} \rangle$ with respect to $\mathbf{p}_{\rm eor}$ (note that both Equations (14) and (15) take the form of finite difference derivatives). Under the approximation that the relation between the two quantities is linear, it does not matter whether this slope is evaluated about $\mathbf{x}$ or $\mathbf{f}$. In reality, one expects some deviations from linearity, but Equation (15) remains a good approximation of Equation (14) as long as $\mathbf{x}$ is dominated by $\mathbf{f}$.

Using this motivation, the numerator of Equation (15) can be mapped to our expression for $P_{\rm out}$, and the denominator to $P_{\rm in}$, where the function g is our OQE power spectrum pipeline.

One may wonder why $P_{\rm out}$ cannot be computed simply as the weighted power spectrum of $\mathbf{e}$ alone, namely $P_{{\rm out},\alpha} = \mathbf{e}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{e}$. In short, this naive expression fails to include correlations between the injected EoR signal and data, which emerges as statistical noise that is added to the estimate, decreasing the value of $P_{\rm out}$. Mathematically, we can show this by expanding out Equation (11):

[1] One specific change from P64 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 (by a factor of ∼10), likely due to the use of the fringe-rate filter as a simulator.

$$P_{\text{out},\alpha} = (\mathbf{x}+\mathbf{e})^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r (\mathbf{x}+\mathbf{e}) - \mathbf{x}^\dagger \mathbf{R}_x \mathbf{Q}^\alpha \mathbf{R}_x \mathbf{x}$$
$$= \mathbf{x}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{x} + \mathbf{e}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{e} + \mathbf{x}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{e}$$
$$+ \mathbf{e}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{x} - \mathbf{x}^\dagger \mathbf{R}_x \mathbf{Q}^\alpha \mathbf{R}_x \mathbf{x}. \quad (16)$$

We see that P_{out} is comprised of multiple terms. Taking the case of very large $\mathbf{e}$ so that any terms involving only $\mathbf{x}$ are small, yields:

$$P_{\text{out},\alpha,\mathbf{e}\gg\mathbf{x}} = \mathbf{e}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{e} + \mathbf{x}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{e}$$
$$+ \mathbf{e}^\dagger \mathbf{R}_r \mathbf{Q}^\alpha \mathbf{R}_r \mathbf{x}. \quad (17)$$

We see that our naive expression for P_{out} is the first term in Equation (17), but there are also two additional terms. An initial assumption would be that the cross-terms that involve both $\mathbf{e}$ and $\mathbf{x}$ should be zero, since foregrounds and the cosmological signal are statistically unrelated. In P64, this assumption is used, and the first term in Equation (17) is directly compared to the input (unweighted) power spectrum to compute signal loss.

However, deeper investigations of these terms reveal that they contain non-negligible power. In fact, foreground modes and signal modes are anti-correlated on average, and this negative cross-term power arises from inverting the covariance matrix $\mathbf{C}_r$. Mathematically, the covariance of the data plus EoR is:

$$\mathbf{C}_r = \langle \mathbf{r}\mathbf{r}^\dagger \rangle \quad (18)$$
$$= \langle \mathbf{x}\mathbf{x}^\dagger \rangle + \langle \mathbf{x}\mathbf{e}^\dagger \rangle + \langle \mathbf{e}\mathbf{x}^\dagger \rangle + \langle \mathbf{e}\mathbf{e}^\dagger \rangle, \quad (19)$$

where the data-only term (first term) and EoR-only term (last term) are approximately diagonal matrices. The middle two terms, however, have a larger amount of power in their off-diagonal terms relative to their diagonal elements, since $\mathbf{e}$ and $\mathbf{x}$ are statistically unrelated. Upon inversion, these off-diagonal terms become negative, driving the cross-terms in Equation 17 to have negative power. **[CC: someone help me write this more elegantly...]**This effect is explained in more depth in Switzer et al. (2015).

The spurious correlation between $\mathbf{e}$ and $\mathbf{x}$ means that we cannot compute signal loss using a signal-only simulation, which would yield greater values for P_{out} and thereby underestimate signal loss. Therefore, in our revised signal loss computation we use the full quantity for P_{out} as defined in Equation (11), which subtracts the weighted power spectrum of the data from the weighted power spectrum of data plus EoR.

The relationship between the input and output power spectra, P_{in} and P_{out}, can be thought of as a transfer function which, for a sampling of P_{in} and P_{out} provides a mapping from an input power spectrum distribution into an output distribution. By viewing data through this signal loss lens, we may then ask the question "what input power spectrum distribution could this (signal-loss affected) data come from?" In the next section, we describe the shape of our signal loss transfer function and detail two different computations used to translate our power spectrum result into one viewed through a signal loss lens. We showcase two methods... **[CC: fill in with some broad statement of why we're showing 2 methods but also state that they yield similar results...]**

3.1.2. *Signal Loss in Practice*

METHOD #1

We now shift our attention towards computing signal loss for the fringe-rate filtered PAPER-64 dataset. While our methodology outlined below is independent of weighting scheme, here we demonstrate the computation using inverse covariance weighting ($\mathbf{R} = \mathbf{C}^{-1}$), the weighting scheme used in P64 which leads to substantial loss. With this weighting, our expressions for P_{in} and P_{out} become (ignoring normalization factors):

$$P_{\text{in},\alpha} = \mathbf{e}^\dagger \mathbf{I}\mathbf{Q}^\alpha \mathbf{I}\mathbf{e} \quad (20)$$
$$P_{\text{out},\alpha} = \mathbf{r}^\dagger \mathbf{C}_r^{-1} \mathbf{Q}^\alpha \mathbf{C}_r^{-1} \mathbf{r} - \mathbf{x}^\dagger \mathbf{C}_x^{-1} \mathbf{Q}^\alpha \mathbf{C}_x^{-1} \mathbf{x} \quad (21)$$

The treatment we outlined in the previous section implicitly assumed that signal loss corrections could be treated in expectation. In practice, however, the signal loss itself is a random variable — some realizations of the cosmological signal may be more correlated with foreground modes than others, leading to more signal loss. This leads to two issues. The first is that Equation (13) may not hold. In particular, a third term that is a quadratic combination of $\mathbf{f}$ and $\mathbf{s}$ (e.g., $\mathbf{f}^\dagger \mathbf{C}^{-1} \mathbf{s}$) may appear. However, we can circumvent this issue by decomposing $\mathbf{f}$ into a sum of two vectors, one that is proportional to $\mathbf{s}$ and one that is orthogonal to $\mathbf{s}$. The former can be absorbed into the EoR term (modifying the signal loss factor in the process), while the latter can be absorbed into the foreground term.

