The Astrophysical Journal, 868:26 (21pp), 2018 November 20 https://doi.org/10.3847/1538-4357/aae833

Characterizing Signal Loss in the 21 cm Reionization Power Spectrum: A Revised Study of PAPER-64

Carina Cheng[1], Aaron R. Parsons[1,2], Matthew Kolopanis[3], Daniel C. Jacobs[3], Adrian Liu[1,4,15], Saul A. Kohn[5], James E. Aguirre[5], Jonathan C. Pober[6], Zaki S. Ali[1], Gianni Bernardi[7,8,9], Richard F. Bradley[10,11,12], Chris L. Carilli[13,14], David R. DeBoer[2], Matthew R. Dexter[2], Joshua S. Dillon[1,16], Pat Klima[11], David H. E. MacMahon[2], David F. Moore[5], Chuneeta D. Nunhokee[8], William P. Walbrugh[7], and Andre Walker[7]

[1] Astronomy Department, University of California, Berkeley, CA, USA; ccheng@berkeley.edu
[2] Radio Astronomy Laboratory, University of California, Berkeley, CA, USA
[3] School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
[4] Berkeley Center for Cosmological Physics, Berkeley, CA, USA
[5] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA
[6] Department of Physics, Brown University, Providence, RI, USA
[7] Square Kilometre Array, Cape Town, South Africa
[8] Department of Physics and Electronics, Rhodes University, South Africa
[9] INAF–Instituto di Radioastronomia, Bologna, Italy
[10] Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, USA
[11] National Radio Astronomy Observatory, Charlottesville, VA, USA
[12] Department of Astronomy, University of Virginia, Charlottesville, VA, USA
[13] National Radio Astronomy Observatory, Socorro, NM, USA
[14] Cavendish Laboratory, Cambridge, UK
Received 2018 June 22; revised 2018 October 9; accepted 2018 October 10; published 2018 November 15

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. Because this signal is very faint and difficult to isolate, it is crucial to develop analysis techniques that maximize sensitivity and suppress contaminants in data. It is also imperative to understand the trade-offs between different analysis methods and their effects on power spectrum estimates. Specifically, with a statistical power spectrum detection in HERA's foreseeable future, it has become increasingly important to understand how certain analysis choices can lead to the loss of the EoR signal. In this paper, we focus on signal loss associated with power spectrum estimation. We describe the origin of this loss 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 signal loss have led to a revised, higher 21 cm power spectrum upper limit from PAPER-64. Additionally, we summarize errors associated with power spectrum error estimation that were previously unaccounted for. We focus on a subset of PAPER-64 data in this paper; revised power spectrum limits from the PAPER experiment are presented in a forthcoming paper by Kolopanis et al. and supersede results from previously published PAPER analyses.

Key words: dark ages, reionization, first stars – early universe – large-scale structure of universe – methods: data analysis – methods: statistical – techniques: interferometric

1. Introduction

By about one billion years after the Big Bang ($z \sim 6$), the first stars and galaxies are thought to have ionized all the neutral hydrogen that dominated the baryonic matter content in the universe. This transition period, during which the first luminous structures formed from gravitational collapse and began to emit intense radiation that ionized the cold neutral gas into a plasma, is known as the Epoch of Reionization (EoR). The EoR is a relatively unexplored era in our universe's history that spans the birth of the first stars to the full reionization of the intergalactic medium (IGM). This epoch encodes important information regarding the nature of the first galaxies and the processes of structure formation. Direct measurements of the EoR would unlock powerful characteristics about the IGM, revealing connections between the matter distribution exhibited via cosmic microwave background (CMB) studies and the highly structured web of galaxies we observe today (for a review, see Barkana & Loeb 2001; Furlanetto et al. 2006; Loeb & Furlanetto 2013).

One promising technique to probe the EoR is to target the 21 cm wavelength signal that is emitted and absorbed by neutral hydrogen via its spin-flip transition (Furlanetto et al. 2006; Barkana & Loeb 2008; Morales & Wyithe 2010; Pritchard & Loeb 2010, 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. Hence, 21 cm tomography offers a unique window into both the spatial and temporal evolution of ionization, temperature, and density fluctuations.

In addition to the first tentative detection of our Cosmic Dawn (pre-reionization era) made by the Experiment to Detect the Global EoR Signature (EDGES; Bowman & Rogers 2010;

[15] Hubble Fellow.
[16] NSF AAPF Fellow.

Bowman et al. 2018), 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 21 cm signal relative to the CMB are the Large Aperture Experiment to Detect the Dark Ages (LEDA; Bernardi et al. 2016), the Dark Ages Radio Explorer (DARE; Burns et al. 2012), the Sonda Cosmológica de las Islas para la Detección de Hidrógeno NeutroSciHi (SCI-HI; Voytek et al. 2014), the Broadband Instrument for Global HydrOgen ReioNisation Signal (BIGHORNS; Sokolowski et al. 2015), and the Shaped Antenna measurement of the background RAdio Spectrum (SARAS; Patra et al. 2015). Radio interferometers that seek to measure statistical power spectra include the Giant Meter-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; Pen et al. 2004; Wu 2009), the Square Kilometer Array (SKA; Koopmans et al. 2015), and the Donald C. Backer Precision Array for Probing the Epoch of Reionization (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 has a forecasted sensitivity capable of a high-significance power spectrum detection with an eventual 350 elements using current analysis techniques (Pober et al. 2014; Dillon & Parsons 2016; 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 (e.g., Santos et al. 2005; Ali et al. 2008; de Oliveira-Costa et al. 2008; Jelić et al. 2008; Bernardi et al. 2009, 2010, 2013; Ghosh et al. 2011; Pober et al. 2013a; Dillon et al. 2014; Kohn et al. 2016). A clean measurement therefore requires an intimate understanding of how data analysis choices, which are often tailored to maximize sensitivity and minimize contaminants, affect power spectrum results. More specifically, it is imperative to develop techniques that ensure the accurate extraction and recovery of the EoR signal, despite the analysis method chosen and how much loss (of both contaminants and the EoR signal) accompanies the method. In this paper, we specifically discuss signal loss—the loss of the cosmological signal—associated with power spectrum estimation. This is an issue that is essential to investigate for a robust 21 cm power spectrum analysis and one that has motivated a revised PAPER analysis. We first approach this topic from a broad perspective and then perform a detailed case study using data from the 64-element configuration of PAPER. In this study, we use a subset of PAPER-64 data to illustrate our revised analysis methods, while a related paper, M. Kolopanis et al. (2018, in preparation), builds off of the methods in this paper to present revised PAPER-64 results for multiple redshifts and baseline types.

Finally, we also highlight several additional errors made in previous PAPER analyses, including those related to bootstrapping and error estimation. This paper accompanies the erratum of Ali et al. (2018) and adds to the growing foundation of lessons that have been documented in, for example, Paciga et al. (2013), Patil et al. (2016), and Jacobs et al. (2016) by the GMRT, LOFAR, and MWA projects, respectively. These lessons are imperative as the community as a whole moves toward higher sensitivities and potential EoR detections.

This paper is organized into four main sections. In Section 2, we use toy models to develop intuition about signal loss and its origin and subtleties. In Section 3, we present a case study using data from the PAPER-64 array, highlighting key changes from the signal loss methods used in the published result in Ali et al. (2015, hereafter A15), which previously underestimated loss. In Section 4, we summarize additional lessons learned since A15 that have shaped our revised analysis. Finally, we conclude in Section 5.

2. Signal Loss Toy Models

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

One type of signal loss can occur when weighting data by itself. Broadly speaking, a data set can be weighted to emphasize certain features and minimize others. For example, one flavor of weighting employed by previous PAPER analyses is inverse covariance weighting in frequency, which is a generalized version of inverse variance weighting that also takes into account frequency correlations (Liu & Tegmark 2011; Dillon et al. 2013, 2014, 2015; Liu et al. 2014a, 2014b). Using such a technique enables the down-weighting of contaminant modes that obey a different covariance structure from that of cosmological modes. However, a challenge of inverse covariance weighting is in estimating a covariance matrix that is closest to the true covariance of the data; the discrepancy between the two, as we will see, can have large impacts on signal loss. In this paper, we focus specifically on loss associated with the use of an empirically estimated covariance matrix with the "optimal quadratic estimator (QE)" formalism. This loss was significantly underestimated in the A15 analysis and is the main reason motivating a revised power spectrum result.

2.1. The QE Method

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

Suppose that the measured visibilities for a single baseline in Jy are arranged as a data vector, $\boldsymbol{x}$. It has length $N_t N_f$, where N_t

is the number of time integrations and N_f is the number of frequency channels. The covariance of the data is given by

$$\boldsymbol{C} \equiv \langle \boldsymbol{x}\boldsymbol{x}^{\dagger} \rangle = \boldsymbol{S} + \boldsymbol{U}, \tag{1}$$

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

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

$$\boldsymbol{S} = \sum_{\alpha} P^{\alpha} \frac{\partial \boldsymbol{C}}{\partial P^{\alpha}} \equiv \sum_{\alpha} P^{\alpha} \boldsymbol{Q}^{\alpha}, \tag{2}$$

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

The optimal QE prescription is then to compute

$$\widehat{P}^{\alpha} = \sum_{\beta} (\boldsymbol{F}^{-1})^{\alpha\beta} (\widehat{q}^{\beta} - \widehat{b}^{\beta}), \tag{3}$$

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

$$F^{\alpha\beta} \equiv \frac{1}{2} \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{Q}^{\alpha}\boldsymbol{C}^{-1}\boldsymbol{Q}^{\beta}), \tag{4}$$

$\widehat{q}$ is the unnormalized power spectrum estimate

$$\widehat{q}^{\alpha} = \frac{1}{2} \boldsymbol{x}^{\dagger}\boldsymbol{C}^{-1}\boldsymbol{Q}^{\alpha}\boldsymbol{C}^{-1}\boldsymbol{x}, \tag{5}$$

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

$$\widehat{b}^{\alpha} = \frac{1}{2} \mathrm{tr}(\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{Q}^{\alpha}\boldsymbol{C}^{-1}). \tag{6}$$

The power spectrum estimator in Equation (3) is the minimum-variance (smallest error bar) estimate of the power spectrum subject to the constraint that it is also unbiased; that is, the ensemble average of the estimator is equal to its true value,

$$\langle \widehat{P}^{\alpha} \rangle = P^{\alpha} \tag{7}$$

(Tegmark 1997; Bond et al. 1998).

Intuitively, the estimator must be capable of "suppressing" or "removing" the effects of contaminants in order to obtain an unbiased estimate of the power spectrum. By construction, the subtraction of the residual foreground and noise bias accomplishes this, removing any additive bias. However, the $\boldsymbol{C}^{-1}$ piece of Equation (5) also has the effect of suppressing residual foregrounds and noise in both the additive bias and any contributions the residuals may have to the variance.

More specifically, the effect of the weighting in Equation (5) is to project out the modes of $\boldsymbol{U}$ with a different covariance structure than $\boldsymbol{S}$ in the power spectrum estimate, and the effect of Equation (6) is to subtract out the remaining bias. Similar effects for a realistic model of the EoR and foregrounds are shown in Liu & Tegmark (2011).

If the covariance structure of the contaminants is sufficiently different from the desired power spectrum, then the linear bias term may be expected to be quite small, and it is only necessary to know $\boldsymbol{C}$ and $\boldsymbol{Q}^{\alpha}$, not $\boldsymbol{U}$. Since the foregrounds are expected to be strongly correlated between frequencies, whereas the EoR is not, we expect different covariance structures and therefore a small linear bias. Moreover, because the linear bias is always positive and there is no multiplicative bias, the quadratic-only term will always produce an estimate that is high relative to the true value and can conservatively be interpreted as an upper limit. These considerations, and the difficulty of obtaining an estimate for $\boldsymbol{U}$, motivate the neglect of the linear bias in the rest of this analysis.

Motivated by the desire to retain the advantageous behavior of suppressing contributions of $\boldsymbol{U}$ to estimates of the EoR power spectrum, we note that is possible to define a modified version of the QE where Equation (5) is replaced by

$$\widehat{q}^{\alpha} = \frac{1}{2} \boldsymbol{x}^{\dagger}\boldsymbol{R}\boldsymbol{Q}^{\alpha}\boldsymbol{R}\boldsymbol{x}, \tag{8}$$

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

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

$$\widehat{\boldsymbol{P}} = \boldsymbol{M}\widehat{\boldsymbol{q}} \tag{9}$$

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

To ensure that $\boldsymbol{M}$ correctly normalizes our power spectrum, one may take the expectation value of Equation (9) to obtain

$$\begin{aligned} \langle \widehat{P}^{\alpha} \rangle &= \frac{1}{2} \sum_{\beta\gamma} M^{\alpha\gamma} \mathrm{tr}(\boldsymbol{R}\boldsymbol{Q}^{\gamma}\boldsymbol{R}\boldsymbol{Q}^{\beta}) P^{\beta} + \frac{1}{2} \sum_{\gamma} \mathrm{tr}(\boldsymbol{U}\boldsymbol{R}\boldsymbol{Q}^{\gamma}\boldsymbol{R}) \\ &\equiv \sum_{\beta} W^{\alpha\beta} P^{\beta} + \frac{1}{2} \sum_{\gamma} \mathrm{tr}(\boldsymbol{U}\boldsymbol{R}\boldsymbol{Q}^{\gamma}\boldsymbol{R}), \end{aligned} \tag{10}$$

where $W^{\alpha\beta}$ are elements of a window function matrix. Considering the first term of this expression (again, we are assuming that the linear bias term is significantly suppressed; if this is not the case, we are simply assuming that we are setting a conservative upper limit), if $\boldsymbol{W}$ ends up being the identity matrix for our choices of $\boldsymbol{R}$ and $\boldsymbol{M}$, then we recover Equation (7) for the first term, and we have an estimator that has no multiplicative matrix bias. However, Equation (7) is a rather restrictive condition, and it is possible to violate it and still have a sensible (and correctly normalized) power spectrum estimate. In particular, as long as the rows of $\boldsymbol{W}$ sum to unity, our power spectrum will be correctly normalized. Beyond this, the data analyst has a choice for $\boldsymbol{M}$, and for simplicity

throughout this paper, we choose $\boldsymbol{M}$ to be diagonal. In a preview of what is to come, we also stress that the derivation that leads to Equation (10) assumes that $\boldsymbol{R}$ and $\boldsymbol{x}$ are not correlated. If this assumption is violated, a simple application of the (now incorrect) formulae in this section can result in an improperly normalized power spectrum estimator that does not conserve power, i.e., one that has signal loss.

