Draft version June 27, 2017
Preprint typeset using LaTeX style AASTeX6 v. 1.0

DATA ANALYSIS METHODS FOR THE DETECTION OF THE EPOCH OF REIONIZATION

Carina Cheng[1], et al.

[1]Astronomy Dept., U. California, Berkeley, CA

ABSTRACT

1. INTRODUCTION

By about one billion years after the Big Bang, the very first stars and galaxies are thought to have ionized all the neutral hydrogen that dominated the Universe's early life. This important transition, during which the first luminous structures formed from gravitational collapse and emit intense radiation, transforming the cold neutral gas into a plasma, is known as the epoch of reionization (EoR). The EoR represents an unexplored era in our cosmic dawn, whose history encodes important information regarding the nature of the first galaxies and the process of structure formation. A direct detection of the EoR would unlock powerful information about the intergalactic medium, revealing connections between the smooth matter distribution exhibited via cosmic microwave background (CMB) studies and the highly structured web of galaxies we observe today.

One promising technique to probe the EoR is to target the 21 cm wavelength emission that is emitted by neutral hydrogen via its spin-flip transition. This technique is powerful because it can be observed 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. The 21 cm line therefore offers a window into following the evolution of ionization, temperature, and density fluctuations on cosmic scales.

Although a detection of the EoR has not currently been made to-date, there are several radio telescope experiments that have succeeded in using the 21 cm signal from hydrogen in order to place constraints on the brightness of the EoR. Examples of experiments investigating the mean brightness temperature of the EoR relative to the CMB are EDGES (Bowman & Rogers 2010), the LWA (Ellingson et al. 2009), LEDA (Greenhill & Bernardi 2012), DARE (Burns et al. 2012), SciHi (Voytek et al. 2014), BIGHORNS (Sokolowski et al. 2015), and SARAS (Patra et al. 2015). Major interferometers, which seek to measure statistical power spectra, include the GMRT (Paciga et al. 2013), LOFAR (van Haarlem et al. 2013), the MWA (Tingay et al. 2013), the 21CMA (Peterson 2004, Wu 2009), and PAPER (Parsons et al. 2010). The Hydrogen Epoch of Reionization Array (HERA), which is currently being built, is a next-generation instrument that hopes to combine lessons learned from previous experiments and is forecasted to be able to make a successful [?σ] detection with an eventual [?] elements.

The major challenge that faces all 21 cm experiments is in isolating a small signal that is buried underneath foregrounds and instrumental systematics that are 4-5 orders of magnitude brighter. A clean measurement therefore requires an intimate understanding of the instrument and a very careful, thorough understanding of data analysis choices. With HERA on the horizon and continual progress being made in the field, it is becoming increasingly more important to develop methods that can both reinforce a potential real detection and explain any false ones. In this paper, we present approaches to both of these issues using data from the 64-element configuration of PAPER. We place special emphasis on the accurate injection and recovery of a fake EoR signal in order to assess signal loss associated with different weighting techniques, as well as the importance of using jack-knife tests to investigate sources of bias.

This paper is organized as follows. In section 2 we give a brief overview of the PAPER-64 array and observations. Section 3 presents two key components of any power spectrum analysis pipeline - accurately quantifying signal loss that may result from data analysis choices, and choosing a weighting scheme to maximum foreground down-weighting without amassing large amounts of signal loss. In Section 4 we highlight useful jack-knife tests that can help discriminate between excess detections from foregrounds. We conclude in Section 5.

2. OBSERVATIONS

The Donald C. Baker Precision Array for Probing the Epoch of Reionization (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, each 2 m on a side. The antenna layout is formatted in a grid layout (Figure 1), with 8 antennas on a side, 30 m spacing between antennas along the East/West direction, and 4 m spacing between antennas along the North/South direction. For the rest of this paper, we focus on data from only the 30 m pure East/West baselines.