The second issue to tackle is how one incorporates the randomness of c_{eor} into our signal loss corrections. Phrased in the context of Bayes' rule, we wish to find the posterior probability distribution of $p(\mathbf{p}_{\text{eor}}|\hat{\mathbf{p}})$ for $\mathbf{p}_{\text{eor}}$ given the uncorrected power spectrum estimate $\hat{\mathbf{p}}$, which is given by

$$p(\mathbf{p}_{\text{eor}}|\hat{\mathbf{p}}) \propto \mathcal{L}(\hat{\mathbf{p}}|\mathbf{p}_{\text{eor}}) p(\mathbf{p}_{\text{eor}}), \quad (22)$$

where $p(\mathbf{p}_{\text{eor}})$ is the prior on $\mathbf{p}_{\text{eor}}$ and $\mathcal{L}$ is the likelihood function. Since the likelihood is defined as the distribution of the measured result $\hat{\mathbf{p}}$ given the theoretical power spectrum $\mathbf{p}_{\text{eor}}$, we may construct this function simply by fixing $\mathbf{p}_{\text{eor}}$, and simulating our analysis pipeline for many realizations of the injected EoR signal consistent with this power spectrum. The resulting distribution can be normalized, and the whole process can then be repeated for a different value of $\mathbf{p}_{\text{eor}}$.

In our code, we simulate 20 realizations of $\mathbf{p}_{\text{eor}}$ by bootstrapping over baselines, as explained in Section 3.2.2, yielding P_{in} values that range from $\sim -10^{13}$ mK^2 (h^{-1} Mpc)^3 (negative P_{in} is injected when one copy of

e in Equation (10) is negative and the other is positive) to $\sim 10^{13}$ mK^2 (h^{-1} Mpc)^3. We also run 40 total EoR injection levels, resulting in a total of 800 data points on our $P_{\rm in}$ vs. $P_{\rm out}$ grid. We smooth the 2D distribution using kernel density estimators. Additionally, we have verified that the transfer function is symmetric for negative and positive $P_{\rm in}$'s, and so we fold all the values into positive ones to increase our signal-to-noise.

The result is shown in the left plot of Figure 15. Bayes' rule then simply instructs us to fix $\hat{\mathbf{p}}$ at our measured value $\hat{\mathbf{p}}_x$, essentially reading Figure 15 horizontally and reinterpreting Equation (22) as a function of $p(\mathbf{p}_{\rm eor})$. The final result can then be normalized to give $p(\mathbf{p}_{\rm eor}|\hat{\mathbf{p}})$. If $\hat{\mathbf{p}}$ itself does not take on a single value but is itself a distribution (for our analysis, this comes from bootstrapping), then one simply repeats the process for every single point on the distribution of $\hat{\mathbf{p}}$, before performing the final summation and normalization. In Figure 15, the peak of our data distribution $\hat{\mathbf{p}}_x$ is marked by the solid gray horizontal lines. From the left plot (inverse covariance weighting), one can eyeball that a data value of $10^5\, mK^2(h^{-1}Mpc)^3$, for example, would map approximately to a value of $\sim 10^8\, mK^2(h^{-1}Mpc)^3$, implying a signal loss factor of ~ 1000. Performing a summation and normalization for the entire distribution of $\hat{\mathbf{p}}_x$ yields a final $P_{\rm in}$ distribution — the distribution of our data as seen through the signal loss lens. We compute power spectrum points from the peak of the histograms, and power spectrum errors from 95% confidence intervals.

As a final complication to our procedure, we note that there will be some intrinsic scatter in our likelihood due to our having only a finite number of simulations. This is shown by the smeared 'heat-map' around the solid black diagonal line (representing unity-transfer) in Figure 15. To separate the intrinsic stochasticity of signal loss from that which arises due to simulation noise, we repeat our analysis for a power spectrum estimator without signal loss ($\mathbf{R} = \mathbf{I}$), shown as the right plot in Figure 15. The width of the lossless scenario's signal loss likelihood is then deconvolved from the lossy scenario's signal loss likelihood. By doing this, we are left with the intrinsic scatter in signal loss, or scatter that stems from how much the random EoR signal **e** happens to look like the data **x**, a quantity we do not know offhand but one that we would like to correct for. As an extreme example, if we are very unlucky, one realization of **e** would have the same shapes, or eigenvectors, as **x**. An empirically-derived covariance would then down-weight these shapes, destroying the entire EoR signal. On the other hand, the less that **e** looks like **x**, the less signal loss that would result. The intrinsic scatter we can get is not a dominant factor in this case but it is important to correct for the fact that a particular $P_{\rm out}$ value could arise from a range of $P_{\rm in}$ values.

One peculiar aspect of Figure 15 is the fact that at low $P_{\rm in}$ values it appears that we can have signal gain ($P_{\rm out} > P_{in}$). This is unphysical in nature but caused due to the dominating cross-terms involving **e** and **x** in Equation (17) once **e** becomes small. As **e** increases, we move into a regime where $P_{\rm out} \sim P_{in}$, and then eventually into a regime where $P_{\rm out} < P_{in}$ when **e** is large enough to be destroyed if weighting the data using itself. Although we only show figures for one k value, we note that the shape of the transfer curve is nearly identical for all k's (though we treat each k separately).

To conclude this method, we also show power spectrum results for fringe-rate filtered PAPER-64 data before and after signal loss correction in Figure 16, using inverse covariance weighting. The blue dashed line represents the unweighted power spectrum, which is identical in both panels (an important check, as we expect no signal loss for this case). The black and gray points are positive and negative power spectrum values plotted with 2σ error bars. Prior to signal loss correction, it is clear that the power spectrum is unfeasible because it is well below the theoretical noise level prediction (solid green curve). Post-correction, the power spectrum errors blow up to be higher than both the theory and unweighted power spectrum, a consequence of not being able to characterize covariances well using the time-averaged fringe-rate filtered data (there is less independent information to use). 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 with minimizing the loss of EoR.

METHOD #2

One of the main difficulties in defining a signal loss procedure is in the interpretation of the injection procedure 'output'. Small differences in approaches were found to cause large differences in the final answer. **[CC: but aren't we arguing the opposite - that they yield similar answers?]**

For Method #1, we used $P_{\rm out}$ as the output of interest. For that method, we had asked the question "What is the input power associated with an excess similar in amplitude to the data power spectrum?" For this second method, we rephrase the question to: "At what point would an injected signal be detectable at significant levels?". Instead of looking at $P_{\rm out}$, we now use P_r, the weighted power spectrum of the data plus signal, as the output of interest.

This question is phrased visually in Figure **??**. In the 2D space relating total P_r to injected $P_{\rm in}$, very small injected levels begin at the level equivalent of no injection (i.e. data only). **[CC: this next sentence does not make sense]** As $P_{\rm in}$ increases, the signal loss transfer curve approaches the power level of the data it does not dominate the covariance but still contributes enough to suppressing itself. Above some level the injected signal is clearly detectable.