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

$$\widehat{\boldsymbol{C}} \equiv \langle \boldsymbol{x}\boldsymbol{x}^{\dagger} \rangle_t \approx \langle \boldsymbol{x}\boldsymbol{x}^{\dagger} \rangle, \tag{11}$$

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

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

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

1. If the covariance matrix is estimated from the data, a strong correlation between the estimated modes and the data will in general produce an estimate of the signal power spectrum that is strongly biased low relative to the true value. In this context, this is what we call "signal loss" (Section 2.2).
2. The effect of the bias is worsened when the number of independent samples used to estimate the covariance matrix is reduced (Section 2.3).
3. The rate at which empirical eigenvectors converge to their true forms depends on the sample variance in the empirical estimate and the shape of the empirical eigenspectrum. In general, larger sample variances lead to more loss (Section 2.3).
4. Knowing these things, there are some simple ways of altering the empirical covariance matrix to decouple it from the data and produce unbiased power spectrum estimates (Section 2.4).

2.2. Empirical Inverse Covariance Weighting

Using a toy model, we will now build intuition into how weighting by the inverse of the empirically estimated covariance, $\widehat{\boldsymbol{C}}^{-1}$, can give rise to signal loss. We construct a simple data set that contains visibility data with 100 time integrations and 20 frequency channels. This model represents

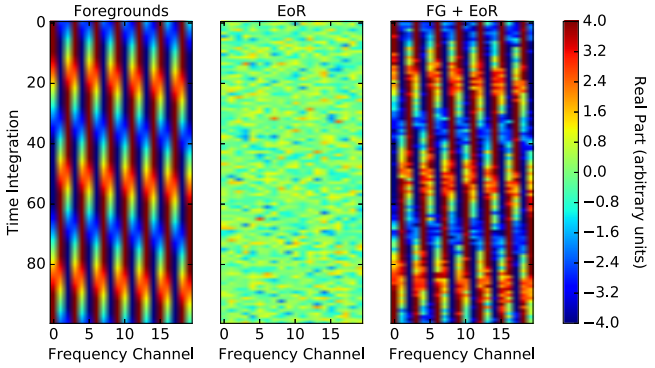

Figure 1. Our toy model data set to which we apply different weighting schemes in order to investigate signal loss. We model a mock foreground-only visibility with a sinusoid signal that varies smoothly in time and frequency. We model a mock visibility of an EoR signal as a random Gaussian signal. We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

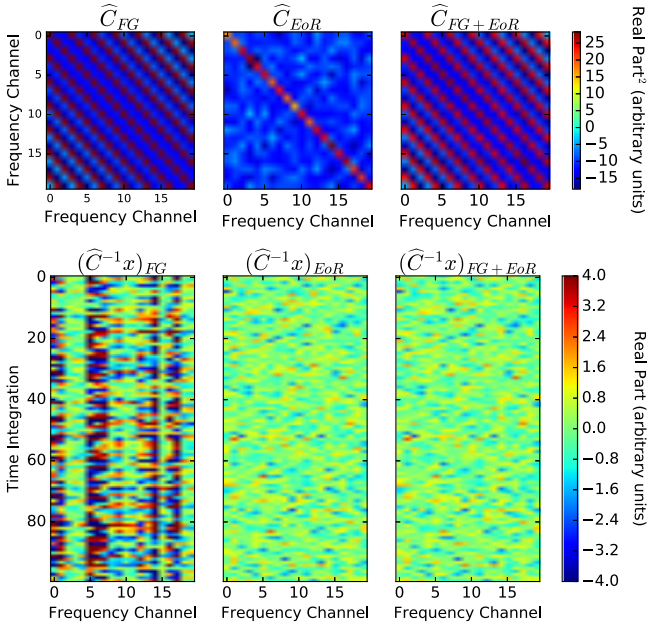

Figure 2. Estimated covariance matrices (top row) and inverse covariance-weighted data (bottom row) for FG only (left), EoR only (middle), and FG + EoR (right). Real parts are shown here.

realistic dimensions of about 1 hr of PAPER data that might be used for a power spectrum analysis. For PAPER-64 (both the A15 analysis and our new analysis), we use ∼8 hr of data (with channel widths of 0.5 MHz and integration times of 43 s), but here we scale it down with no loss of generality.

We create mock visibilities, $\boldsymbol{x}$, and assume a nontracking, drift-scan observation. Hence, flat-spectrum sources (away from zenith) lead to measured visibilities that oscillate in time and frequency. We therefore form a mock visibility measurement of a bright foreground signal, $\boldsymbol{x}_{\rm FG}$, as a complex sinusoid that varies smoothly in time and frequency, a simplistic but realistic representation of a single bright source. We also create a mock visibility measurement of an EoR signal, $\boldsymbol{x}_{\rm EoR}$, as a complex, Gaussian random signal. A more realistic EoR signal would have a sloped power spectrum in $p(k)$ (instead of flat, as in the case of white noise), which could be simulated by introducing frequency correlations into the mock EoR signal. However, here we treat all k-values separately, so a simplistic

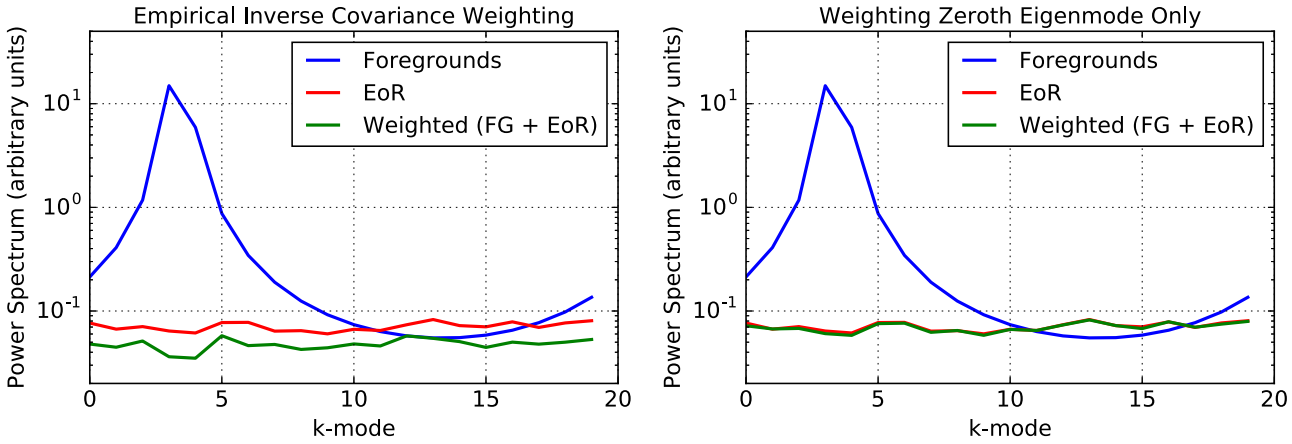

Figure 3. Resulting power spectrum estimates for the toy model simulation described in Section 2.2—foregrounds only (blue), EoR only (red), and weighted FG + EoR data set (green). The power spectrum of the foregrounds peaks at a k-mode based on the frequency of the sinusoid used to create the mock foreground signal. In the two panels, we compare using empirically estimated inverse covariance weighting, where $\boldsymbol{C}$ is derived from the data (left) and projecting out the zeroth eigenmode only (right). In the former case, signal loss arises from the coupling of the eigenmodes of $\widehat{\boldsymbol{C}}$ to the data. There is negligible signal loss when all eigenmodes besides the foreground one are no longer correlated with the data.

white noise approximation can be used. Our combined data vector is then $\boldsymbol{x} = \boldsymbol{x}_{\rm FG} + \boldsymbol{x}_{\rm EoR}$, to which we apply different weighting schemes throughout Section 2. The three data components are shown in Figure 1.

We compute the power spectrum of our toy model data set $\boldsymbol{x}$ using Equations (8) and (9), with $\boldsymbol{R} \equiv \widehat{\boldsymbol{C}}^{-1}$. Figure 2 shows the estimated covariances of our toy model data sets, along with the $\widehat{\boldsymbol{C}}^{-1}$ weighted data. The foreground sinusoid is clearly visible in $\widehat{\boldsymbol{C}}_{\rm FG}$. The power spectrum result is shown in green in the left plot of Figure 3. Also plotted in the figure are the uniform-weighted ($\boldsymbol{R} \equiv \boldsymbol{I}$) power spectra of the individual components $\boldsymbol{x}_{\rm FG}$ (blue) and $\boldsymbol{x}_{\rm EoR}$ (red). As shown, our $\widehat{\boldsymbol{C}}^{-1}$ weighted result successfully suppresses foregrounds, demonstrated in Figure 3 by the missing foreground peak in the weighted power spectrum estimate (green). It is also evident that our result fails to recover the EoR signal—it exhibits the correct shape, but the amplitude level is slightly low. It is this behavior that we describe as signal loss.

As discussed in Section 2.1, this behavior is not expected in the case where we use a true $\boldsymbol{C}^{-1}$ weighting. Rather, we would obtain a nearly unbiased estimate of the power spectrum. The key difference is that since $\widehat{\boldsymbol{C}}$ is estimated from the data, its eigenvectors and eigenvalues are strongly coupled to the particular data realization that was used to compute it, and this coupling leads to loss.

For the case of an eigenmode that can be safely assumed to be predominantly a foreground, its presence in the true covariance matrix will result in the desired suppression via a kind of projection; whether or not it is strongly correlated with the actual data vector is irrelevant. However, in the case of an empirically estimated covariance matrix, the eigenmodes of $\widehat{\boldsymbol{C}}_{\rm EoR}$ will both be incorrect and can be correlated with the data. If these incorrect eigenmodes are not correlated with the data, it will lead to nonminimum variance estimates but will not produce the suppression of the power spectrum amplitude as seen in the left plot of Figure 3. As described in Section 3.1, however, if $\widehat{\boldsymbol{C}}_{\rm EoR}$ is correlated with the data vector $\boldsymbol{x}$, there is a kind of projection of power in the nonforeground modes from the resulting power spectrum estimate, thus producing an estimate that is biased low. In short, if the covariance is computed from the data itself, it carries the risk of overfitting information in the data and introducing a multiplicative bias (per k) to estimates of the signal.

The danger of an empirically estimated covariance matrix comes mostly from not being able to describe the EoR-dominated eigenmodes of $\boldsymbol{C}$ accurately, for which the EoR signal is brighter than foregrounds. In such a case, the coupling between these modes to the data realization leads to the overfitting and subtraction of the EoR signal. More specifically, the coupling between the estimated covariance and the data is anticorrelated in nature (explained in more detail in Section 3.1), which leads to loss. Misestimating $\boldsymbol{C}$ for EoR-dominated eigenmodes is therefore more harmful than for foreground-dominated modes, and since the lowest-valued eigenmodes of an eigenspectrum are typically EoR-dominated, using this part of the spectrum for weighting is most dangerous.

Armed with this information, we can tweak the covariance in a simple way to suppress foregrounds and yield minimal signal loss. Recall that our toy model foreground can be perfectly described by a single eigenmode. Using the full data set's (foreground plus EoR signal) empirical covariance, we can project out the zeroth eigenmode and then take the remaining covariance to be the identity matrix. This decouples the covariance from the data for the EoR modes. The resulting power spectrum estimate for this case is shown in the right plot of Figure 3. In this case, we recover the EoR signal, demonstrating that if we can disentangle the foreground-dominated modes and EoR-dominated modes, we can suppress foregrounds with negligible signal loss.

Altering $\widehat{\boldsymbol{C}}$ as such is one specific example of a regularization method for this toy model, in which we are changing $\widehat{\boldsymbol{C}}$ in a way that reduces its coupling to the data realization. There are several other simple ways to regularize $\widehat{\boldsymbol{C}}$, and we will discuss some in Section 2.4.

2.3. Fringe-rate Filtering

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

A fringe-rate filter is an analysis technique designed to maximize sensitivity by integrating in time (Parsons et al. 2016).

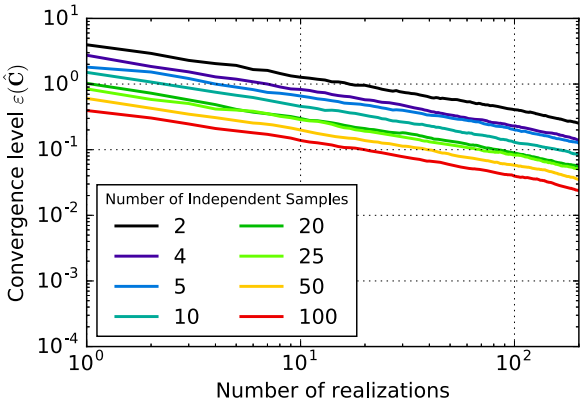

Figure 4. Convergence level, as defined by Equation (12), of empirically estimated covariances of mock EoR signals with different numbers of independent samples. In red, the mock EoR signal is comprised entirely of independent samples (100 of them). Subsequent colors show time-averaged signals. As the number of realizations increases, we see that the empirical covariances approach the true covariances. With more independent samples, the quicker an empirical covariance converges (i.e., the quicker it decouples from the data), the less signal loss we would expect to result.

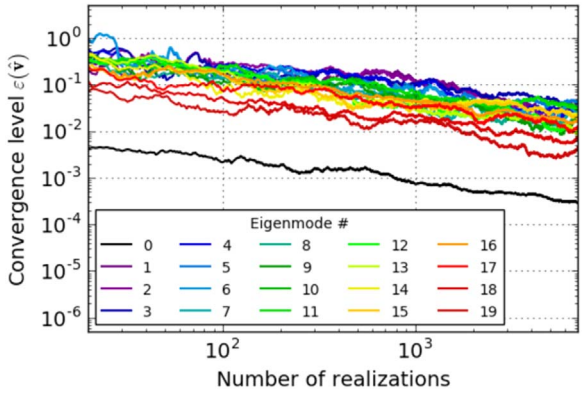

Figure 5. Convergence level, as defined by Equation (13), of empirically estimated eigenvectors for different numbers of mock data realizations. The colors range from the zeroth (highest eigenvalue) to 19th (lowest eigenvalue) eigenmode, where they are ordered by eigenvalue in descending order. This figure shows that the zeroth eigenmode converges the quickest, implying that eigenvectors with eigenvalues that are substantially different than the rest (the foreground-dominated mode has a much higher eigenvalue than the EoR modes) are able to converge to the true eigenvectors the quickest. On the other hand, eigenmodes 1–19 have similar eigenvalues and are slower to converge because of degeneracies between them.

Rather than a traditional boxcar average in time, a time-domain filter can be designed to up-weight temporal modes consistent with the sidereal motion on the sky while down-weighting modes that are noise-like.