PAPER-64 observed for a total of 135 nights between 2012-2013. The correlator processes a bandwidth of 100-

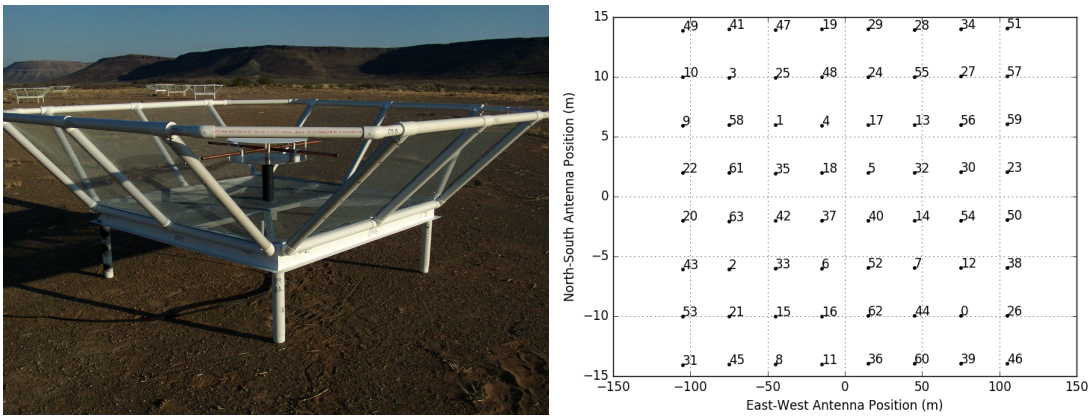

Figure 1. PAPER dipole in South Africa (left) and PAPER-64 antenna layout (right).

200 MHz, corresponding to a redshift range of 6-12. For more information about the backend system of PAPER-64 and its observations, we refer the reader to [??] and Ali et al. (2015).

Because there is a detailed discussion of the PAPER-64 data reduction pipeline in Ali et al. (2015), here we will only briefly summarize the data processing steps prior to the power spectrum analysis.

[TO DO: summarize redundant&absolute calibration, fg-clean, LST-binning, FRF]

3. POWER SPECTRUM ANALYSIS

We use optimal quadratic estimators as explained by [??] in order to form our final power spectrum quantities. For the reader's convenience, we summarize the method here, as well as highlight changes that we have made for our most recent analysis.

[TO DO: summarize OQE formalism and bootstrapping] [Don't forget to explain new PS channels, like noise and eor, and how they're created]

3.1. *Signal Loss Formalism*

Based on our analysis pipeline, potential signal loss is a real and significant issue. More specifically, when applying inverse covariance weighting, $\mathbf{C}^{-1}$ is empirically estimated from the data itself, which has the consequence of over-fitting the noise in the data, producing power spectra values well below the thermal noise limit that is predicted based on observation parameters. This is especially prevalent when weighting fringe-rate-filtered data, which has so few independent time modes to begin with, leading to a noisier dataset. Being able to accurately quantify this loss is crucial in interpreting and providing credibility to any power spectrum limits.

New to the PAPER-64 analysis is a robust method to estimate signal loss associated with inverse covariance weighting. This method, explained below, is now a standard analysis step for all PAPER analyses and one that will be used for HERA moving forward.

As discussed previously, our power spectrum pipeline runs on a standardized set of channels (pure data, pure noise, pure EoR, and combinations of the three). As part of our signal loss routine, we also compute power spectra with various levels of the created EoR signal, dialing its amplitude from well-below the data level, to well-above. Suppose that $\mathbf{e}$ is the injected EoR (at some amplitude level), and $\mathbf{x}$ is our data vector. We define $\mathbf{r}$ to be:

$$\mathbf{r} = \mathbf{x} + \mathbf{e} \tag{1}$$

Using our OQE formalism, we are interested in the following two quantities: P_{in} and P_{out}. The input power spectrum, P_{in} represents the unweighted power spectrum of only $\mathbf{e}$, our simulated EoR signal. The output power spectrum, P_{out}, is the weighted power spectrum of $\mathbf{e}$ that would result from our pipeline if the signal was mixed with our data. Comparing the two quantities yields insight into how much of $\mathbf{e}$ is lost due to our choice of weighting. Ignoring normalization, factors:

$$P_{in} \propto \mathbf{e}^\dagger \mathbf{I}^{-1}\mathbf{Q}\mathbf{I}^{-1}\mathbf{e} \tag{2}$$

$$\begin{aligned} P_{out} \equiv \mathbf{P}_e &= \mathbf{P}_r - \mathbf{P}_x \\ &\propto \mathbf{r}^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}\mathbf{r} - \mathbf{x}^\dagger \mathbf{C}_x^{-1}\mathbf{Q}\mathbf{C}_x^{-1}\mathbf{x} \end{aligned} \tag{3}$$

It is noted that the output power spectrum is comprised of two terms: the covariance treated power spectrum associated with $\mathbf{r}$, and that of data $\mathbf{x}$ alone.