One difference in this method is that we are exclusively asking about signals detectable *above* the level of the observed data. Thus it is explicitly limited to plac-

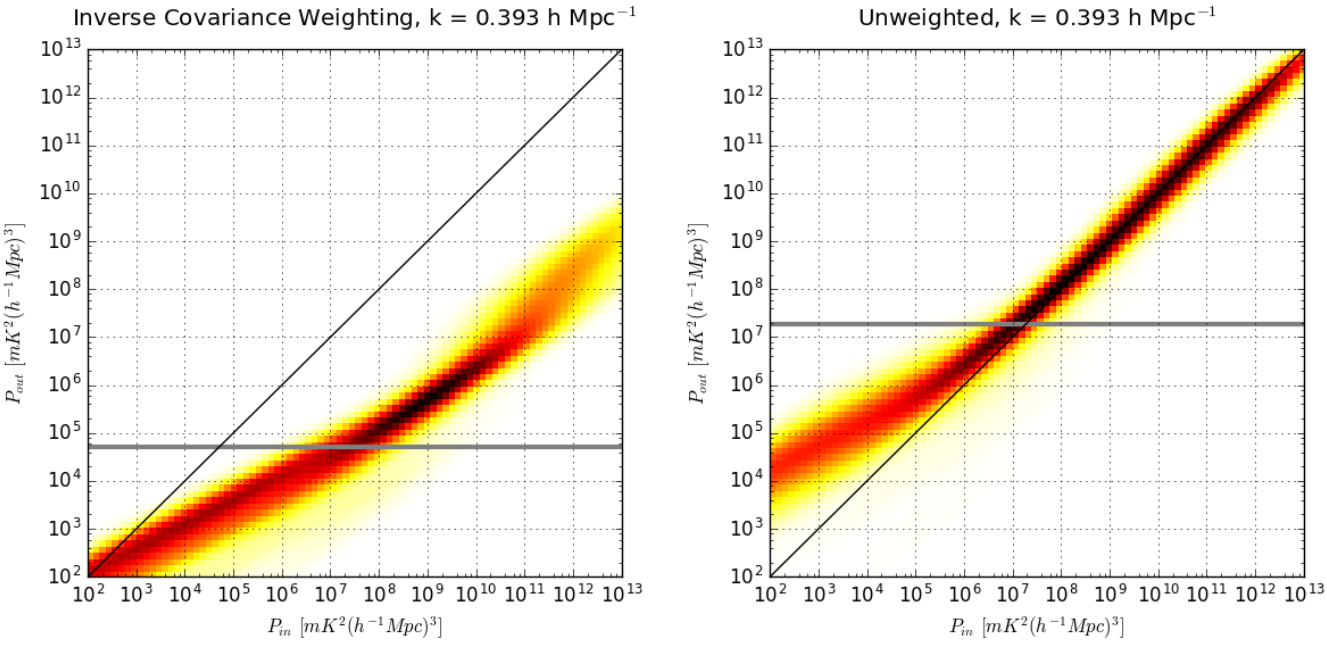

Figure 15. Signal loss transfer functions showing the relationship of $P_{\rm in}$ and $P_{\rm out}$, as defined by Equations (10) and (11). Kernel density estimations of the power spectrum transfer functions are shown as colored heat-maps for the cases of inverse covariance weighted PAPER-64 data (left) and unweighted data (right). The solid black diagonal line marks a perfect unity mapping, and the solid gray horizontal line denotes the peak of $\hat{\mathbf{p}}$, the data distribution. From these plots, it is clear that inverse covariance weighting results in ~ 3 orders of magnitude of signal loss for power spectrum values above $\sim 2 \times 10^4\ mK^2\,(h^{-1}\,Mpc)^3$, whereas the unweighted case does not exhibit loss. The peculiar feature of 'signal gain' at low injection levels is explained in the text.

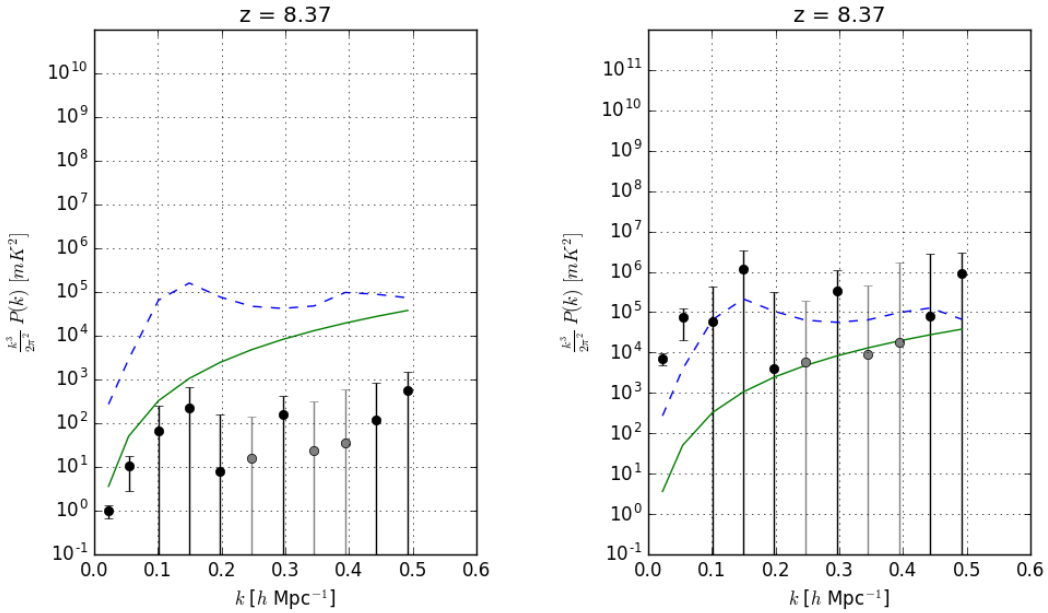

Figure 16. Full inverse covariance weighted power spectrum of PAPER-64 data (positive (black) and negative (gray) points, with 2σ error bars) before signal loss correction (left) and after (right). The dashed blue line is the unweighted power spectrum (2σ upper limit). The solid green line is the theoretical 2σ noise level prediction based on observational parameters, whose calculation is detailed in Section 3.2.1.

ing an upper limit, even if in some k-modes one has actual detections above the noise. This might be seen as more conservative, however, a similar assumption has already been made when choosing to down-weight the data using itself. In such a down-weighting step, one has assumed that significant excess power is associated with a foreground residual. We note that a high SNR detection will require a different method, but we show this one to... **[CC: fill in..]**

Formally we can express our test as: "What is the $P_{\rm in}$ such that

$$P_r > P_x\Lambda, \tag{23}$$

where Λ represents a choice of threshold." Of course, both sides of this equation are actually a distribution of values associated with data variance and injected model variance. One way to interpret this test for the two distributions is to compute the probability that the two distributions are not[2] the same.

$$\mathbf{P}(P_{in}) = 1 - \int_{r=x} \mathbf{P}(P_x)\mathbf{P}(P_r, P_{in})dr \tag{24}$$

where $\mathbf{P}(P_x)$ is the probability of obtaining the power spectrum P_x found via bootstrapping and $\mathbf{P}(P_r, P_{in})$ is the probability of obtaining the summed power spectrum P_r given a range of input power levels, as sampled in Figure **??**.

Probability vs. injection level is shown in Figure **??**. At low injection levels the variation between x and r is such that the null test is ruled out to within 20%. This is the level expected from two distributions with ∼10 effective samples. **[DCJ: checked with a little python test but needs a cite]** As the injected signal approaches the level of the data, the probability of null rejection rises steeply.