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

$$\varepsilon(\widehat{\boldsymbol{C}}) \equiv \sqrt{\frac{\sum_{ij}(\widehat{C}_{ij} - C_{ij})^2}{\sum_{ij} C_{ij}^2}}, \tag{12}$$

where $\boldsymbol{C}$ is the true covariance matrix. To compute this metric, we draw different numbers of realizations (different draws of Gaussian noise) of our toy model EoR measurement, $\boldsymbol{x}_{\rm EoR}$, and take their ensemble average. We then compare this to the "true" covariance, which in our simulation is set to be the empirical covariance after a large number (500) of realizations. As shown in Figure 4, we perform this computation for a range of total independent ensemble realizations (horizontal axis) and number of independent samples in the data following time-averaging, or "fringe-rate filtering" (different colors). With more independent time samples (i.e., more realizations) in the data, one converges to the true fringe-rate filtered covariance more quickly.

The situation here with using a finite number of time samples to estimate our covariance is analogous to a problem faced in galaxy surveys, where the nonlinear covariance of the matter power spectrum is estimated using a large—but finite—number of expensive simulations. There, the limited number of independent simulations results in inaccuracies in estimated covariance matrices (Dodelson & Schneider 2013; Taylor & Joachimi 2014), which in turn result in biases in the final parameter constraints (Hartlap et al. 2007). In our case, the empirically estimated covariances are used for estimating the power spectrum, and as we discussed in the previous section (and will argue more thoroughly in Section 3.1), couplings between these covariances and the data can lead to power spectrum estimates that are biased low—which is precisely signal loss. In future work, it will be fruitful to investigate whether advanced techniques from the galaxy survey literature for estimating accurate covariance matrices can be successfully adapted for 21 cm cosmology. These techniques include the imposition of sparsity priors (Padmanabhan et al. 2016), the fitting of theoretically motivated parametric forms (Pearson & Samushia 2016), covariance tapering (Paz & Sánchez 2015), marginalization over the true covariance (Sellentin & Heavens 2016), and shrinkage methods (Pope & Szapudi 2008; Joachimi 2017).

The overall convergence of the covariance is important, but also noteworthy is the fact that different eigenvectors converge to their true forms at different rates. This is illustrated by Figure 5, which shows the convergence of eigenvectors in an empirical estimate of a covariance matrix. For this particular toy model, we construct a covariance whose true form combines the same mock foreground from the previous toy models with an EoR component that is modeled as a diagonal matrix with eigenvalues spanning one order of magnitude (more specifically, we construct the EoR covariance as a diagonal matrix in the Fourier domain, where the signal is expected to be uncorrelated; its Fourier transform is then the true covariance of the EoR in the frequency domain, or $\boldsymbol{C}_{\rm EoR}$). For different numbers of realizations, we draw random EoR signals that are consistent with $\boldsymbol{C}_{\rm EoR}$, add them to the mock foreground data, and compute the combined empirical covariance by averaging over the realizations. The eigenvectors of this empirical covariance are then compared to the true eigenvectors $\hat{\boldsymbol{v}}$, where we use as a convergence metric $\varepsilon(\hat{\boldsymbol{v}})$, defined as

$$\varepsilon(\hat{\boldsymbol{v}}) \equiv \sqrt{\sum_{i}^{N_f} |\boldsymbol{v} - \hat{\boldsymbol{v}}|_i^2}, \tag{13}$$

where N_f is the number of frequencies (Equation (20)) in the mock data. The eigenmode convergence curves in Figure 5 are

ranked by eigenvalue, such that "eigenmode 0" illustrates the convergence of the eigenvector with the largest eigenvalue, "eigenmode 1" for the second-largest eigenvalue, and so on. We see that the zeroth eigenmode—the mode describing the foreground signal—is quickest to converge.

Our numerical test reveals that the convergence rates of empirical eigenvectors are related to the sample variance in our empirical estimate. In general, computing an empirical covariance from a finite ensemble average means that the empirical eigenmodes have sample variances. Consider first a limiting case where all eigenvalues are equal. In such a scenario, any linear combination of eigenvectors is also an eigenvector, and thus there is no sensible way to define the convergence of eigenvectors. In our current test, aside from the zeroth mode, the eigenvalues have similar values but are not precisely equal. Hence, there is a well-defined set of eigenvectors to converge to. However, due to the sample variance of our empirical covariance estimate, there may be accidental degeneracies between modes, where some modes are mixing and swapping with others. Therefore, the steeper an eigenspectrum, the easier it is for the eigenmodes to decouple from each other and approach their true forms. A particularly drastic example of this can be seen in the behavior of mode 0 (the foreground mode), whose eigenvalue differs enough from the others that it is able to converge reasonably quickly despite substantial sample variance in our empirical covariance estimate. To break degeneracies in the remaining modes, however, requires many more realizations.

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

In the toy models throughout Section 2, we exploit the fact that the strongest eigenmode (highest-eigenvalue mode) is dominated by foregrounds in order to purposely incur signal loss for that mode. Even for the case of real PAPER data (Section 3), we make the assumption that the strongest eigenmodes are likely the most contaminated by foregrounds. However, in general, foregrounds need not be restricted to the strongest eigenmodes, and as we have seen, it is really the degeneracies between modes that determine how quickly they converge, and hence how much signal loss can result.

With Figures 4 and 5 establishing the connection between convergence rates (of empirical covariances and eigenvectors) and number of realizations, we now turn back to our original toy model used in Section 2.2, which is comprised of a mock foreground and mock EoR signal. We mimic a fringe-rate filter by averaging every four time integrations of our toy model data set together, yielding 25 independent samples in time (Figure 6). We choose these numbers so that the total number of independent samples is similar to the number of frequency channels; hence, our matrices will be full rank. We use this "fringe-rate filtered" mock data for the remainder of Section 2.

The power spectrum results for this model are shown in Figure 7, and as expected, there is a much larger amount of signal loss for this time-averaged data set, since we do a worse job estimating the true covariance. In addition, as a result of having fewer independent samples, we obtain an estimate with more scatter. This is evident by noticing that the green curve in Figure 7 fails to trace the shape of the uniform-weighted EoR power spectrum (red).

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

2.4. Other Weighting Options

In Section 2.2, we showed one example of how altering $\widehat{\boldsymbol{C}}$ can make the difference between nearly zero and some signal loss. We will now use our toy model to describe several other ways to tailor $\widehat{\boldsymbol{C}}$ in order to minimize signal loss. We choose four independent regularization methods to highlight in this section that have been chosen due to their simplicity in implementation and straightforward interpretation. We illustrate the resulting power spectra for the different cases in Figure 8. These examples are not meant to be taken as suggested analysis methods but rather as illustrative cases.

As a first test, we model the covariance matrix of EoR as a proof of concept that if perfect models are known, signal loss can be avoided. We know that our simulated EoR signal should have a covariance matrix that mimics the identity matrix, with its variance encoded along the diagonal. We model $\boldsymbol{C}_{\rm EoR}$ as such (i.e., the identity), instead of computing it based on $\boldsymbol{x}_{\rm EoR}$ itself. Next, we add $\boldsymbol{C}_{\rm EoR} + \widehat{\boldsymbol{C}}_{\rm FG}$ (where $\widehat{\boldsymbol{C}}_{\rm FG} = \langle \boldsymbol{x}_{\rm FG}\boldsymbol{x}_{\rm FG}^{\dagger}\rangle_t$) to obtain a final $\widehat{\boldsymbol{C}}_{\rm reg}$ (regularized empirical covariance matrix) to use in weighting. In Figure 8 (upper left), we see that there is negligible signal loss. This is because, by modeling $\boldsymbol{C}_{\rm EoR}$, we avoid overfitting EoR fluctuations in the data that our model does not know about (but an empirically derived $\widehat{\boldsymbol{C}}_{\rm EoR}$ would know about the fluctuations). In practice, such a weighting option is not feasible, as it is difficult to model $\boldsymbol{C}_{\rm EoR}$, and $\widehat{\boldsymbol{C}}_{\rm FG}$ is unknown because we do not know how to separate out the foregrounds from the EoR in our data.

The upper right panel in Figure 8 uses a regularization method of setting $\widehat{\boldsymbol{C}}_{\rm reg} \equiv \widehat{\boldsymbol{C}} + \gamma \boldsymbol{I}$, where $\gamma = 5$ (an arbitrary strength of $\boldsymbol{I}$ for the purpose of this toy model). By adding the identity matrix, element-wise, we are weighting the diagonal elements of the estimated covariance matrix more heavily than those off diagonal. Since the identity component does not know anything about the data realization, it alters the covariance to be less coupled to the data, and there is no loss.

The lower left panel in Figure 8 minimizes signal loss by only using the first three eigenmodes of the estimated covariance. Recalling that our toy model foregrounds can be described entirely by the zeroth eigenmode, this method intentionally projects out the highest-valued modes only by replacing all but the three highest weights in the eigenspectrum with 1s (equal weights). Again, avoiding the overfitting of EoR-dominated modes that are coupled to the data results in negligible signal loss. While this case is illuminating for the toy model, in practice, it is not obvious which eigenmodes are foreground or EoR-dominated (and they could be mixed as

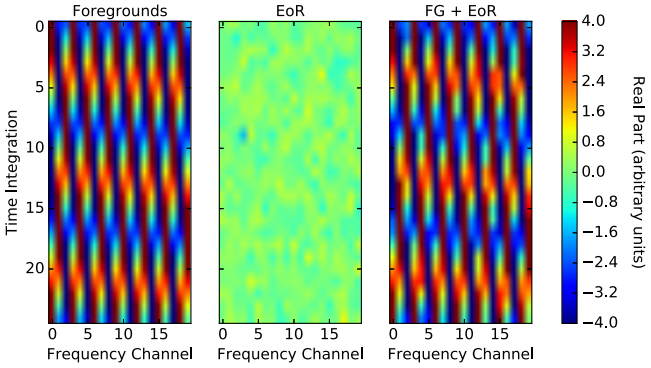

Figure 6. Our "fringe-rate filtered" (time-averaged) toy model data set. We average every four samples together, yielding 25 independent samples in time. Real parts are shown here.

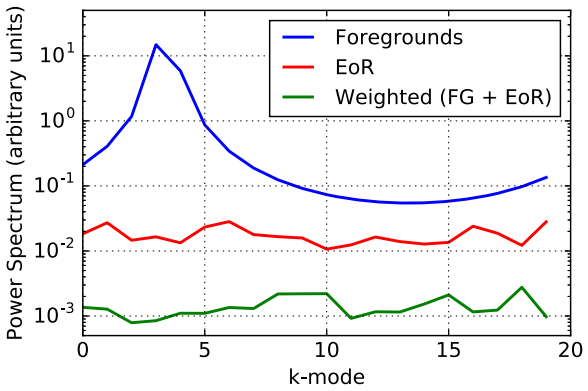

Figure 7. Resulting power spectrum estimate for the "fringe-rate filtered" (time-averaged) toy model simulation—foregrounds only (blue), EoR only (red), and weighted FG + EoR data set (green). We use empirically estimated inverse covariance weighting, where $\boldsymbol{C}$ is computed from the data. There is a larger amount of signal loss than for the nonaveraged data, a consequence of weighting by eigenmodes that are more strongly coupled to the data due to there being fewer independent modes in the data.

well), so determining which subset of modes to down-weight is not trivial. We experiment with this idea using PAPER data in Section 3.3.

The last regularization scheme we are highlighting here is setting $\widehat{\boldsymbol{C}}_{\rm reg} \equiv \widehat{\boldsymbol{C}} \circ \boldsymbol{I}$ (element-wise multiplication), or inverse variance weighting (i.e., keeping only the diagonal elements of $\widehat{\boldsymbol{C}}$). In the lower right panel of Figure 8, we see that this method does not down-weight the foregrounds at all—this regularization altered $\widehat{\boldsymbol{C}}$ in a way where it is no longer coupled to any of the empirically estimated eigenmodes, including the foreground-dominated one. To understand this, we recall that our foregrounds are spread out in frequency and therefore have nonnegligible frequency–frequency correlations. Multiplying by the identity matrix element-wise results in a diagonal matrix, meaning we do not have any correlation information. Because of this, we do a poor job suppressing the foreground. But because we decoupled the whole eigenspectrum from the data, we also avoid signal loss. Although this method did not successfully recover the EoR signal for this particular simulation, it is important that we show that there are many options for estimating a covariance matrix, and some may down-weight certain eigenmodes more effectively than others based on the spectral nature of the components in a data set.

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

3. Signal Loss in PAPER-64

We now turn to a detailed signal loss investigation using a subset of the PAPER-64 data set from A15. In the previous section, we showed how signal loss arises when weighting data with empirically estimated covariances; in this section, we highlight how the amount of this loss was underestimated in the previous analysis. Additionally, we illustrate how we have revised our analysis pipeline in light of our growing understanding.

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 9). All data are compressed, calibrated (using self-calibration and redundant calibration), delay-filtered (to remove foregrounds inside the wedge), binned in local sidereal time (LST), and fringe-rate filtered. For detailed information about the back-end system of PAPER-64 and its observations and data reduction pipeline, we refer the reader to Parsons et al. (2010) and A15. We note that all data-processing steps are identical to those in A15 until after the LST-binning step in Figure 3 of A15.

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

$$\Delta^2(k) = \frac{k^3}{2\pi^2}\,\hat{P}(k), \tag{14}$$

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

For the analysis in this paper, we use 8.1 hr of LST, namely an R.A. range of 0.5–8.6 hr (A15 uses a slightly longer R.A. range of 0–8.6 hr; we found that some early LSTs were more severely foreground-contaminated). We also use only 10 baselines, a subset of the 51 total East/West baselines used in A15, in order to illustrate our revised methods. All power spectrum results are produced for a center frequency of 151 MHz using a width of 10 MHz (20 channels), identical to the analysis in A15. In the case study in this paper, we only use one baseline type instead of the three in A15, but M. Kolopanis et al. (2018, in preparation) use the full data set presented in A15 to revise the result and place limits on the EoR at multiple redshifts (using a straightforward and not lossy approach to avoid many of the issues that will be made clear later on).