One may wonder why P_{out} cannot be computed simply as the weighted power spectrum of $\mathbf{e}$ alone, namely $P_{out} \propto \mathbf{e}^\dagger \mathbf{C}_e^{-1}\mathbf{Q}\mathbf{C}_e^{-1}\mathbf{e}$. Expanding Equation 3:

$$\begin{aligned} P_{out} &\propto (\mathbf{x}+\mathbf{e})^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}(\mathbf{x}+\mathbf{e}) - \mathbf{x}^\dagger \mathbf{C}_x^{-1}\mathbf{Q}\mathbf{C}_x^{-1}\mathbf{x} \\ &\propto \mathbf{x}^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}\mathbf{x} + \mathbf{e}^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}\mathbf{e} + \mathbf{x}^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}\mathbf{e} \\ &\quad + \mathbf{e}^\dagger \mathbf{C}_r^{-1}\mathbf{Q}\mathbf{C}_r^{-1}\mathbf{x} - \mathbf{x}^\dagger \mathbf{C}_x^{-1}\mathbf{Q}\mathbf{C}_x^{-1}\mathbf{x} \end{aligned}$$

And taking the case of very large $\mathbf{e}$, so that $\mathbf{C}_r^{-1} \sim \mathbf{C}_e^{-1}$ and any terms involving only $\mathbf{x}$ are small:

$$\begin{aligned} P_{out,\mathbf{e}\gg\mathbf{x}} \propto \mathbf{e}^\dagger \mathbf{C}_e^{-1}\mathbf{Q}\mathbf{C}_e^{-1}\mathbf{e} + \mathbf{x}^\dagger \mathbf{C}_e^{-1}\mathbf{Q}\mathbf{C}_e^{-1}\mathbf{e} \\ + \mathbf{e}^\dagger \mathbf{C}_e^{-1}\mathbf{Q}\mathbf{C}_e^{-1}\mathbf{x} \end{aligned} \tag{4}$$

We see that our naive expression for P_{out} is the first term, but there also two additional terms. An initial assumption would be that the cross-terms that involve both $\mathbf{e}$ and $\mathbf{x}$ should be zero, since the two quantities are un-correlated. However, **[need explanation about power in cross-terms here]**. Therefore, in our investigation of signal loss, we use the full quantity for P_{out} as in Equation 3.

For the unweighted case ($\mathbf{C} \equiv \mathbf{I}$), we expect P_{out} and P_{in} to be equal, and hence the ratio of P_{in}/P_{out} to be 1. For the weighted case, this is not true due to signal loss. In order to quantify the loss, we look at the ratio of P_{in} to P_{out} as the amplitude level of the injected signal $\mathbf{e}$ is increased. In the next section, we highlight two methods that yield similar results for the determination of signal loss using P_{in} and P_{out}

3.2. *Signal Loss in Practice*

Recall that fringe-rate-filtered noise, which mimics the level of noise in our actual PAPER-64 dataset, is a channel in our power spectrum pipeline. We can compute signal loss quantities of interest for the noise $\mathbf{n}$ similar to the expressions we featured previously for data $\mathbf{x}$.

$$P_{in} \propto \mathbf{e}^\dagger \mathbf{I}^{-1}\mathbf{Q}\mathbf{I}^{-1}\mathbf{e} \tag{5}$$

$$\mathbf{s} = \mathbf{n} + \mathbf{e} \tag{6}$$

$$P_{out} \propto \mathbf{s}^\dagger \mathbf{C}_s^{-1}\mathbf{Q}\mathbf{C}_s^{-1}\mathbf{s} - \mathbf{n}^\dagger \mathbf{C}_n^{-1}\mathbf{Q}\mathbf{C}_n^{-1}\mathbf{n} \tag{7}$$

Using the input and output power spectra for range of EoR amplitudes, we use two methods to determine signal loss associated with the over-fitting of noise during inverse covariance weighting. We first look at signal loss for the pure noise case (no data), to show that we can successfully inject EoR signals, determine signal loss, and then recover the signal.