Lastly, this method requires that we choose the probability at which we choose to reject the null hypothesis. As can be seen in Figure **??**, the difference between 80 and 90% can mean a difference of almost an order of magnitude in the power spectrum level that is ruled out. We can also see that the probability does not perfectly converge at high injected power levels. **[CC: why?]** Rather than insist on a rejection at the 99th percent level, we compromise at 90% **[DCJ: Matt, put in actual value here]** **[MJK: The current image uses 90th percentile!]** .

The resulting limits for inverse covariance weighted PAPER-64 data (the same data used in Method #1), using this second signal loss method, are shown in Figure 17 next to the limits from Method #1. In the rest of this paper, we use Method #1 **[CC: why?]**.

3.1.3. *Minimizing Signal Loss*

With a signal loss formalism established (Method #1, for the rest of this paper), we now have the capability of experimenting with different weighting options for

[2] i.e. the null test

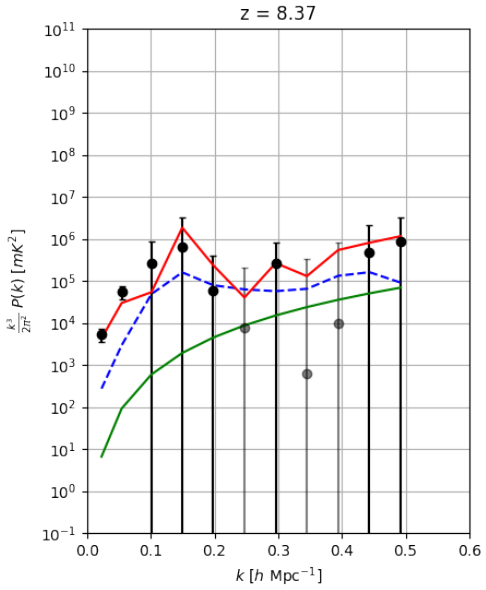

Figure 17. A comparison of the Method #1 signal loss corrected power spectrum (black and gray points) and Method #2 uppers limits at 90% confidence (red curve). Also shown are the upper limits of the unweighted power spectrum (blue dashed) and theoretical noise model (green).

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. We have found that the balance between the two is a delicate one and requires a careful understanding and altering of covariance matrices.

We saw in Section 2.1.3 how limiting the number of down-weighted eigenmodes (i.e. flattening out part of the eigenspectrum) can help minimize signal loss. We experiment with this idea on PAPER-64 data, dialing the number of modes down-weighted from zero (which is equivalent to identity-weighting, or the unweighted case) to 21 (which is full inverse covariance weighting of our 21 channels). The power spectrum results for one k value, both before and after signal loss correction, are shown in Figure 18. 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 correction, increases with the number of down-weighted modes (black curve). Because our data is fringe-rate filtered, we have less information to characterize the covariance matrix and therefore do a worse job estimating it. Hence, the more modes we use in down-weighting, the greater the error we have in estimating the power spectrum.

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 describe foreground modes, but not EoR modes. We see in Figure 18 that this occurs when only the strongest $\sim$ 3-4 eigenmodes are down-weighted, though the improvement from the unweighted case is not significant.

In addition to our thermal noise prediction (green), two additional horizontal lines are shown in Figure 18 which denote power spectrum values, post-signal loss correction, for two other regularization schemes. We multiply an identity matrix element-wise to $\hat{\mathbf{C}}$ (i.e. inverse variance weighting, shown in red), and also add an arbitrarily chosen level of $\mathbf{I}$ to $\hat{\mathbf{C}}$ (cyan). We see that all three regularization schemes (black, red, cyan) perform similarly at their best (i.e. when $\sim$ 3-4 eigenmodes are down-weighted in the case of the black curve).

For the remainder of this paper, we choose to use the weighting option of $\hat{\mathbf{C}} = \hat{\mathbf{C}} \circ \mathbf{I}$, or inverse variance weighting, which we will denote as $\hat{\mathbf{C}}_{eff}$. We choose this weighting scheme due to its simplicity in practice and because its effectiveness is comparable to other weighting schemes that produce minimal signal loss.

Our revised, best to-date PAPER-64 power spectrum (using only one baseline separation type) is shown in Figure 19. Again, black and gray points correspond to positive and negative power spectrum values respectively, with 2σ errors bars. Also plotted are the unweighted power spectrum upper limit (dashed blue) and theoretical prediction of noise (solid green). From this result, we quote a best 2σ upper limit of $(158.0\ \mathrm{mK})^2$ at $k = 0.34\ h\mathrm{Mpc}^{-1}$, a higher limit than P64 by a factor of $\sim$ 7 in mK (though we only use one baseline type in our analysis as opposed to 3). **[CC: replace with kolopanis limit instead later]**

In this section we have shown three simple ways of regularizing $\hat{\mathbf{C}}$ 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 $\hat{\mathbf{C}}$ using information from different subsets of baselines. For redundant arrays this could mean calculating $\hat{\mathbf{C}}$ from a different but similar baseline type, such as the $\sim$ 30 m diagonal PAPER baselines (instead of the horizontal E/W ones). Alternately, covariances 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) in order to avoid suppressing the 21 cm signal, and it's worth noting that the PAPER-32 results are likely safe 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., *in prep*.

Another possible way to regularize $\hat{\mathbf{C}}$ is to use information from different ranges of LST. For example, one could calculate $\hat{\mathbf{C}}$ with data from LSTs where foregrounds are stronger (earlier or later LSTs than the 'foreground- quiet' range used in forming power spectra) — doing so may yield a better description of the foregrounds that we desire to down-weight. Fundamentally, each of these examples are similar in that they rely on a computation of $\hat{\mathbf{C}}$ 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.

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 a derivation 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 P64 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.

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

$$p(k) = \frac{X^2 Y \Omega_{eff} T_{sys}^2}{\sqrt{2 N_{lsts} N_{seps}}\, t_{int} N_{days} N_{bls} N_{pols}} \tag{25}$$