The most significant changes from A15 occur in our revised power spectrum analysis, which is explained in the rest of this paper, but we also note that the applied fringe-rate filter is also slightly different. In A15, the applied filter was not equivalent to the optimal fringe-rate filter (which is designed to maximize power spectrum sensitivity). Instead, the optimal filter was degraded slightly by widening it in fringe-rate space. This was chosen in order to increase the number of independent modes and reduce signal loss associated with the QE, though as we

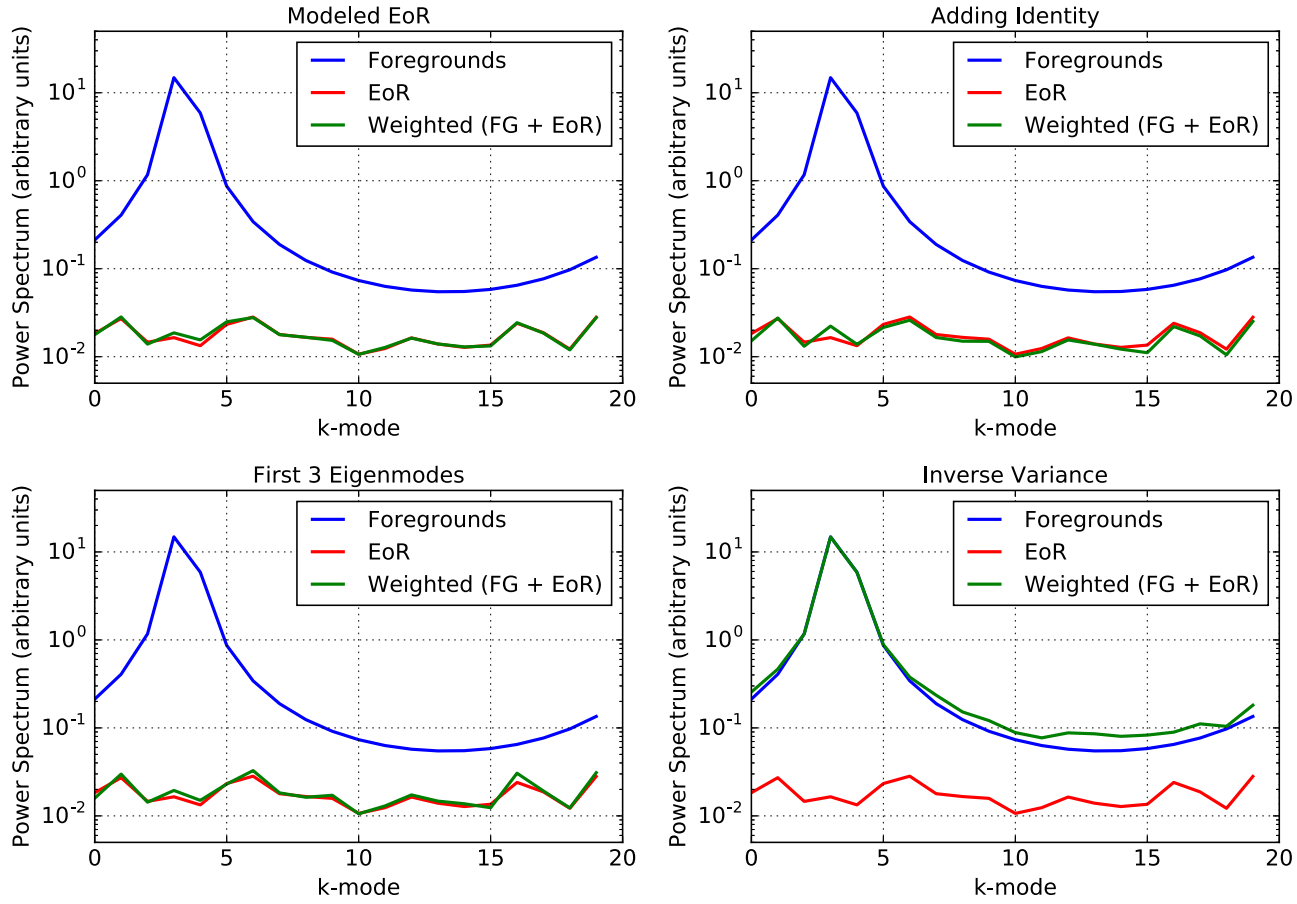

Figure 8. Resulting power spectrum estimates for our "fringe-rate filtered" (time-averaged) toy model simulation—foregrounds only (blue), EoR only (red), and weighted FG + EoR data set (green). We show four alternate weighting options that each minimize signal loss, including modeling the covariance matrix of EoR (upper left), regularizing $\widehat{C}$ by adding an identity matrix to it (upper right), using only the first three eigenmodes of $\widehat{C}$ (lower left), and keeping only the diagonal elements of $\widehat{C}$ (lower right). The first case (upper left) is not feasible in practice, since we do not know $\boldsymbol{C}_{\rm FG}$ and $\boldsymbol{C}_{\rm EoR}$ like we do in the toy model.

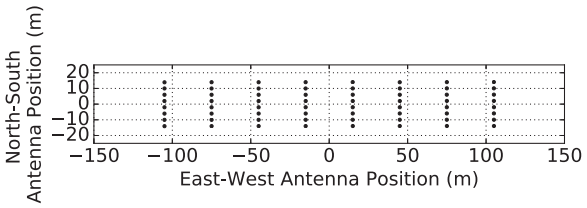

Figure 9. PAPER-64 antenna layout. We use only 10 of the 30 m East/West baselines for the analysis in this paper (i.e., a subset of the shortest horizontal spacings).

will explain in the next section, this signal loss was still underestimated. With the development of a new, robust method for assessing signal loss, we choose to use the optimal filter in order to maximize sensitivity. This filter is computed for a fiducial 30 m baseline at 150 MHz, the center frequency in our band. The filter in both the fringe-rate domain and time domain is shown in Figure 10.

Finally, we emphasize that the discussion that follows is solely focused on signal loss associated with empirical covariance weighting. As mentioned in Section 2, there are a number of steps in our analysis pipeline that could lead to loss, including gain calibration, delay filtering, and fringe-rate filtering, which have been investigated at various levels of detail in Parsons et al. (2014) and A15 but are clearly the subject of future work. Here we only focus on the most significant source of loss we have identified and note that M. Kolopanis et al. (2018, in preparation) and other future work will consider additional sources of signal loss and exercise increased caution in reporting results.

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

3.1. Signal Loss Methodology

In short, our method for estimating signal loss consists of adding an EoR-like signal into visibility data and then measuring how much of this injected signal would be detectable given any attenuation of this signal by the (lossy) data analysis pipeline. To capture the full statistical likelihood of signal loss, one requires a quick way to generate many

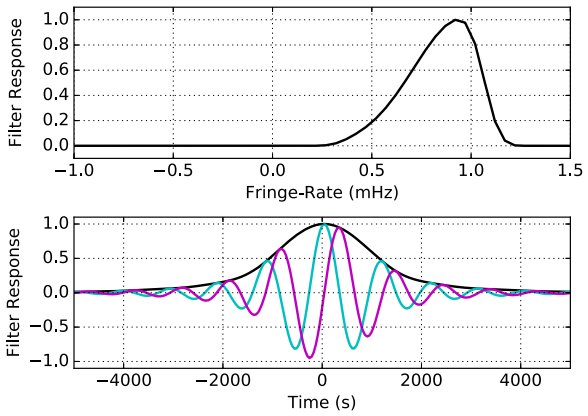

Figure 10. Top: normalized optimal power spectrum sensitivity weighting in fringe-rate space for our fiducial baseline and Stokes I polarization beam. Bottom: time-domain convolution kernel corresponding to the top panel. Real and imaginary components are illustrated in cyan and magenta, respectively, with the absolute amplitude in black. The fringe-rate filter acts as an integration in time, increasing sensitivity but reducing the number of independent samples in the data set.

realizations of simulated 21 cm signal visibilities. Here we use the same method as in A15, where mock Gaussian noise visibilities (mock EoR signals) are filtered in time using an optimal fringe-rate filter to retain only "sky-like" modes. Since the optimal filter has a shape that matches the rate of the sidereal motion of the sky, this transforms the Gaussian noise into a measurement that PAPER could make. This signal is then added to the visibility data.[17]

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

$$\boldsymbol{r} = \boldsymbol{x} + \boldsymbol{e}, \tag{15}$$

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

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

$$\widehat{P}^{\alpha}_{\rm in} \equiv M^{\alpha}_{\rm in}\boldsymbol{e}^{\dagger}\boldsymbol{I}\boldsymbol{Q}^{\alpha}\boldsymbol{I}\boldsymbol{e} \tag{16}$$

and

$$\begin{aligned}\widehat{P}^{\alpha}_{\rm out} &\equiv \widehat{\boldsymbol{P}}^{\alpha}_{r}\\ &= M^{\alpha}_{r}\boldsymbol{r}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{r},\end{aligned} \tag{17}$$

where, for illustrative purposes and notational simplicity, we have written these equations with scalar normalizations M, even though, for our numerical results, we choose a diagonal matrix normalization using $\boldsymbol{M}$ as in Equation (9).

The quantity $\widehat{P}_{\rm in}$, defined by Equation (16), is a uniformly weighted estimator of the power spectrum of $\boldsymbol{e}$. It can be considered the power spectrum of this particular realization of the EoR; alternatively, it can be viewed as the true power spectrum of the injected signal up to cosmic variance fluctuations. The role of $\widehat{P}_{\rm in}$ in our analysis is to serve as a reference for the power spectrum that would be measured if there were no signal loss or other systematics. The input power spectrum is then to be compared to $\widehat{P}_{\rm out}$, which approximates the (lossy) power spectrum estimate that is output by our analysis pipeline prior to any signal loss adjustments.

Under this injection framework, we can begin to see explicitly why there can be large signal loss. Expanding Equation (17), $\widehat{P}_{\rm out}$ becomes

$$\begin{aligned}\widehat{P}^{\alpha}_{\rm out} &= M^{\alpha}_{r}(\boldsymbol{x}+\boldsymbol{e})^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}(\boldsymbol{x}+\boldsymbol{e})\\ &= M^{\alpha}_{a}\boldsymbol{x}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{x} + M^{\alpha}_{b}\boldsymbol{e}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{e}\\ &\quad + M^{\alpha}_{c}\boldsymbol{x}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{e} + M^{\alpha}_{d}\boldsymbol{e}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{x}.\end{aligned} \tag{18}$$

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

$$\begin{aligned}\widehat{P}^{\alpha}_{\rm out} &= M^{\alpha}_{a}\boldsymbol{x}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{x} + M^{\alpha}_{b}\boldsymbol{e}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{e}\\ &\quad + 2M^{\alpha}_{c}\boldsymbol{x}^{\dagger}\boldsymbol{R}_{r}\boldsymbol{Q}^{\alpha}\boldsymbol{R}_{r}\boldsymbol{e}.\end{aligned} \tag{19}$$

One of the key takeaways of this section is that the A15 analysis estimated signal loss by comparing only the signal-only term (second term in Equation (19)) with $\widehat{P}_{\rm in}$, whereas in fact, the cross-term (third term in Equation (19)) can substantially lower $\widehat{P}_{\rm out}$. In order to investigate the effect of each of these terms on signal loss, all three components are plotted in Figure 11 for two cases: empirically estimated inverse covariance weighting ($\boldsymbol{R}_r \equiv \widehat{\boldsymbol{C}}_r^{-1}$) and uniform weighting ($\boldsymbol{R}_r \equiv \boldsymbol{I}$). We will now go into further detail and examine the behavior of this equation in three different regimes of the injected signal: very weak (left ends of the $P_{\rm in}$ axes in Figure 11), very strong (right ends), and in between (middle portions).

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

Large injection. When the injected signal is much larger than the measured power spectrum, the data-only components can be neglected, as they are many orders of magnitude smaller.

[17] One specific change from A15 is that we add this simulated signal—which has been fringe-rate filtered once already in order to transform it into a "sky-like" signal—into the analysis pipeline before a fringe-rate filter is applied to the data (i.e., prior to the analysis step of fringe-rate filtering). Previously, the addition was done after the fringe-rate filter analysis step. This change results in an increased estimate of signal loss, likely due to the use of the fringe-rate filter as a simulator. However, this pipeline difference, while significant, is not the dominant reason why signal loss was underestimated in A15 (the dominant reason is explained in the main text in Section 3.1).