The first method is very straightforward. Post-bootstrapping, our final power spectra are 1-dimensional and only a function of k. For each k, we simply look at P_{out} as a function of P_{in}, as shown in Figure 2. The shape of this function can be explained as follows. At small injection levels (small $\mathbf{e}$), P_{out} and P_{in} are equal, and there is no signal loss. As the amplitude of EoR increases, we then move into a regime where the final output power spectrum is lower than the unweighted input one. This is dangerous, because without correcting for this effect one might be led to underestimate the EoR signal. The peculiar tail at very low injection levels (where $P_{out} > P_{in}$) is an unphysical feature, but rather illustrates that there is some non-negligible cross-term power between $\mathbf{n}$ and $\mathbf{e}$.

For this method, we interpolate the signal loss factor (per k), computed as P_{in}/P_{out}, at a P_{out} value equal to the 2σ power spectrum upper limit of noise alone. In other words, we look at $P_{noise} \propto \mathbf{n}^\dagger \mathbf{C}_n^{-1}\mathbf{Q}\mathbf{C}_n^{-1}\mathbf{n}$, compute its 2σ upper limit (mean over bootstraps + 2×standard deviation over bootstraps), and interpolate the value of P_{in}/P_{out} at this value. We therefore end up with one signal loss correction factor per k.

Figure 3 shows the power spectrum of our noise simulation, using full inverse covariance weighting, both before and after signal loss correction. Prior to signal loss correction, it is obvious that the power spectrum is unfeasible because it is well below the theoretical noise level prediction. Post-correction, the power spectrum values blow up much higher than both the theory and unweighted power spectrum. This is an effect caused by the steep nature of the eigenspectrum of $\mathbf{C}$, and is explained more in Section 3.3.

[Need a plot that shows our signal loss factors are CORRECT. How to do that??]

Our second method for estimating signal loss is similar to the first, but more comprehensive in a statistical sense. Instead of looking at input and output power spectra after bootstrapping, we now look at their values for every bootstrap in order to get a sense of their distributions. Figure 4 plots P_{in} vs. P_{out} for 20 bootstraps, and as expected, the function now has a spread in the width-direction in comparison to what was plotted in Figure 2, but otherwise shows a familiar trend. Similarly, our weighted noise power spectra also has a defined spread due to bootstrapping.

Using these two distributions (P_{in}/P_{out} and P_{noise}), we can create bins along the P_{noise} axis to yield histograms of signal loss factors for each bin. We similarly sort the values of P_{noise} into the same bins, and multiply the probability of P_{noise} per bin (the number of values falling into that bin, divided by the total) with the signal loss factors in that bin, essentially computing a weighted average across all bins to obtain a final signal loss factor per k. As shown in Figure 5, the results are very similar to the previous method. For future power spectrum results, we choose to use the second method because it computes final signal loss values using our full distributions of measurements.

One thing to note is that for both methods, we have been careful to validate that the computations yield no signal loss (signal loss factors of 1) for the unweighted power spectrum case, as is expected. This is important in confirming that signal loss is a direct result of the choice for $\mathbf{C}$.

3.3. *Data Weighting*

With our signal loss formalism established, we now have the capability of experimenting with different weighting options for $\mathbf{C}$. Our goal here is to choose a weighting method that successfully down-weights foregrounds and systematics in our data without generating large amounts of signal loss. We have found that the balance between the two is a delicate one and requires **[?? finish sentence...]**.

We now turn our attention to power spectra using the 30 m East/West baselines of PAPER-64. Our dataset spans 8.5 hours of LST (.1-8.6 hrs), includes a total of 51 baselines, and is fringe-rate-filtered using an optimal fringe-rate-filter. We have two datasets (even days and odd days), and only cross-multiply data from different