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

- Ω_{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 parts of the

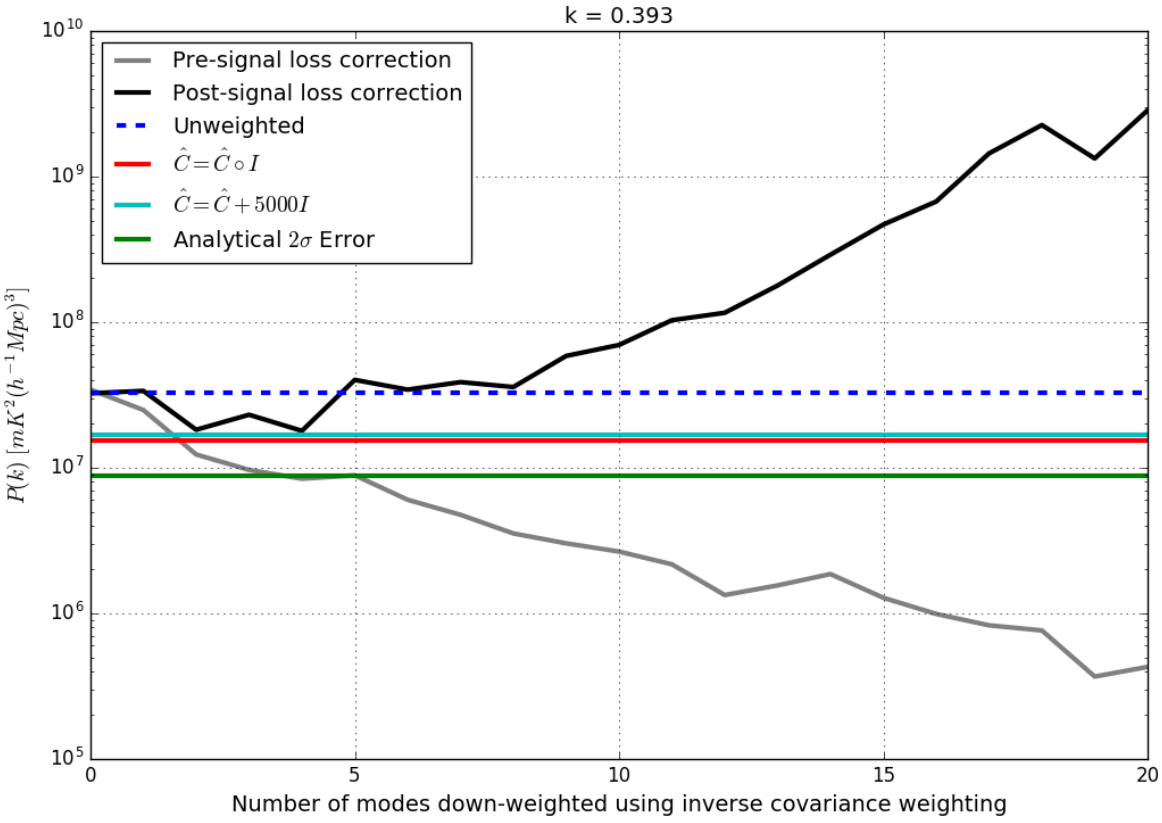

Figure 18. Power spectra 2σ upper limits for $k = 0.393$ h Mpc^{-1} for fringe-rate filtered PAPER-64 data. Values are shown before (gray) and after (black) signal loss correction as a function of number of eigenmodes of $\hat{\mathbf{C}}$ that are down-weighted. This regularization knob is tuned from 0 modes (i.e. unweighted) to 21 modes (i.e. full inverse covariance weighting). Over ~ 3 orders of magnitude of signal loss results when using inverse covariance weighting. Also plotted for comparison are 2σ power spectrum upper limits for the unweighted case (dashed blue), inverse variance weighted case (red), and added-identity case (cyan). All three regularizations shown (black, red, cyan) perform similarly (when $\sim$3 modes are down-weighted for the black curve). Finally, a theoretical prediction for noise (2σ error) is plotted as solid green.

beam are up-weighted and down- weighted. Using numbers from Table 1 in Parsons et al. (2016), $\Omega_{eff} = 0.74^2/0.24$ for an optimal fringe-rate filter.

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

$$T_{sys} = 180\left(\frac{\nu}{0.18}\right)^{-2.55} + T_{rcvr}, \qquad (26)$$

where ν are frequencies in GHz. We use a receiver temperature of 144 K, yielding $T_{sys} = 431$ K at 150 MHz. This is lower than the T_{sys} of 500 K used in P64 because of several small miscalculation errors that were identified[3].

- $\sqrt{2}$: This factor in the denominator of the sensitivity equation comes from taking the real part of the power spectrum estimates after squaring visibility measurements. In P64, a factor of 2 was mistakenly used.

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

- N_{seps}: The number of baseline separation types averaged incoherently in a final power spectrum estimate. For the analysis in this paper, we only use one type of baseline, hence $N_{seps} = 1$. The updated limits in Kolopanis et al., *in prep* use three separation types.

- t_{int}: The integration time 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

[3] such as missing a square root in going from a variance to a standard deviation

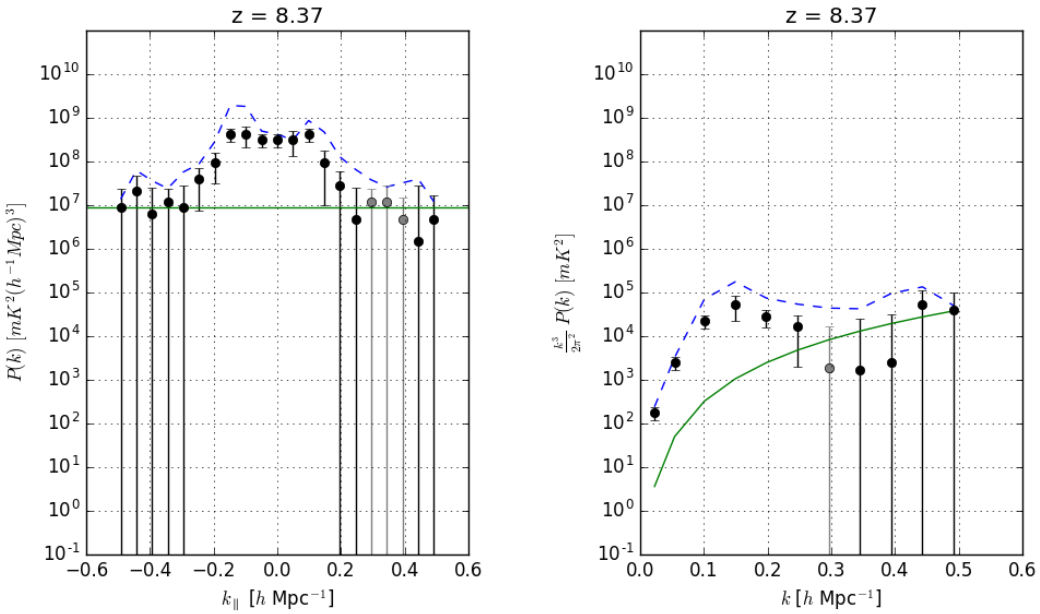

Figure 19. Power spectrum of PAPER-64 using $\hat{\mathbf{C}}_{eff}$. Black and gray points correspond to positive and negative power spectrum values, respectively, with 2σ error bars also plotted. The dashed blue line is the unweighted power spectrum (2σ upper limit). The solid green line is the theoretical 2σ noise level prediction based on observational parameters. This power spectrum result differs from P64 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. We quote a best 2σ upper limit of $(158.0\ \mathrm{mK})^2$ at $k = 0.34\ h\mathrm{Mpc}^{-1}$, a higher limit than P64 by a factor of ~ 7 in mK **[CC: replace with kolopanis limit later]**.

using the following:

$$t_{frf} = t_{int} \frac{\int 1\, df}{\int w^2(f)\, df}, \tag{27}$$

where $t_{int} = 43$ seconds, t_{frf} 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 $t_{int} = 3857$ s.