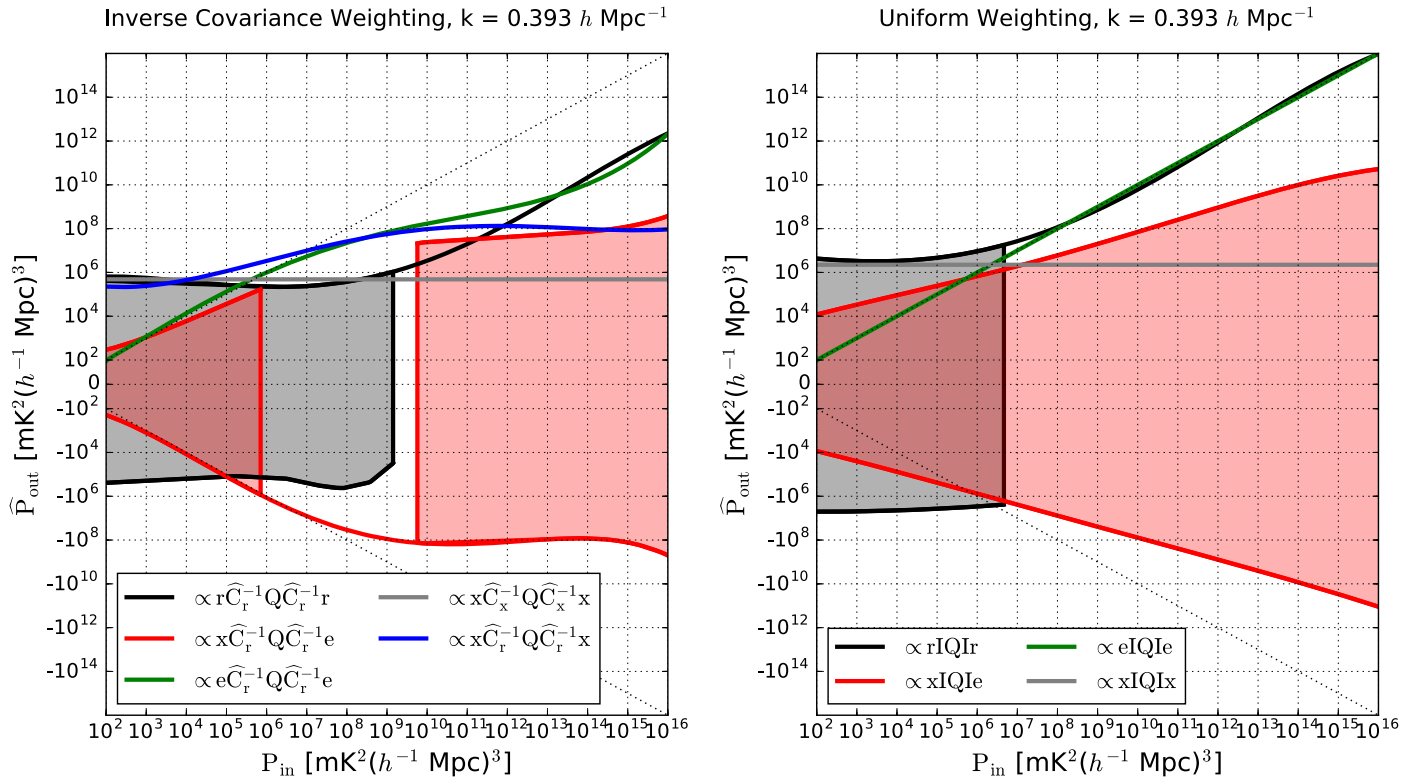

Figure 11. Illustration of the power spectrum amplitude of five different power spectrum terms, each a function of visibility data ($\boldsymbol{x}$), simulated injected EoR signal ($\boldsymbol{e}$), or both ($\boldsymbol{r}$). The figure shows how these quantities behave as the power level of the injected EoR signal increases (along the x-axis). The details of the simulation used to generate the figure are explained in Section 3.2; here we sample a larger $P_{\rm in}$ range and fit smooth polynomials to our data points to make an illustrative example. We emphasize that the output power spectrum in black ($\widehat{P}_{\rm out} = \widehat{P}_r$) approximates the (lossy) power spectrum estimate that is output by our analysis pipeline prior to any signal loss adjustments. Roughly speaking, it can be compared to the input signal level ($P_{\rm in}$) to estimate the amount of signal loss. In the left panel, empirical inverse covariance weighting is used in power spectrum estimation, as in A15. The dotted black diagonal line indicates perfect 1:1 input-to-output mapping (no signal loss). The gray horizontal line is the power spectrum value of data alone, $\widehat{P}_x$ (it does not depend on injected power). The green signal–signal component is the term used in A15 to estimate signal loss. It is significantly higher than $\widehat{P}_r$ (black) when the cross-terms (red) are large and negative (black = green + red + blue). In the regime where cross-correlations between signal and data are not dominant (small and large $P_{\rm in}$), the cross-terms have a noise-like term with width $\sqrt{\widehat{P}_e}\sqrt{\widehat{P}_x}$. However, at power levels comparable to the data (middle region), the cross-terms can produce large, negative estimates due to couplings between $\boldsymbol{x}$ and $\boldsymbol{e}$ that affect $\widehat{\boldsymbol{C}}_r$. This causes the difference between the green curve (which exhibits negligible loss at the data-only power spectrum value) and the black curve (which exhibits ∼4 orders of magnitude of loss). Right: same power spectrum terms illustrated for the uniform-weighted case.

We include a description of this regime for completeness in our discussion but note that the upper limits that we compute are typically not determined by simulations in this regime (i.e., in using an empirical weighting scheme, we have assumed that the data are dominated by foregrounds rather than the cosmological signal). However, it is useful as a check of our system in a relatively simple case. As we can see from Figure 11, the cross-terms (red) are small in comparison to the signal-only term (green). Only here does the signal-only term used in A15 dominate the total power output. We again see that, in the empirical inverse covariance-weighted case, the cross-terms behave as noise (positive and negative fluctuations around zero mean). This is for the same reason as at small injections—here $\widehat{\boldsymbol{C}}_r$ is dominated by the signal $\boldsymbol{e}$. The cross-correlation can again be modeled as a symmetric distribution of zero mean and width $\sqrt{\widehat{P}_e}\sqrt{\widehat{P}_x}$.

In between. When the injected signal is of a similar amplitude to the data by itself, the situation becomes less straightforward. We see that the weighted injected power spectrum component mirrors the input power indicating little loss (i.e., the green curve follows the dotted black line), eventually departing from unity when the injected amplitude is well above the level of the data power spectrum. However, in this regime, the cross-term (red) has nearly the same amplitude but with a negative sign. As explained below, this negativity is the result of cross-correlating inverse covariance-weighted terms. This negative component drives down the $\widehat{P}_{\rm out}$ estimator (black). Again, we emphasize that in A15, signal loss was computed by only looking at the second term in Equation (19) (green), which incorrectly implies no loss at the data-only power spectrum level. Ignoring the effect of the negative power from the cross-terms is the main reason for underestimating power spectrum limits in A15.

The source of the strong negative cross-term is not immediately obvious; however, it is an explainable effect. When $\boldsymbol{R}_r$ is taken to be $\widehat{\boldsymbol{C}}_r^{-1}$, the third term of Equation (19) is a cross-correlation between $\widehat{\boldsymbol{C}}_r^{-1}\boldsymbol{x}$ and $\widehat{\boldsymbol{C}}_r^{-1}\boldsymbol{e}$. As shown in Switzer et al. (2015), this cross-correlation term is nonzero and, in fact, negative in expectation. This negative cross-term power arises from a coupling between the inverse of $\widehat{\boldsymbol{C}}_r$ and $\boldsymbol{x}$. Intuitively, we can see this by expanding the empirical covariance of $\boldsymbol{r} = \boldsymbol{x} + \boldsymbol{e}$,

$$\begin{aligned}\widehat{\boldsymbol{C}}_r &= \langle \boldsymbol{r}\boldsymbol{r}^\dagger \rangle_t \\ &= \langle \boldsymbol{x}\boldsymbol{x}^\dagger \rangle_t + \langle \boldsymbol{x}\boldsymbol{e}^\dagger \rangle_t + \langle \boldsymbol{e}\boldsymbol{x}^\dagger \rangle_t + \langle \boldsymbol{e}\boldsymbol{e}^\dagger \rangle_t, \end{aligned} \quad (20)$$

where we can neglect the first term because $\boldsymbol{x}$ is small (i.e., the large negative cross-term power in the left panel of Figure 11 occurs when the injected amplitude surpasses the level of the data-only power spectrum). Without loss of generality, we will assume an eigenbasis of $\boldsymbol{e}$, so that $\langle \boldsymbol{e}\boldsymbol{e}^{\dagger}\rangle_t$ is diagonal. The middle two terms, however, can have power in their off-diagonal terms due to the fact that, when averaging over a finite ensemble, $\langle \boldsymbol{x}\boldsymbol{e}^{\dagger}\rangle_t$ is not zero. As shown in Appendix C of Parsons et al. (2014), to leading order, the inversion of a diagonal-dominant matrix like $\widehat{\boldsymbol{C}}_r$ (from $\langle \boldsymbol{e}\boldsymbol{e}^{\dagger}\rangle_t$) with smaller off-diagonal terms results in a new diagonal-dominant matrix with negative off-diagonal terms. These off-diagonal terms depend on both $\boldsymbol{x}$ and $\boldsymbol{e}$. Then, when $\widehat{\boldsymbol{C}}_r^{-1}$ is multiplied into $\boldsymbol{x}$, the result is a vector that is similar to $\boldsymbol{x}$ but contains a residual correlation to $\boldsymbol{e}$ from the off-diagonal components of $\widehat{\boldsymbol{C}}_r^{-1}$. The correlation is negative because the product $\widehat{\boldsymbol{C}}_r^{-1}\boldsymbol{x}$ effectively squares the $\boldsymbol{x}$-dependence of the off-diagonal terms in $\widehat{\boldsymbol{C}}_r^{-1}$ while retaining the negative sign that arose from the inversion of a diagonal-dominant matrix.

In general. Another way to phrase the shortcomings of the empirical inverse covariance estimator is that it is not properly normalized. Signal loss due to couplings between the data and their weightings arise because our unnormalized QE from Equation (8) ceases to be a quadratic quantity and instead contains higher-order powers of the data. However, the normalization matrix $\boldsymbol{M}$ is derived assuming that the unnormalized estimator is quadratic in the data. The power spectrum estimate will therefore be incorrectly normalized, which manifests as signal loss. We leave a full analytic solution for $\boldsymbol{M}$ for future work, since our simulations already capture the full phenomenology of signal loss and have the added benefit of being more easily generalizable in the face of non-Gaussian systematics.

3.2. Signal Loss in Practice

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

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

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

$$p(P_{\rm in}|\widehat{P}_{\rm out}) \propto \mathcal{L}(\widehat{P}_{\rm out}|P_{\rm in})\, p(P_{\rm in}), \tag{21}$$

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

The implementation details of the injection process require some more detailed explanation. In our code, we add a new realization of EoR to each independent bootstrap of data (see Section 4.1 for a description of PAPER's bootstrapping routine) with the goal of simultaneously capturing cosmic variance, noise variance, and signal loss. To limit computing time, we perform 20 realizations of each $P_{\rm in}$ level. We also run 50 total EoR injection levels, yielding $P_{\rm in}$ values that range from $\sim 10^5$ to $\sim 10^{11}$ mK^2 $(h^{-1}$ Mpc$)^3$, resulting in a total of 1000 data points on our $P_{\rm in}$ versus $\widehat{P}_{\rm out}$ grid.

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

Finally, our transfer functions are shown in Figure 12 for both the weighted (left) and unweighted (right) cases. Our bootstrapped power spectrum outputs are shown as black points, and the colored heat map overlaid on top is the likelihood function modified by our prior. Although we only show figures for one k-value, we note that the shape of the transfer curve is similar for all k-values. We then invoke Bayes's interpretation and reinterpret it as the posterior $p(P_{\rm in}|\widehat{P}_{\rm out})$, where we recall that $\widehat{P}_{\rm out}$ represents a (lossy) power spectrum. To do this, we make a horizontal cut at the data value $\widehat{\boldsymbol{P}}_x$ (setting $\widehat{P}_{\rm out} = \widehat{\boldsymbol{P}}_x$), shown by the gray solid line, to yield a posterior distribution for the signal. We normalize this final distribution and compute the 95% confidence interval (an upper limit on EoR).

By-eye inspection of the transfer function in Figure 12 gives a sense of what the signal loss result should be. The power spectrum value of our data, $\widehat{\boldsymbol{P}}_x$, is marked by the solid gray horizontal lines. From the left plot (empirically estimated inverse covariance weighting), one can eyeball that a data value of 10^5 mK^2 $(h^{-1}$ Mpc$)^3$, for example, would map approximately to an upper limit of $\sim 10^9$ mK^2 $(h^{-1}$ Mpc$)^3$, implying a signal loss factor of $\sim 10^4$.

The loss-corrected power spectrum limit for empirically estimated inverse covariance-weighted PAPER-64 data is shown in Figure 13 (solid red), which we can compare to the original lossy result (dashed red). Post–signal loss estimation, the power spectrum limits are higher than both the theoretical noise level (green) and uniform-weighted power spectrum

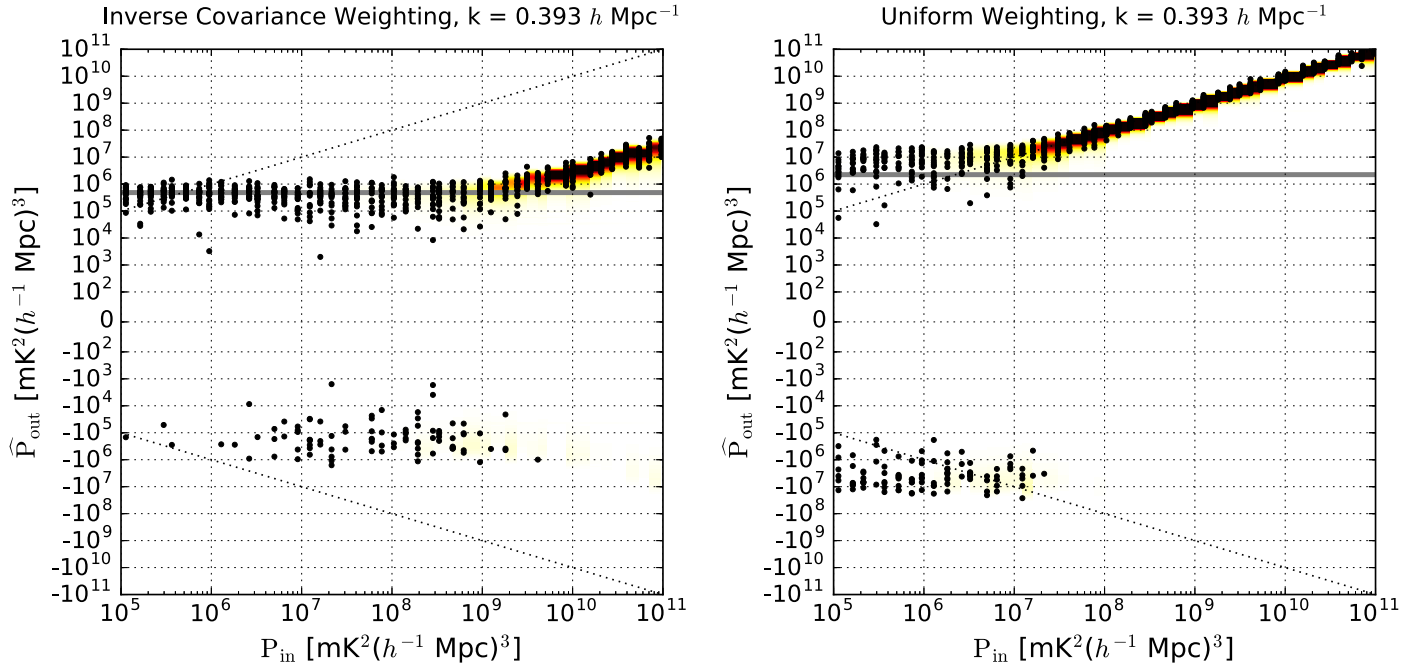

Figure 12. Signal loss transfer functions showing the relationship of P_{in} and $\widehat{P}_{\mathrm{out}}$, as defined by Equations (16) and (17). Power spectra values (black points) are generated for 20 realizations of $\boldsymbol{e}$ per signal injection level. Since our $\widehat{P}_{\mathrm{out}}$ values are well approximated by a Gaussian distribution, we fit Gaussians to each injection level based on the mean and variance of the simulation outputs. This entire likelihood function is then multiplied by a Jeffreys prior for $p(P_{\mathrm{in}})$, with the final result shown as the colored heat maps on top of the points. Two cases are displayed: empirically estimated inverse covariance-weighted PAPER-64 data (left) and uniform-weighted data (right). The dotted black diagonal lines mark a perfect unity mapping, and the solid gray horizontal line denotes the power spectrum value of the data $\widehat{\boldsymbol{P}}_x$ from which a posterior distribution for the signal is extracted. From these plots, it is clear that the weighted case results in ∼4 orders of magnitude of signal loss at the data-only power spectrum value, whereas the uniform-weighted case does not exhibit loss. The general shapes of these transfer functions are also shown by the black curves in Figure 11 for comparison.