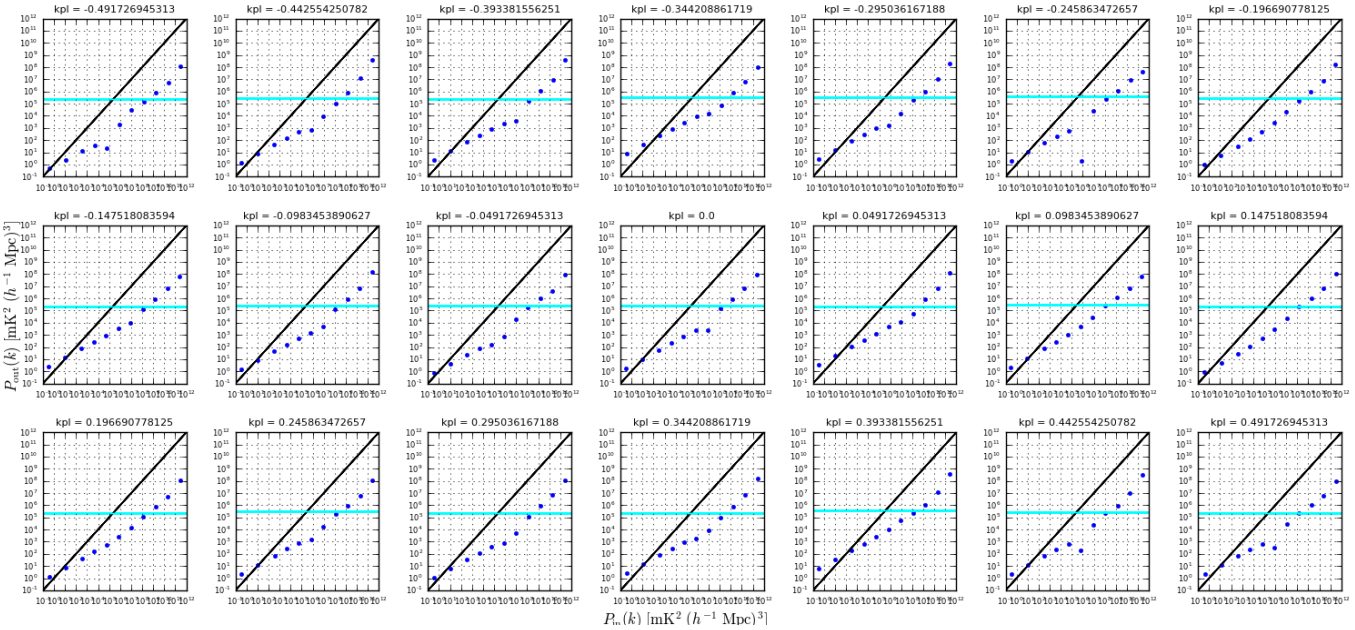

Figure 2. P_{in} vs. P_{out} (blue points) for 15 injection levels and 21 k's. The solid cyan line is the 2σ upper limit for the weighted power spectrum of noise alone, and it is at this level where the signal loss factor P_{in}/P_{out} is computed by interpolation. **[Maybe re-do this plot with closer-together points]**

days and different baselines. We are interested in 21 frequency channels (channels 95-115), which yields a power spectrum for a redshift of $z = 8.4$.

Using full inverse covariance weighting, our results are not too dissimilar to that of pure noise. Signal loss factors (Figure 6) are of similar order of magnitude, and our power spectrum blows up past the unweighted version after signal loss correction (Figure 7).

Looking into this behavior in more detail, we investigate the shape of the eigenspectrum of $\mathbf{C}$ for a typical baseline used in the analysis. Figure 8 shows this spectrum for baseline (1,4). Most obviously, the spectrum is steep, spanning 4 orders of magnitude. Not as obvious is the effect of this shape on our results. When the matrix $\mathbf{C}$ is inverted to form $\mathbf{C}^{-1}$, the effect of the steepness of the eigenspectrum is to up-weight very few modes of the sky while the rest are drastically down-weighted. More specifically, our fringe-rate-filtered data contains a finite, small number of independent modes, thereby resulting in a covariance matrix that can be described by just a few modes. Beyond the first few modes, the eigenvalues of each additional mode falls of dramatically. When inverting, we end up not only down-weighting those initial modes but severely up-weighting a few insignificant ones. Because of our weighting choice, signal loss blows up as it thinks we only have a couple modes in our data.

Clearly the full inverse covariance treatment of our data is suboptimal to even the unweighted case, but we would like to find a weighting method that does successfully down-weight contaminants in our data and make some improvement over the unweighted power spectrum. There are many choices for determining the covariance matrix $\mathbf{C}$, but here we will illustrate [?] promising ones as applied to PAPER-64.