- N_{days}: The total number of days of data analyzed. In P64, 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 realistic value of N_{days} as:

$$\frac{1}{N_{days}} = \sqrt{\left\langle \frac{1}{N_i^2} \right\rangle_i}, \tag{28}$$

where i is over LST (Jacobs et al. 2015). For PAPER-64, our revised estimate of N_{days} is ~ 34 days.

- N_{bls}: The number of baselines contributing to the sensitivity of a power spectrum estimate. In P64, this number was the total number of 30 m East/West baselines used in the analysis. However, using the total number of baselines neglects the fact that baselines are averaged into groups before cross-multiplying data. Our revised estimate for the parameter is:

$$N_{bls} = \frac{N_{bls}}{N_{gps}} \sqrt{N_{gps}^2 - N_{gps}}, \tag{29}$$

where, for the P64 analysis, $N_{gps} = 5$. Each baseline group averages down linearly as the number of baselines entering the group (N_{bls}/N_{gps}) and then as the square root of the number of cross-multiplied pairs ($\sqrt{N_{gps}^2 - N_{gps}}$). For the revised PAPER-64 analysis with only one baseline separation type, this becomes $N_{bls} \sim 46$ instead of 51.

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

An additional factor of $\sqrt{2}$ is gained in sensitivity when folding our power spectra into $\Delta^2(k)$, due to averaging together positive and negative k's.

Our revised sensitivity estimate for PAPER-64 is shown in comparison with that of P64 in Figure 20. 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_{sys} (neglecting sky variation) but accounting for the true N_{days} as determined by LST sampling counts for each

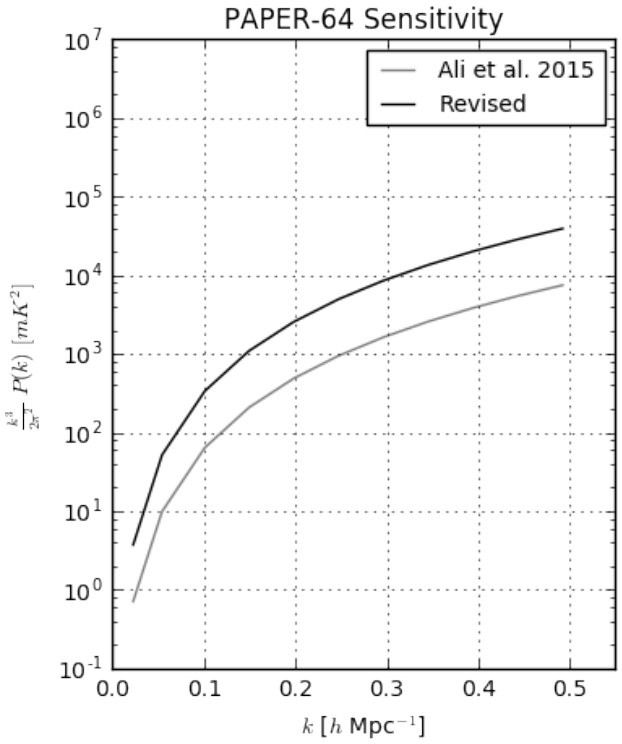

Figure 20. An updated prediction for the noise-level of PAPER-64 data (black) is shown in comparison to previously published sensitivity limits (gray). Both sensitivity analyses plotted assume only one baseline type (an additional factor of $\sqrt{3}$ for 3 baseline types is needed to match P64 exactly). Major factors that contribute to the discrepancy are Ω_{eff}, N_{days} and N_{bls}, as described in Section 3.2.1, which when combined decreases our sensitivity (higher noise floor) by a factor of ~ 5 in mK^2.

time and frequency in the LST-binned data. We convert T_{sys} into a variance statistic using:

$$T_{rms} = \frac{T_{sys}}{\sqrt{\Delta\nu \Delta t N_{days} N_{pols}}}, \tag{30}$$

where $\Delta\nu$ is channel spacing, Δt is integration time, N_{days} is the number of daily counts for a particular time and frequency that went into our LST-binned set, and N_{pols} is the number of polarizations (2 for Stokes I). This RMS 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 21, where the black and gray points represent positive and negative power spectrum values, respectively (with 2σ error bars and weighting matrix $\hat{\mathbf{C}}_{eff}$), the dashed blue line represents the unweighted power spectrum, and the solid green line denotes our 2σ theoretical noise prediction as calculated by Equation (25). All three are in 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 two major changes in the way we estimate errors since P64, using the lessons we've learned about sampling with replacement and 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. The PAPER-64 pipeline outputs 20 bootstraps (over baselines), each a 2-dimensional power spectrum that is a function of k and time.

In P64, a second round of bootstrapping occurred over the time axis. A total of 400 bootstraps were created in this step ($N_{boot} = 400$), 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, which we expect to greatly exceed the approximate number of independent samples after fringe-rate filtering. Medians were then taken of the values in each bootstrap (with the appropriate median correction factor applied). 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_{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 with the "single replacement" strategy described in Section 2.2.

The single replacement strategy also provides a higher sensitivity estimate as only at most one baseline is repeated in each group. In P64, baselines were chosen for groups randomly with replacement. In doing so, $\sim$3-4 baselines per group were repeated (Figure 10 shows the fraction of independent samples to be $\sim$55% for 10 total independent samples) which sacrificed some sensitivity.

Power spectrum estimates with these bootstrapping changes for PAPER-64 fringe-rate filtered data are shown in Figure 22. The estimates are unweighted in order to disentangle the effects of bootstrapping from signal loss. As shown in the figure, when the baseline average includes many repeated samples our sensitivity is not maximized, while when only one sample in the same average is allowed the possibility of being a repeat, the error on the noise-dominated higher k modes converges to the theoretical noise level. However, if 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 below the predicted noise level.