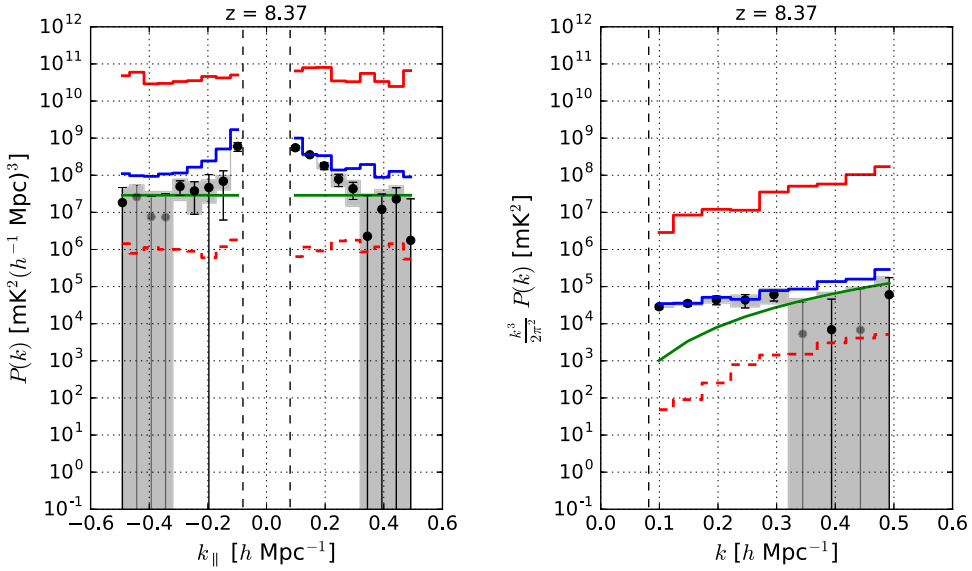

Figure 13. Power spectrum of a subset of PAPER-64 data illustrating the use of empirical inverse covariance weighting. The solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework using empirical inverse covariance weighting. Shown for comparison is the lossy limit prior to signal loss estimation (dashed red). The theoretical 2σ thermal noise level prediction based on observational parameters is shown in green, and its calculation is detailed in Section 4.2. Additionally, the power spectrum result for the uniform-weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), the 2σ upper limit on the EoR signal using our full signal injection framework (solid blue lines), and the measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. In this example, we see that the lossy power spectrum limit is ∼4 orders of magnitude too low when using empirical inverse covariance weighting.

(which is shown three ways: black and gray points are positive and negative power spectrum values, respectively, with 2σ error bars from bootstrapping; the solid blue line is the upper limit on the EoR signal using the full signal injection framework; and the shaded gray region shows the power spectrum values with thermal noise errors). We elaborate on this point in the next section and investigate alternate weighting schemes to inverse covariance weighting, with the goal of

finding one that balances the aggressiveness of down-weighting contaminants and minimizing the loss of the EoR signal.

3.3. Minimizing Signal Loss

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

We saw in Section 2.4 how limiting the number of down-weighted eigenmodes (i.e., flattening out part of the eigenspectrum and effectively decoupling the lowest-valued eigenmodes, which are typically EoR-dominated, from the data) can help minimize signal loss. We experiment with this idea on PAPER-64 data, dialing the number of modes that are down-weighted from zero (which is equivalent to identity-weighting, or the uniform-weighted case) to 21 (which is the full inverse covariance estimator). The power spectrum results for one k-value, both before and after signal loss estimation, are shown in the top panel of Figure 14. We see that the amount of signal loss increases as weighting becomes more aggressive (dashed red line). In other words, more EoR-dominated fluctuations are being overfit and subtracted as more modes are down-weighted. We also find that the power spectrum upper limit, post–signal loss estimation, increases with the number of down-weighted modes (solid red line). The more modes we use in down-weighting, the stronger the coupling between the weighting and the data and the greater the error we have in estimating the power spectrum. Switzer et al. (2013) took a similar approach in determining the optimal number of modes to down-weight in GBT data, finding similar trends and noting that removing too few modes is limited by residual foregrounds and removing too many modes is limited by large error bars and signal loss.

Optimistically, we expect there to be a "sweet spot" as we dial our regularization knob: a level of regularization where weighting is beneficial compared to uniform weighting (blue). In other words, we would like a weighting scheme that down-weights eigenmodes that predominantly describe foreground modes but not EoR modes. We see in Figure 14 that this occurs roughly when only the approximately three highest-valued eigenmodes are down-weighted and the rest are given equal weights (though for the case shown, weighting only slightly outperforms uniform weighting). For a similar discussion of projecting out modes (zeroing out eigenmodes, rather than just ignoring their relative weightings, as we do in this study), see Switzer et al. (2013).

We also saw in Section 2.4 how adding the identity matrix to the empirical covariance can minimize signal loss. We experiment with this idea as well, shown in the bottom panel of Figure 14. The dashed red and solid red lines represent power spectrum limits pre– and post–signal loss estimation, respectively, as a function of the strength of $\boldsymbol{I}$ that is added to $\widehat{\boldsymbol{C}}$, quantified as a percentage of $\mathrm{Tr}(\widehat{\boldsymbol{C}})\boldsymbol{I}$ added to $\widehat{\boldsymbol{C}}$. We parameterize this "regularization strength" parameter as γ, namely $\widehat{\boldsymbol{C}} \equiv \widehat{\boldsymbol{C}} + \gamma\,\mathrm{Tr}(\widehat{\boldsymbol{C}})\boldsymbol{I}$. From this plot, we see that only a small percentage of $\mathrm{Tr}(\widehat{\boldsymbol{C}})$ is needed to significantly reduce loss. We expect that as the strength of $\boldsymbol{I}$ is increased (going to the left), both of the red curves will approach the uniform-weighted case. We also notice that the post–signal loss limit hovers around the uniform-weighted limit for a large range of regularization strengths, and while an overall trend from high-to-low signal loss is seen as the strength increases, there does not appear to be a clear "minimum" that produces the least loss.

In addition to our thermal noise prediction (green) and uniform-weighted power spectrum limit (blue), one additional horizontal line is shown in Figure 14 in both panels and represents a third regularization technique. This line (black) denotes the power spectrum value, post–signal loss estimation, for inverse variance weighting (multiplying an identity matrix element-wise to $\widehat{\boldsymbol{C}}$). This result is single-valued and not a function of the horizontal axis. We see that all three regularization schemes shown (solid red top panel, solid red bottom panel, black) perform similarly at their best (i.e., when about three eigenmodes are down-weighted, in the case of the top panel's solid red curve). However, for the remainder of this paper, we choose to use the weighting option of $\widehat{\boldsymbol{C}} + 0.09\,\mathrm{Tr}(\widehat{\boldsymbol{C}})\boldsymbol{I}$, or $\gamma = 0.09$, which we will denote as $\widehat{\boldsymbol{C}}_{\mathrm{eff}}$. We choose this weighting scheme merely as a simple example of regularizing PAPER-64 covariances, noting that the power spectrum upper limit remains roughly constant for a broad range of values of γ.

It is important to note that our signal injection methodology for assessing loss makes the assumption that we know the true signal's strength and structure. Realistically, these details about the EoR signal are unknown, and our signal loss framework is limited by our simulations. Therefore, while this paper employs this methodology as an example of one way of estimating loss, M. Kolopanis et al. (2018, in preparation) use uniform weightings in order to produce more trustworthy, straightforward power spectrum limits that do not suffer from loss.

The power spectrum result for our subset of PAPER-64 data (using only one baseline separation type, 10 baselines, and $\widehat{\boldsymbol{C}}_{\mathrm{eff}}$) using the analysis presented in this paper is shown in Figure 15. Again, the solid red curve represents our upper limit on the EoR signal using the full signal injection framework. The uniform-weighted case is shown as the black and gray points, which correspond to positive and negative power spectrum values, respectively (with 2σ errors bars from bootstrapping). It is also shown as an upper limit using the signal injection framework (solid blue line), which is, interestingly, larger than the errors computed from bootstrapping, likely because the full injection framework takes into account additional sample variance, whereas the bootstrapped errors do not. Finally, the gray shaded regions combine the measured uniform-weighted power spectrum values with thermal noise errors. We show this power spectrum result as one example of how a simple regularization of an empirical covariance matrix can minimize signal loss, though we also note that this weighting does not produce more stringent limits than the uniform-weighted case, thus further motivating uniform weighting for M. Kolopanis et al. (2018, in preparation).

In this section, we have shown three simple ways of regularizing $\widehat{\boldsymbol{C}}$ to minimize signal loss using PAPER-64 data. There are many other weighting schemes that we leave for consideration in future work. For example, one could estimate $\widehat{\boldsymbol{C}}$ using information from different subsets of baselines. For redundant arrays, this might mean calculating $\widehat{\boldsymbol{C}}$ from a different but similar baseline type, such as the $\sim$30 m diagonal

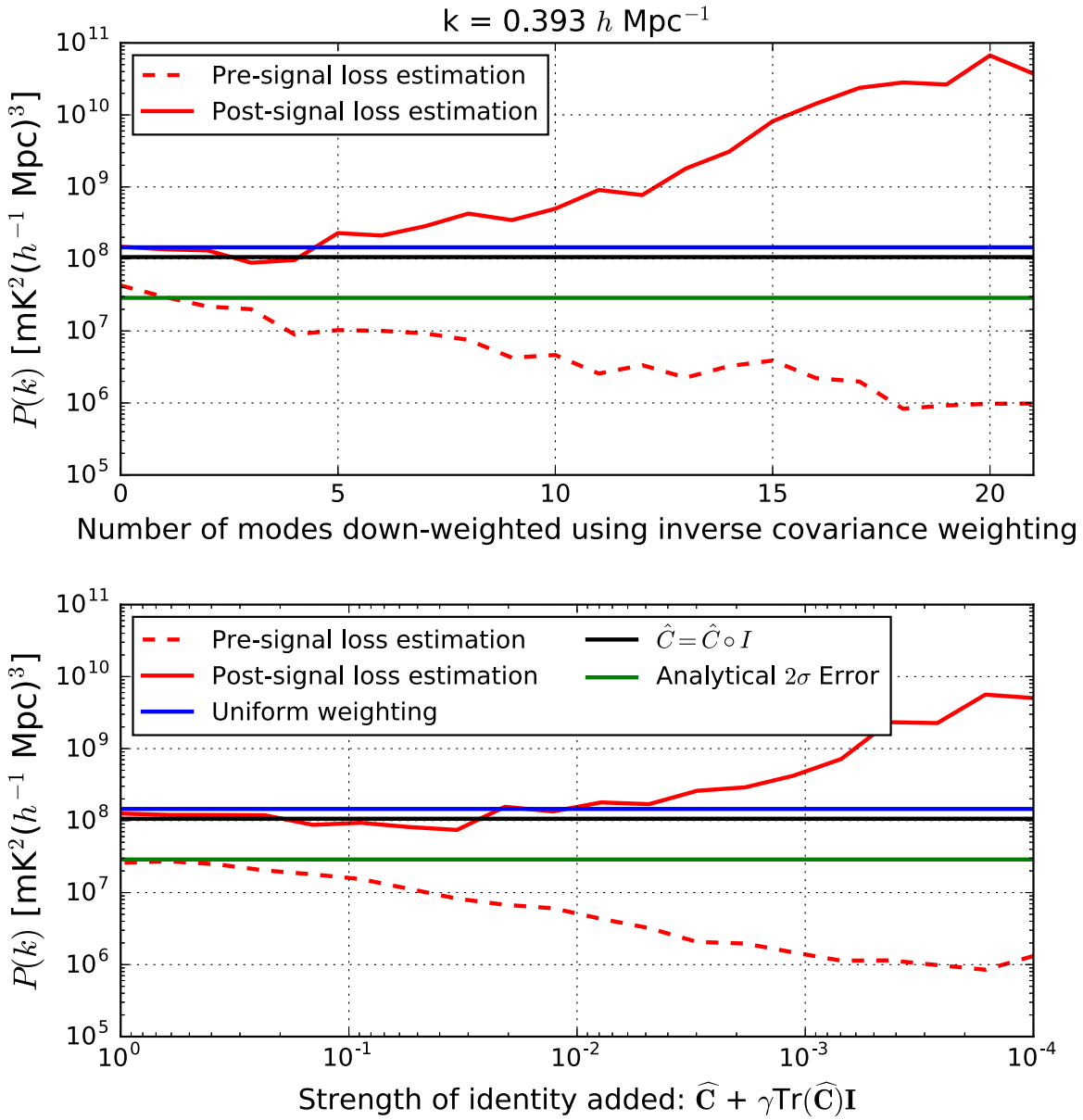

Figure 14. Power spectra 2σ upper limits for $k = 0.393$ h Mpc^{-1} for fringe-rate filtered PAPER-64 data. Top: values are shown before (dashed red) and after (solid red) signal loss estimation via our signal injection framework as a function of the number of eigenmodes of $\widehat{\boldsymbol{C}}$ that are down-weighted. This regularization knob is tuned from zero modes on the left (i.e., unweighted) to 21 modes on the right (i.e., the full inverse covariance estimator). About 4 orders of magnitude of signal loss results when using empirically estimated inverse covariance weighting. Bottom: power spectrum upper limits before (dashed red) and after (solid red) signal loss estimation as a function of identity added to the empirical covariance. This regularization knob is tuned from $\gamma = 10^{-4}$ on the right (i.e., very little regularization) to $\gamma = 1$ on the left (see main text for the definition of γ). Also plotted in both panels for comparison are 2σ power spectrum upper limits for the uniform-weighted case (blue) and inverse variance-weighted case (black); both are after signal loss estimation. Finally, a theoretical prediction for noise (2σ error) is plotted in green. In the PAPER-64 analysis in this paper, we choose to use a regularization scheme of $\widehat{\boldsymbol{C}}_{\rm eff} \equiv 0.09\ \mathrm{Tr}(\widehat{\boldsymbol{C}})\boldsymbol{I} + \widehat{\boldsymbol{C}}$ ($\gamma = 0.09$) as a simple example of regularization that minimizes loss, and we note that the power spectrum limits using this type of regularization are roughly constant across a large range of values of γ.

PAPER baselines (instead of the horizontal E/W ones). Alternatively, covariances could be estimated from all baselines except the two being cross-multiplied when forming a power spectrum estimate. This method was used in Parsons et al. (2014; a similar method was also used in Dillon et al. 2015) in order to avoid suppressing the 21 cm signal, and it is worth noting that the PAPER-32 results are likely less impacted by the issue of signal loss underestimation for this very reason (however, they are affected by the error estimation issues described in Section 4.2, so we also regard those results as suspect and superseded by those of M. Kolopanis et al. 2018, in preparation).