[TO DO: decide on/explain/show different weightings]

4. JACK-KNIVES

5. CONCLUSION

6. ACKNOWLEDGEMENTS

[NSF Graduate Research Fellowship Program (GRFP) Fellowship] [UC Berkeley Chancellor's Fellowship]

REFERENCES

Ali, Z. S., et al. 2015, ApJ, 809, 61
Bowman, J. D., & Rogers, A. E. E. 2010, Nature, 468, 796
Burns, J. O., et al. 2012, Advances in Space Research, 49, 433
Ellingson, S. W., Clarke, T. E., Cohen, A., Craig, J., Kassim, N. E., Pihlstrom, Y., Rickard, L. J., & Taylor, G. B. 2009, IEEE Proceedings, 97, 1421
Greenhill, L. J., & Bernardi, G. 2012, ArXiv e-prints
Paciga, G., et al. 2013, MNRAS
Parsons, A. R., et al. 2010, AJ, 139, 1468
Patra, N., Subrahmanyan, R., Sethi, S., Udaya Shankar, N., & Raghunathan, A. 2015, ApJ, 801, 138
Peterson, U.-L. P. X.-P. W. J. 2004, ArXiv Astrophysics e-prints
Sokolowski, M., et al. 2015, PASA, 32, e004
Tingay, S. J., et al. 2013, PASA, 30, 7
van Haarlem, M. P., 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

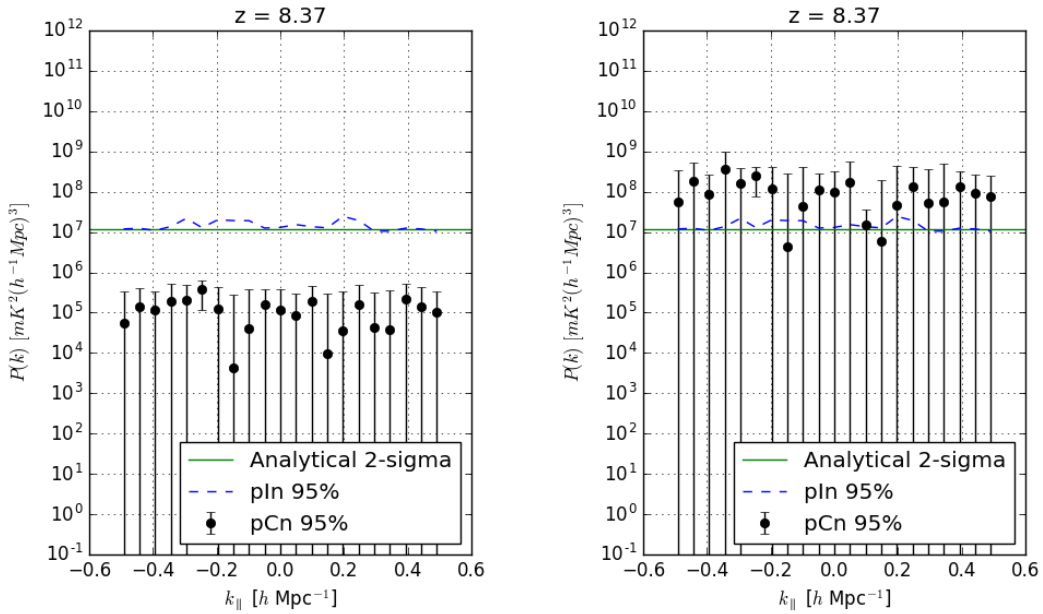

Figure 3. Full inverse covariance weighted power spectrum of pure noise (black points, with 2σ error bars) before signal loss correction (left) and after (right). The dashed blue line is the unweighted power spectrum (2σ upper limit). The solid green line is the theoretical noise level prediction based on observational parameters. **[Color negative points grey]**