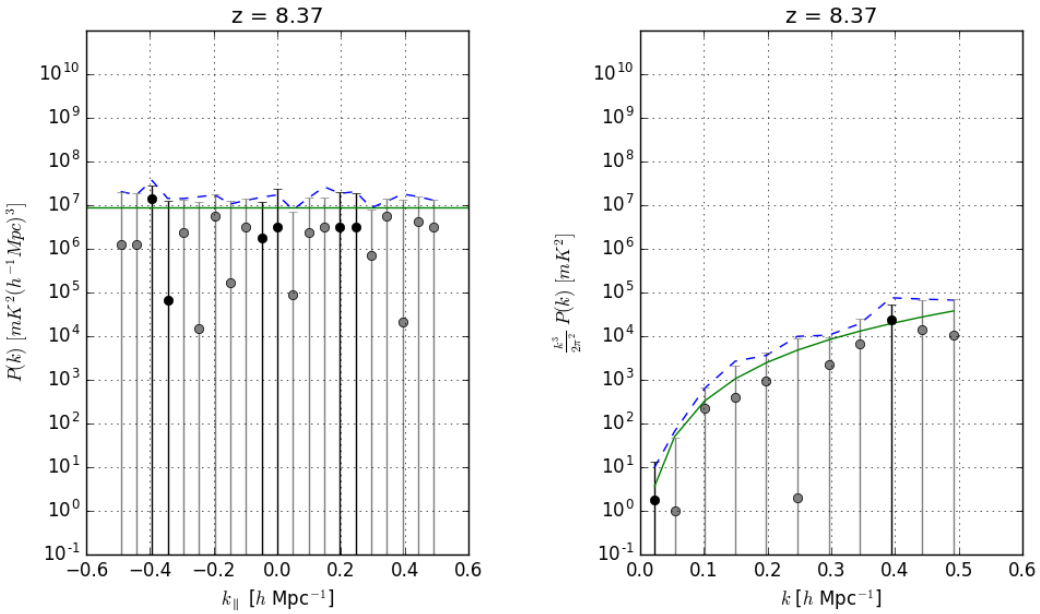

Figure 21. Power spectrum estimate for a noise simulation that mimics the noise level of PAPER-64 data. The weighted power spectrum points and their 2σ errors are show in black and gray (positive and negative values), where we use $\hat{\mathbf{C}}_{eff}$ to minimize signal loss. The dashed blue line is the unweighted power spectrum (also 2σ upper limit). The solid green line is the theoretical 2σ noise level prediction as calculated by Equation (25). All three estimates agree (the analytic curve should encompass 95% of the points).

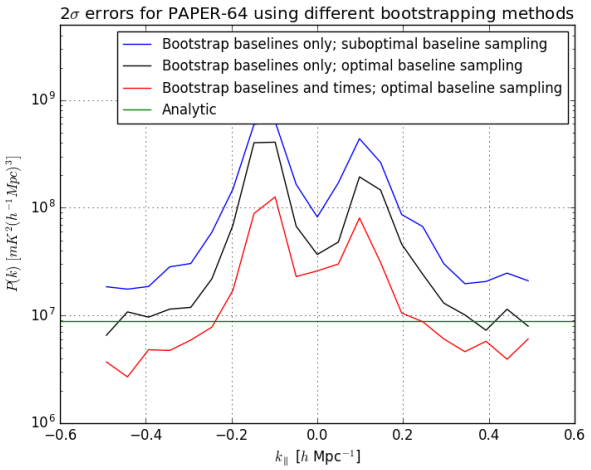

Figure 22. 2σ power spectrum errors (not upper limits) using PAPER-64 data and three different bootstrapping methods. The data is fringe-rate filtered and a weighting matrix of $\mathbf{I}$ (unweighted) is used in order to disentangle the effects of bootstrapping from signal loss. Three different bootstrapping methods are shown: the blue and black use methods where bootstrapping occurs on the baseline axis only (not the time axis), and they differ by how the baselines are randomly sampled. It is evident that sensitivity is maximized when baselines are sampled in a way that ensures mostly independent samples (only the last baseline spot is sampled randomly), as shown by the black curve. In red, bootstrapping occurs over both the baseline and time axes, illustrating how errors can be under-estimated if sampling more values than independent ones (fringe-rate filtering reduces the number of independent samples). We use the method illustrated by the black curve in our updated analysis, which is shown to agree with the 2σ analytic prediction for noise (green).

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 EoR. 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 in which we mitigate foreground leakage in a power spectrum estimate, and two in which 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. 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 data, we compute $\mathbf{M}$ using the Fisher matrix $\mathbf{F}$, defined as:

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{tr}[\mathbf{R}\mathbf{Q}^\alpha\mathbf{R}\mathbf{Q}^\beta] \tag{31}$$

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:

$$\mathbf{F} = \mathbf{L}\mathbf{L}^\dagger. \tag{32}$$

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

$$\mathbf{M} = \mathbf{D}\mathbf{L}^{-1} \tag{33}$$

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

$$\mathbf{W} = \mathbf{D}\mathbf{L}^{\dagger}. \tag{34}$$

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.

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 $\hat{q}$ 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 $\hat{q}$ 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.

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 14 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 P64. Most notably, power spectrum detections 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 Tests*

The highest sensitivity power spectrum result for PAPER-64 using the updated analysis presented in this paper, shown in Figure 19, has positive biases at low k values. As discussed in Section 2.3.1, these detections are most likely attributable to foreground leakage. Here we demonstrate 3 null tests performed on PAPER-64 data that verify that the positive detections are indeed due to foreground variation and not attributable to EoR.

The 3 null test results are shown in Figure 23, with each test described as the following:

- Original (black): This is identical to the power spectrum in Figure 19 and represents the best revised PAPER-64 power spectrum (one baseline type only) with weighting matrix $\hat{\mathbf{C}}_{eff}$. There are clear detections at low k's.

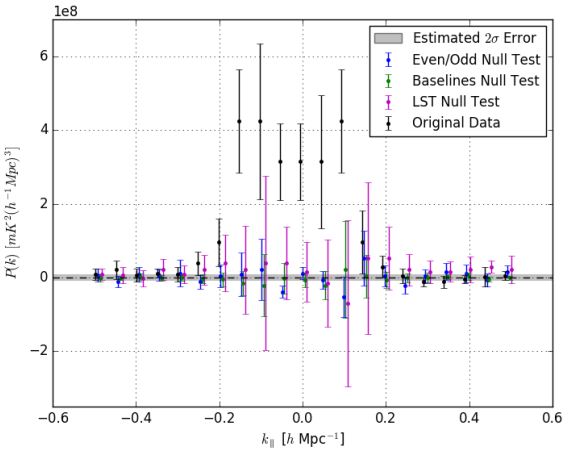

Figure 23. Results for 3 null tests compared to the PAPER-64 revised power spectrum (black) and analytically estimated 2σ errors (gray shaded region). For each, we took jackknives along different axes of the dataset - by julian days (separating even and odd days; blue), by baselines (green), and by LST (purple). We expect the sky signal to disappear for a 'passing' null test. We find that the biases in our power spectrum are likely caused by variation of foregrounds in LST, and are not attributable to particular baselines or days of data.

- Even/Odd (blue): We split our dataset into even (e) and odd (o) days, where $\mathbf{x}_1 = e + o$ and $\mathbf{x}_2 = e - o$. We form datasets in this way to ensure that we use the full sensitivity of our data. When cross-multiplied, we obtain:

$$\mathbf{x}_1\mathbf{x}_2^{\dagger} = ee^{\dagger} - eo^{\dagger} + oe^{\dagger} - oo^{\dagger} \tag{35}$$

  If the same sky signal is in both the even and odd datasets, we expect it to cancel out.

- Baselines (green): We split our dataset into two halves, where each contains half of the total baselines used in the analysis. Again, we form $\mathbf{x}_1 = b_1 + b_2$ and $\mathbf{x}_2 = b_1 - b_2$, where b_1 is the first baseline group and b_2 is the second baseline group.

- LST (purple): We split our dataset into two halves based on LST, namely t_1 (LSTs 0.5-4.6 hours) and t_2 (LSTs 4.6-8.6 hours). We form our two datasets as $\mathbf{x}_1 = t_1 + t_2$ and $\mathbf{x}_2 = t_1 - t_2$.