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

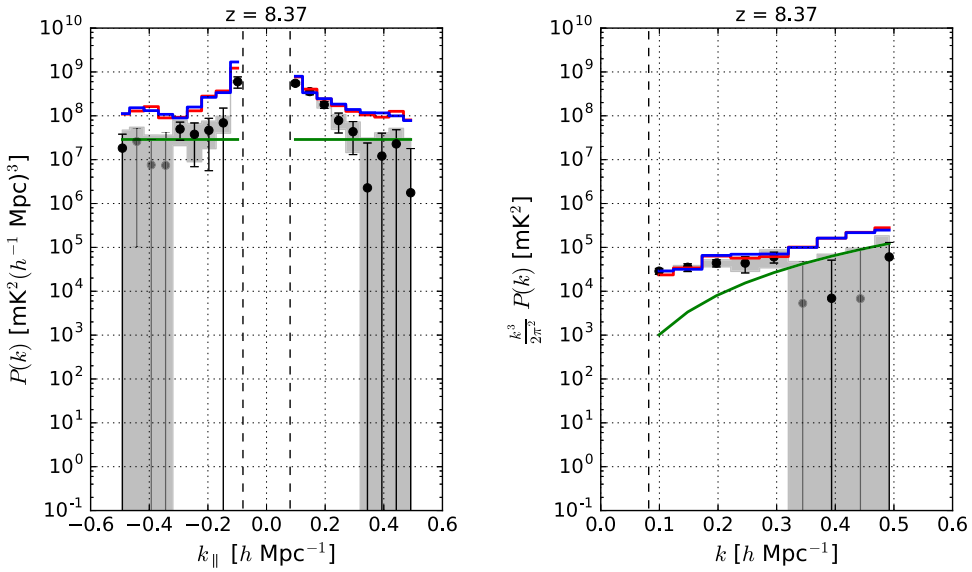

Figure 15. Power spectrum of a subset of PAPER-64 data illustrating the use of $\widehat{\boldsymbol{C}}_{\rm eff}$ to minimize signal loss. The solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework. The theoretical 2σ thermal noise level prediction based on observational parameters is in green. Additionally, the power spectrum result for the uniform-weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), 2σ upper limit on the EoR signal using our full signal injection framework (solid blue lines), and measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. This power spectrum result does not use the full data set's sensitivity as in A15 and M. Kolopanis et al. (2018, in preparation), though we include all analysis changes, which have mostly stemmed from revisions regarding signal loss, bootstrapping, and the theoretical error computation. We see that the regularization scheme used here produces limits similar to the unweighted limits.

from overfitting EoR modes in the power spectrum data set itself.

In Section 3, we have detailed several aspects of signal loss in PAPER-64: how the loss arises, how it can be estimated from an injection framework, and ways it can be minimized. We again emphasize that these lessons learned about signal loss are largely responsible for shaping our revised analysis of PAPER data. In the remainder of this paper, we will transition to other aspects of our analysis that have been revised since A15.

4. Additional PAPER-64 Revisions

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

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

4.1. Bootstrapping

Broadly speaking, we desire robust methods for determining accurate confidence intervals for our measurements. For PAPER's analysis, we choose a data-driven method of error estimation, computing error bars that have been derived from the inherent variance of our measurements. A common technique used to do this is bootstrapping, which we first define below and then discuss its application to PAPER.

Bootstrapping uses sampling with replacement to estimate a posterior distribution. For example, bootstrap measurements (of power spectra, for example) can be made from different random samples of data. Each of these bootstraps is a different realization drawn from some underlying distribution, and realizations are correlated with each other to a degree set by the fraction of sampled points that are held in common between them. Through the process of resampling and averaging along different axes of a data set, such as along baselines or times, we can estimate error bars for our results that represent the underlying distribution of values that are allowed by our measurements (Efron & Tibshirani 1994; Andrae 2010).

One major caveat of bootstrapping arises when working with correlated data. If, for example, a data set has many repeated values inside it, this would be reflected in each bootstrap. The same value would be present multiple times within a bootstrap and also between bootstraps, purely because it has a more likely chance of being drawn if there are repeats of itself. Therefore, bootstrapping correlated data results in a smaller variation between bootstraps and hence underestimates errors.

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

More specifically, the PAPER-64 pipeline outputs 20 bootstraps (over baselines), each a two-dimensional power spectrum that is a function of k and time. In A15, a second round of bootstrapping occurred over the time axis, and a total

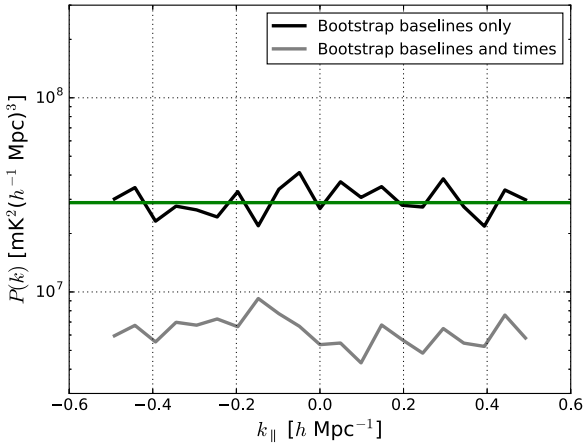

Figure 16. The 2σ power spectrum errors (from bootstrap variances) for a noise simulation (computed via Equation (27) using PAPER-64 observing parameters) using two different bootstrapping methods. The noise is fringe-rate filtered, and a weighting matrix of $\boldsymbol{I}$ (uniform-weighted) is used in order to disentangle the effects of bootstrapping from signal loss. The bootstrapping method used in A15 is shown in gray, where bootstrapping occurs along both the baseline and time axes. This underestimates the errors by sampling more values than independent ones in the data set (fringe-rate filtering reduces the number of independent samples along time). We use the method illustrated by the black curve in our updated analysis, where bootstrapping only occurs along the baseline axis. We find that these revised limits agree with the 2σ analytic prediction for noise (green).

of 400 bootstraps were created in this step, each comprised of randomly selected values sampled with replacement (i.e., each of these bootstraps contained the same number of values as the number of time integrations, which, at ∼700, greatly exceeds the approximate number of independent samples after fringe-rate filtering). Means were then taken of the values in each bootstrap. Finally, power spectrum limits were computed by taking the mean and standard deviation over all the bootstraps. We emphasize again that in this previous analysis, the number of elements sampled per bootstrap greatly exceeded the number of independent LST samples, underestimating the errors. A random draw of 700 measurements from this data set has many repeated values, and the variance between hundreds of these random samples is smaller than the true underlying variance of the data.

Given our new understanding of the sensitivity of bootstraps to the number of elements sampled, we have removed the second bootstrapping step along time entirely and now simply bootstrap over the baseline axis. Power spectrum 2σ errors (computed from bootstrap variances) with and without this bootstrapping change for a fringe-rate filtered noise simulation are shown in Figure 16 in black and gray, respectively. The estimates are uniformly weighted in order to disentangle the effects of bootstrapping from signal loss. As shown in the figure, when more elements are drawn for each bootstrap than the number of independent samples (by oversampling elements along the time axis), repeated values begin to crop up and the apparent variation between bootstraps drops, resulting in limits (gray) below the predicted noise level (green). Using the revised bootstrapping method, where bootstrapping only occurs over the baseline axis, the limits (black) are shown to agree with the analytic prediction for noise. While Figure 16 implies that errors computed prior to our bootstrapping change (gray) are underestimated by a factor of ∼5 in mK^2 for the noise simulation (whose creation details are outlined in the next section), in practice, this factor is lower for the case of real data (a factor of ∼3 in mK^2 instead), possibly due to the data being less correlated in time than the fringe-rate filtered noise in the simulation.

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

1. A second change to our bootstrapping procedure is that we now bootstrap over baseline cross-products, instead of the baselines themselves. In the previous analysis, baselines were bootstrapped prior to forming cross power spectra, and using this particular ordering of operations (bootstrapping, then cross-multiplication) yields variances that have been found to disagree with predicted errors from bootstrapping using simulations. On the contrary, bootstrapping over cross power spectra ensures that we are estimating the variance of our quantity of interest (i.e., the power spectrum). This change, while fundamental in retaining the integrity of the bootstrapping method in general, alters the resulting power spectrum errors by factors of <2 in practice.
2. In A15, individual baselines were divided into five independent groups, where no baselines were repeated in each group. Then, baselines within each group were averaged together, and the groups were cross-multiplied to form power spectra. This grouping method was used to reduce computational time. However, upon closer examination, it has been found that the initial grouping introduces an element of randomness into the final measurements; more specifically, the power spectrum value fluctuates depending on how baselines are assigned into their initial groups. Our new approach removes this element of randomness at the cost of computational expense, as we now perform all baseline cross-products.
3. Finally, the last change from the A15 method is that our power spectrum points (previously computed as the mean of all bootstraps) are now computed as the power spectrum estimate resulting from not bootstrapping at all. More specifically, we compute one estimate without sampling, and this estimate is propagated through our signal loss computation (this estimate is $\widehat{\boldsymbol{P}}_x$). The difference between taking the mean of the bootstrapped values and using the estimate from the no-bootstrapping case is small, but doing the latter ensures that we are forming results that reflect the estimate preferred by all our data.

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

4.2. Theoretical Error Estimation

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

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

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

$$N(k) = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{\sqrt{2 N_{\rm lst} N_{\rm seps}}\, t_{\rm int} N_{\rm days} N_{\rm bls} N_{\rm pols}}. \quad (22)$$

We will now explain each factor in Equation (22) and highlight key differences from the numbers used in A15.

1. X^2Y: conversion factors from observing coordinates (angles on the sky and frequency) to cosmological coordinates (comoving distances). For $z = 8.4$, $X^2Y = 5 \times 10^{11}\, h^{-3}$ Mpc^3 str^{-1} GHz^{-1}.
2. $\Omega_{\rm eff}$: effective primary beam area in steradians (Parsons et al. 2010; Pober et al. 2012). The effective beam area changes with the application of a fringe-rate filter, since different parts of the beam are up-weighted and down-weighted. Using numbers from Table 1 in Parsons et al. (2016), $\Omega_{\rm eff} = 0.74^2/0.24$ for an optimal fringe-rate filter and the PAPER primary beam.
3. $T_{\rm sys}$: system temperature defined as

$$T_{\rm sys} = 180\left(\frac{\nu}{0.18}\right)^{-2.55} + T_{\rm rcvr}, \quad (23)$$

where ν are frequencies in GHz (Thompson et al. 2001). We use a receiver temperature of 144 K, yielding $T_{\rm sys} = 431$ K at 150 MHz. This is lower than the $T_{\rm sys}$ of 500 K used in A15 because of several small miscalculation errors that were identified.[18]
4. $\sqrt{2}$: factor in the denominator of the sensitivity equation that comes from taking the real part of the power spectrum estimates after cross-multiplying two independent visibility measurements. In A15, a factor of 2 was mistakenly used.
5. $N_{\rm lst}$: number of independent LST bins that go into a power spectrum estimation. The sensitivity scales as the square root because we integrate incoherently over time. For PAPER-64, $N_{\rm lst} = 8$.
6. $N_{\rm seps}$: number of baseline separation types (where baselines of a unique separation type have the same orientation and length) averaged incoherently in a final power spectrum estimate. For the analysis in this paper, we only use one type of baseline (PAPER's 30 m East/West baselines). However, both the updated limits in M. Kolopanis et al. (2018, in preparation) and the sensitivity prediction in Figure 17 use three separation types ($N_{\rm seps} = 3$) to match A15.
7. $t_{\rm int}$: length of an independent integration of the data. It is crucial to adapt this number if filtering is applied along the time axis (i.e., a fringe-rate filter). We compute the effective integration time of our fringe-rate filtered data by scaling the original integration time t_i using

$$t_{\rm int} = t_i \frac{\int 1\, df}{\int w^2(f)\, df}, \quad (24)$$

where $t_i = 43$ s, $t_{\rm int}$ is the fringe-rate filtered integration time, w is the fringe-rate profile, and the integral is taken over all fringe rates. For PAPER-64, this number is $t_{\rm int} = 3857$ s.
8. $N_{\rm days}$: total number of days of data analyzed. In A15, this number was set to 135. However, because we divide our data in half (to form "even" and "odd" data sets, or $N_{\rm data\ sets} = 2$), this number should reflect the number of days in each individual data set instead of the total. Additionally, this number should be adjusted to reflect the actual number of cross-multiplications that occur between data sets ("even" with "odd" and "odd" with "even," but not "odd" with "odd" or "even" with "even," in order to

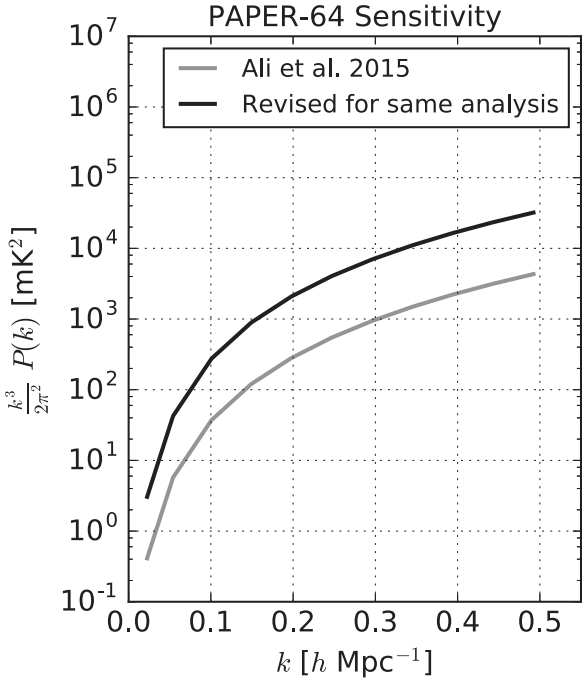

Figure 17. Updated prediction for the thermal noise level of PAPER-64 data (black) in comparison to previously published sensitivity limits (gray), both computed for the parameters and methods used in A15. Major factors that contribute to the discrepancy are $\Omega_{\rm eff}$, $N_{\rm days}$, and $N_{\rm bls}$, as in Equation (22) and described in Section 4.2, which when combined decrease our sensitivity (higher noise floor) by a factor of $\sim$7 in mK^2.