Investigating Figure 23, we find that both the baselines and even/odd null tests are consistent with noise. We note that the error bars for the baselines test (green) are especially small — this is because the signal is effectively cancelled out prior to cross-multiplication (when forming baseline groups).

The null test in LST (purple) is much more illuminating. The fact that there are many power spectrum points outside the estimated error region signifies that the bias we see is likely from foregrounds that vary in LST. In other words, t_1 and t_2 differ by an amount greater than thermal noise and they do not effectively cancel out the sky signal during cross- multiplication.

The errors for this null test are also large, also reflective of residual foregrounds.

We do not investigate the detailed nature of the residual foregrounds in this paper, though one could imagine performing many similar null tests with a sliding window of t_1's and t_2's. This would illuminate the particular LST range that introduces the foreground bias (or whether the bias is constant in LST), and could potentially be traced to an individual bright radio source. This type of detailed analysis will be critical at EoR sensitivities, however, for our current analysis we are not surprised that the bias we see comes from foregrounds in regions where we expect leakage.

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 if more eigenmodes are downweighted than exist in the data (having an underconstrained covariance matrix). Hence, there exists a trade-off between sensitivity driven time-averaging techniques such as fringe-rate filtering and signal loss.
- Signal injection and recovery simulations can be used to quantify signal loss (Section 3.1). However, a signal-only simulation (i.e. comparing an unweighted vs. weighted power spectrum of EoR only) can underestimate loss by failing to account for spurious correlations between the data and signal.
- Errors that are estimated via bootstrapping can be underestimated if samples in the dataset are significantly correlated (Section 2.2). 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. For an analysis of only one baseline type, we quote a new 2σ upper limit of $(158.0\ \mathrm{mK})^2$ at $k = 0.34\ h\mathrm{Mpc}^{-1}$, a higher limit than P64 by a factor of ~ 7 in mK **[CC: replace with kolopanis limit later]**. The reasons for a previously underestimated limit and ways in which our new analysis differs can be summarized by the following:

- Signal loss, previously found to be $< 2\%$ in P64, was underestimated by a factor of $\sim$1000 for 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 ~ 7 mK higher than previous limits. Underestimated signal loss therefore represents the bulk of our revision. We note that PAPER's analysis is not the first to underestimate loss; results from the GMRT (Paciga et al. 2013) were also revised from new signal loss calculations associated with their singular value decomposition foreground filter.
- Power spectrum errors, originally computed by bootstrapping, were underestimated by a factor of $\sim$2 due to oversampling data whose effective number of independent samples was reduced from fringe-rate filtering. They were also overestimated by a factor of $\sim$2 from using a suboptimal way of combining baselines into groups when forming power spectra (Section 3.2.2). Hence, the net change for our error estimation is negligible, though the rationale behind the previous estimation was incorrect.
- 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 $\sim$2 in mK (Section 3.2.1). We have verified our revised prediction extensively using pure noise simulations. 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 P64 was too high for its data points.

The 21 cm future is exciting, as new experiments have sensitivities that expect to reach and surpass EoR levels, improved foreground mitigation/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. Looking ahead, the future promises to bring higher sensitivity measurements, better foreground and instrumental models, more

effective and less lossy weighting techniques — and with all that, an improved understanding of reionization and our early Universe.

5. ACKNOWLEDGEMENTS

CC would like to acknowledge the UC Berkeley Chancellor's Fellowship, National Science Foundation Graduate Research Fellowship (Division of Graduate Education award 1106400), and thank Eric Switzer 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). We also 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 build intuition by deriving an analytic formula for power spectrum signal loss. We will also show that in general, signal loss appears as a multiplicative bias on one's power spectrum estimate.

The minimum-variance quadratic estimator $\hat{p}_\alpha$ for the αth bandpower of the power spectrum is given by

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

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[4]. In our case, however, we do not have *a priori* knowledge of the covariance matrix. Thus, we replace $\mathbf{C}$ with $\hat{\mathbf{C}}$, its data-derived approximation. Our estimator then becomes

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

where

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

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

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

where η is parameter quantifying our success at estimating the true covariance matrix. If $\eta = 0$, our covariance estimate has perfectly modeled the true covariance and $\hat{\mathbf{C}} = \mathbf{C}$. 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 $\hat{p}_\alpha^{\mathrm{loss}}$ in terms of $\hat{p}_\alpha$. We begin by expressing $\hat{\mathbf{C}}^{-1}$ in terms of $\mathbf{C}^{-1}$ using the Woodbury identity so that

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

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

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

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

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

[4] For this section only, with no loss of generality, we assume that the data $\mathbf{x}$ are real.

Both for the purposes of analytical tractability and to provide intuition, we expand this expression to leading order in η. The result is

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

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

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

where N is the length of $\mathbf{x}$. 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. Now, recall that p_α is the *true* power spectrum. This means that it can be decomposed into the sum of foreground and EoR power spectra, since the foregrounds and EoR are uncorrelated in expectation. This toy example, while not definitive, serves to motivate the form of Equation (13).

REFERENCES

Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166
Ali, Z. S., et al. 2015, ApJ, 809, 61
Bernardi, G., et al. 2009, A&A, 500, 965
—. 2010, A&A, 522, A67+
Bond, J. R., Jaffe, A. H., & Knox, L. 1998, PhRvD, 57, 2117
Bowman, J. D., & Rogers, A. E. E. 2010, Nature, 468, 796
Burns, J. O., et al. 2012, Advances in Space Research, 49, 433
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., et al. 2015, Phys. Rev. D, 91, 123011
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
Greenhill, L. J., & Bernardi, G. 2012, ArXiv e-prints
Jacobs, D. C., et al. 2015, ApJ, 801, 51
Jelić, V., et al. 2008, MNRAS, 389, 1319
Keating, G. K., Marrone, D. P., Bower, G. C., Leitch, E., Carlstrom, J. E., & DeBoer, D. R. 2016, The Astrophysical Journal, 830, 34
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
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. 2013, MNRAS
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
Patra, N., Subrahmanyan, R., Sethi, S., Udaya Shankar, N., & Raghunathan, A. 2015, ApJ, 801, 138
Peterson, U.-L. P. X.-P. W. J. 2004, ArXiv Astrophysics e-prints
Pober, J. C., et al. 2012, The Astronomical Journal, 143, 53
Pober, J. C., et al. 2013, AJ, 145, 65
—. 2014, ApJ, 782, 66
Pober, J. C., et al. 2016, The Astrophysical Journal, 819, 8
Pritchard, J. R., & Loeb, A. 2012, Reports on Progress in Physics, 75, 086901
Santos, M. G., Cooray, A., & Knox, L. 2005, ApJ, 625, 575
Sokolowski, M., et al. 2015, PASA, 32, e004
Switzer, E. R., Chang, T.-C., Masui, K. W., Pen, U.-L., & Voytek, T. C. 2015, The Astrophysical Journal, 815, 51
Tegmark, M. 1997, PhRvD, 55, 5895
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
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
Wu, X. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, 474