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

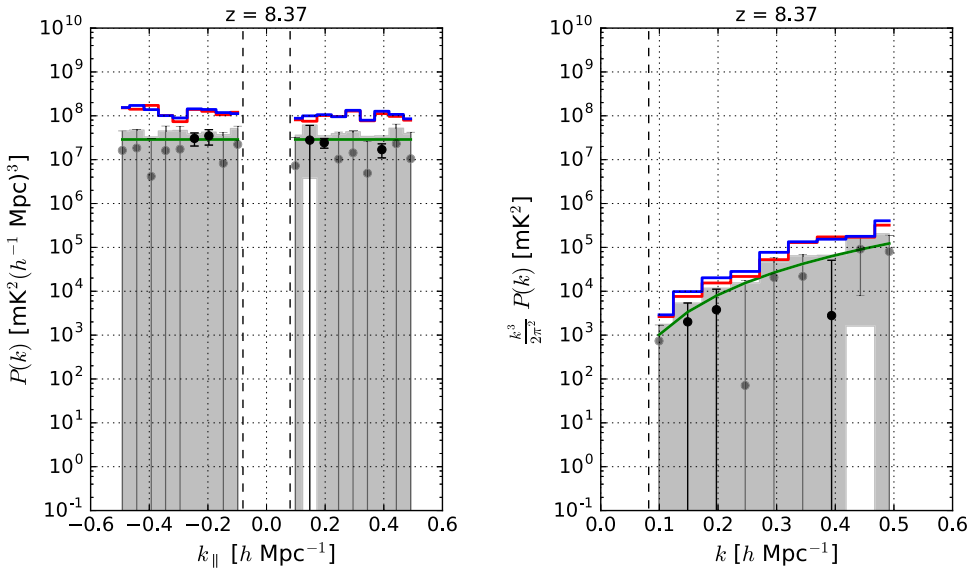

Figure 18. Power spectrum for a noise simulation that mimics the noise level of a subset of PAPER-64 data, where the solid red curve is the 2σ upper limit on the EoR signal estimated from our signal injection framework using $\widehat{C}_{\rm eff}$. The theoretical 2σ thermal noise level prediction based on observational parameters (calculated by Equation (22)) is in green. Additionally, the power spectrum result for the uniform-weighted case is shown in three different ways: power spectrum values (black and gray points as positive and negative values, respectively, with 2σ error bars from bootstrapping), the 2σ upper limit on the EoR signal using our full signal injection framework (solid blue line), and the measured power spectrum values with 2σ thermal noise errors (gray shaded regions). The vertical dashed black lines signify the horizon limit for this analysis using 30 m baselines. We highlight that the bootstrapped data points and thermal noise prediction show good agreement, while the limits from the full injection framework (red and blue) are inflated due to the additional inclusion of sample variance that comes from the injection simulations.

avoid noise biases). Finally, because our LST coverage is not 100% complete (it does not overlap for every single day), we incorporate an rms statistic in computing a realistic value of $N_{\rm days}$. Our expression therefore becomes

$$N_{\rm days} = \sqrt{\langle N_i^2 \rangle}\sqrt{(N^2_{\text{data sets}} - N_{\text{data sets}})}, \tag{25}$$

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

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

$$N_{\rm bls} = \frac{N_{\rm bls\_total}}{N_{\rm gps}}\sqrt{\frac{N_{\rm gps}^2 - N_{\rm gps}}{2}}, \tag{26}$$

where, in the A15 analysis, $N_{\rm gps} = 5$. Each baseline group averages down linearly as the number of baselines entering the group ($N_{\rm bls\_total}/N_{\rm gps}$) and then as the square root of the number of cross-multiplied pairs $\left(\sqrt{\frac{N_{\rm gps}^2 - N_{\rm gps}}{2}}\right)$. A revised A15 analysis should therefore use $N_{\rm bls} \sim 32$ instead of 51, and this change is taken into account in Figure 17. However, the analysis in this paper and M. Kolopanis et al. (2018, in preparation) no longer averages baselines into groups ($N_{\rm gps} = 1$). For the subset of data presented in this paper, $N_{\rm bls} = 10$.

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

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

Our revised sensitivity estimate for the A15 analysis of PAPER-64 is shown in Figure 17. Together, the revised parameters yield a decrease in sensitivity (higher noise floor) by a factor of ∼7 in mK^2.

To verify our thermal noise prediction, we form power spectra estimates using a pure noise simulation. We create Gaussian random noise assuming a constant $T_{\rm rcvr}$ (translated into $T_{\rm sys}$ via Equation (23)) but accounting for the true $N_{\rm days}$ as determined by LST sampling counts for each time and frequency in the LST-binned data. We convert $T_{\rm sys}$ into an rms variance statistic using

$$T_{\rm rms} = \frac{T_{\rm sys}}{\sqrt{\Delta\nu\Delta t N_{\rm days} N_{\rm pols}}}, \tag{27}$$

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

Power spectrum results for the noise simulation, which uses our full power spectrum pipeline, are shown in Figure 18. We highlight that the bootstrapped data (black and gray points with 2σ error bars) and thermal noise prediction (solid green line) show good agreement, as bootstrapping provides an accurate estimate of the noise variance. However, the limits from the full signal loss framework (weighted and unweighted in red and blue, respectively) are inflated, likely due to the additional inclusion of sample variance that comes from the EoR simulations. While the noise simulation provides an important indicator of the accuracy of our theoretical noise calculation, we note that the calculation did not take into account additional sources of error associated with earlier analysis steps (for example, Trott & Wayth 2017 showed how calibration

specifically can add errors to visibilities). Additionally, we recommend that future work investigate possible error correlations between baseline pairs and any interaction effects between signal and noise that may effect error calculations. For these reasons, we interpret our noise prediction as the sensitivity floor for our measurements.

5. Conclusion

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

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

1. Substantial signal loss can result when weighting data using empirically estimated covariances due to couplings with the data realizations (Section 2). Loss of the 21 cm signal is especially significant when fewer independent modes exist in the data. Hence, there exists a trade-off between sensitivity-driven time-averaging techniques such as fringe-rate filtering and signal loss when using empirically estimated covariances.
2. Signal injection and recovery simulations can be used to quantify signal loss (Section 3.1). However, a signal-only simulation (i.e., comparing a uniformly weighted versus weighted power spectrum of EoR only) can underestimate loss by failing to account for correlations between the data and signal that can be large and negative.
3. Errors that are estimated via bootstrapping can be underestimated if samples in the data set are significantly correlated (Section 4.1). However, if the number of independent samples in a data set is well-determined, bootstrapping is a simple and accurate way of estimating errors.

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

1. Signal loss, previously found to be $<2\%$ in A15, was underestimated by a factor of >1000 for the case of empirically estimated inverse covariance weighting. Using a regularized covariance weighting method can minimize loss (Section 3.3). However, because a regularized weighting method is not as aggressive as the former, it produces limits that are still higher than the lossy empirical inverse covariance limits. Underestimated signal loss therefore represents the bulk of our revision.
2. Power spectrum errors, originally computed by bootstrapping, were underestimated for the data by a factor of ~ 2 in mK due to oversampling data whose effective number of independent samples was reduced from fringe-rate filtering (Section 4.1). Several other errors were also found regarding error estimation, though with smaller effects.
3. Several factors used in an analytic expression to predict the noise level in PAPER-64 data were revised, yielding a decrease in predicted sensitivity level by a factor of ~ 3 in mK (Section 4.2). We note that our sensitivity prediction is revised by a factor less than our overall power spectrum result, implying that if taken at face value, the theoretical prediction for noise in A15 was too high for its data points.

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

CC would like to acknowledge the UC Berkeley Chancellor's Fellowship and National Science Foundation Graduate Research Fellowship (Division of Graduate Education award 1106400). She would also like to thank Phil Bull, Bryna Hazelton, Miguel Morales, and 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). AL acknowledges support for this work by NASA through Hubble Fellowship grant no. *HST*-HF2-51363.001-A, awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. SAK is supported by the University of Pennsylvania School of Arts and Sciences Dissertation Completion Fellowship. JSD acknowledges NSF AAPF award 1701536. GB acknowledges support from the Royal Society and the Newton Fund under grant NA150184. This work is based on research supported in part by the National Research Foundation of South Africa (award 103424). We graciously thank SKA-SA for site infrastructure and observing support.

ORCID iDs

Daniel C. Jacobs https://orcid.org/0000-0002-0917-2269
Adrian Liu https://orcid.org/0000-0001-6876-0928
Saul A. Kohn https://orcid.org/0000-0001-6744-5328
James E. Aguirre https://orcid.org/0000-0002-4810-666X
Gianni Bernardi https://orcid.org/0000-0002-0916-7443

Chris L. Carilli https://orcid.org/0000-0001-6647-3861
Joshua S. Dillon https://orcid.org/0000-0003-3336-9958
Chuneeta D. Nunhokee https://orcid.org/0000-0002-5445-6586

References

Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2018, ApJ, 863, 201
Andrae, R. 2010, arXiv:1009.2755
Barkana, R., & Loeb, A. 2001, PhR, 349, 125
Barkana, R., & Loeb, A. 2008, MNRAS, 384, 1069
Bernardi, G., de Bruyn, A. G., Brentjens, M. A., et al. 2009, A&A, 500, 965
Bernardi, G., de Bruyn, A. G., Harker, G., et al. 2010, A&A, 522, A67
Bernardi, G., Greenhill, L. J., Mitchell, D. A., et al. 2013, ApJ, 771, 105
Bernardi, G., Zwart, J. T. L., Price, D., et al. 2016, MNRAS, 461, 2847
Bond, J. R., Jaffe, A. H., & Knox, L. 1998, PhRvD, 57, 2117
Bowman, J. D., & Rogers, A. E. E. 2010, Natur, 468, 796
Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, Natur, 555, 67
Burns, J. O., Lazio, J., Bale, S., et al. 2012, AdSpR, 49, 433
Chang, T.-C., Pen, U.-L., Bandura, K., & Peterson, J. B. 2010, Natur, 466, 463
de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247
DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001
Dillon, J. S., Liu, A., & Tegmark, M. 2013, PhRvD, 87, 043005
Dillon, J. S., Liu, A., Williams, C. L., et al. 2014, PhRvD, 89, 023002
Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015, PhRvD, 91, 123011
Dillon, J. S., & Parsons, A. R. 2016, ApJ, 826, 181
Dodelson, S., & Schneider, M. D. 2013, PhRvD, 88, 063537
Efron, B., & Tibshirani, R. 1994, An Introduction to the Bootstrap (London: Taylor and Francis)
Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, PhR, 433, 181
Ghosh, A., Bharadwaj, S., Ali, S. S., & Chengalur, J. N. 2011, MNRAS, 418, 2584
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Jacobs, D. C., Hazelton, B. J., Trott, C. M., et al. 2016, ApJ, 825, 114
Jacobs, D. C., Pober, J. C., Parsons, A. R., et al. 2015, ApJ, 801, 51
Jaynes, E. 1968, IEEE Trans. Syst. Sci. Cybern., 4, 227
Jelić, V., et al. 2008, MNRAS, 389, 1319
Joachimi, B. 2017, MNRAS, 466, L83
Kohn, S. A., Aguirre, J. E., Nunhokee, C. D., et al. 2016, ApJ, 823, 88
Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14) (Trieste: SISSA), 1
Liu, A., & Parsons, A. R. 2016, MNRAS, 457, 1864
Liu, A., Parsons, A. R., & Trott, C. M. 2014a, PhRvD, 90, 023018
Liu, A., Parsons, A. R., & Trott, C. M. 2014b, PhRvD, 90, 023019
Liu, A., & Tegmark, M. 2011, PhRvD, 83, 103006
Loeb, A., & Furlanetto, S. 2013, The First Galaxies in the Universe (Princeton, NJ: Princeton Univ. Press)
Masui, K. W., Switzer, E. R., Banavar, N., et al. 2013, ApJL, 763, L20
Moore, D. F., Aguirre, J. E., Parsons, A. R., Jacobs, D. C., & Pober, J. C. 2013, ApJ, 769, 154
Morales, M. F., & Wyithe, J. S. B. 2010, ARA&A, 48, 127
Paciga, G., lbert, J. G., Bandura, K., et al. 2013, MNRAS, 433, 639
Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. 2016, MNRAS, 460, 1567
Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, AJ, 139, 1468
Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106
Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51
Parsons, A. R., Pober, J. C., Aguirre, J. E., et al. 2012b, ApJ, 756, 165
Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012a, ApJ, 753, 81
Patil, A. H., Yatawatta, S., Zaroubi, S., et al. 2016, MNRAS, 463, 4317
Patra, N., Subrahmanyan, R., Sethi, S., Udaya Shankar, N., & Raghunathan, A. 2015, ApJ, 801, 138
Paz, D. J., & Sánchez, A. G. 2015, MNRAS, 454, 4326
Pearson, D. W., & Samushia, L. 2016, MNRAS, 457, 993
Pen, U. L., Wu, X. P., & Peterson, J. 2004, arXiv:astro-ph/0404083
Pober, J. C., Parsons, A. R., Aguirre, J. E., et al. 2013a, ApJL, 768, L36
Pober, J. C., Parsons, A. R., DeBoer, D. R., et al. 2013b, AJ, 145, 65
Pober, J. C., Parsons, A. R., Jacobs, D. C., et al. 2012, AJ, 143, 53
Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, ApJ, 782, 66
Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766
Pritchard, J. R., & Loeb, A. 2010, PhRvD, 82, 023006
Pritchard, J. R., & Loeb, A. 2012, RPPh, 75, 086901
Santos, M. G., Cooray, A., & Knox, L. 2005, ApJ, 625, 575
Sellentin, E., & Heavens, A. F. 2016, MNRAS, 456, L132
Sokolowski, M., Tremblay, S. E., Wayth, R. B., et al. 2015, PASA, 32, e004
Switzer, E. R., Chang, T.-C., Masui, K. W., Pen, U.-L., & Voytek, T. C. 2015, ApJ, 815, 51
Switzer, E. R., Masui, K. W., Bandura, K., et al. 2013, MNRAS, 434, L46
Taylor, A., & Joachimi, B. 2014, MNRAS, 442, 2728
Tegmark, M. 1997, PhRvD, 55, 5895
Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 2001, Interferometry and Synthesis in Radio Astronomy (2nd ed.; New York: Wiley)
Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, 7
Trott, C. M., Pindor, B., Procopio, P., et al. 2016, ApJ, 818, 139
Trott, C. M., & Wayth, R. B. 2017, PASA, 34, e061
Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101
van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
Voytek, T. C., Natarajan, A., Jáuregui García, J. M., Peterson, J. B., & López-Cruz, O. 2014, ApJL, 782, L9
Wu, X. 2009, BAAS, 41, 474