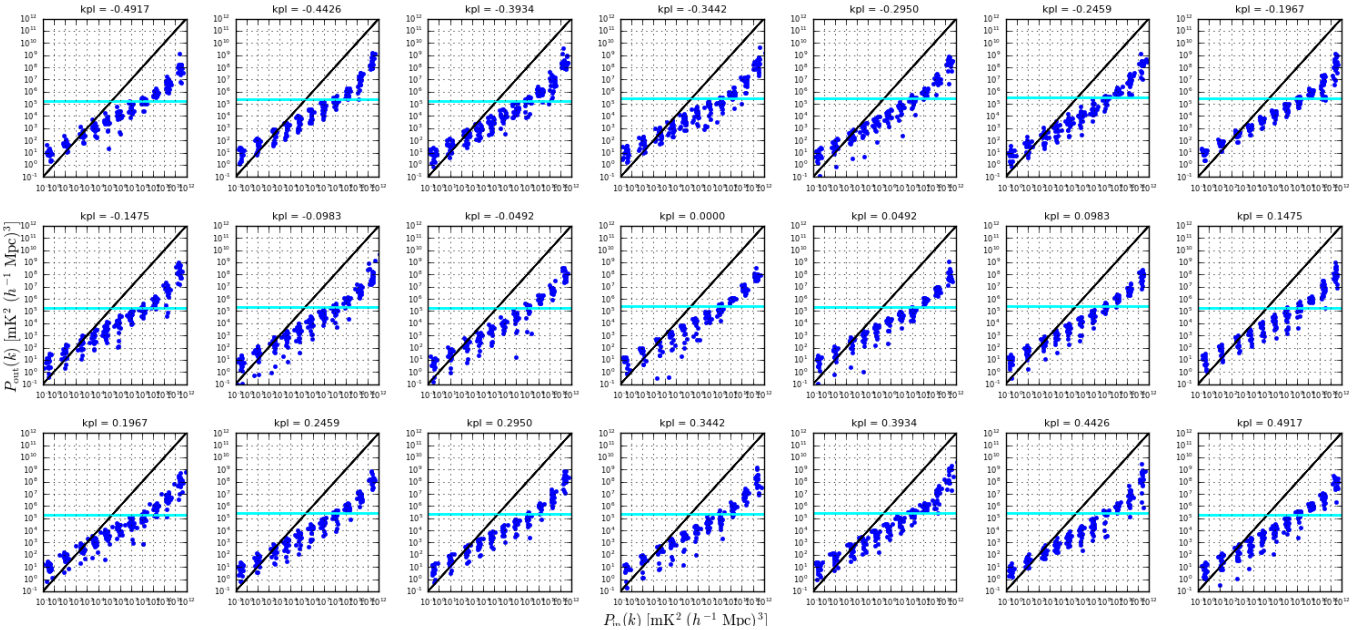

Figure 4. P_{in} vs. P_{out} (blue) for 15 injection levels, 20 bootstraps, and 21 k's. **[Plot a semi-transparent cyan range of pCn values instead of just the max]**

Wu, X. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, 474

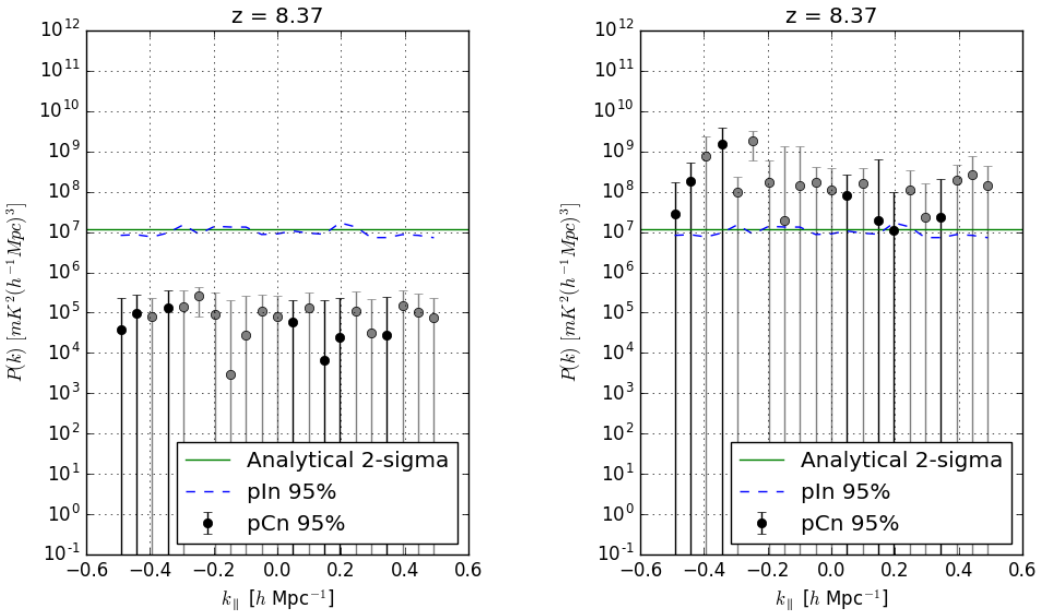

Figure 5. Full inverse covariance weighted power spectrum of pure noise (black and grey points, with 2σ error bars) before signal loss correction (left) and after (right). Black points correspond to positive values, while grey points correspond to originally negative values that have been made positive for plotting. The dashed blue line is the unweighted power spectrum (2σ upper limit). The solid green line is the theoretical noise level prediction based on observational parameters.

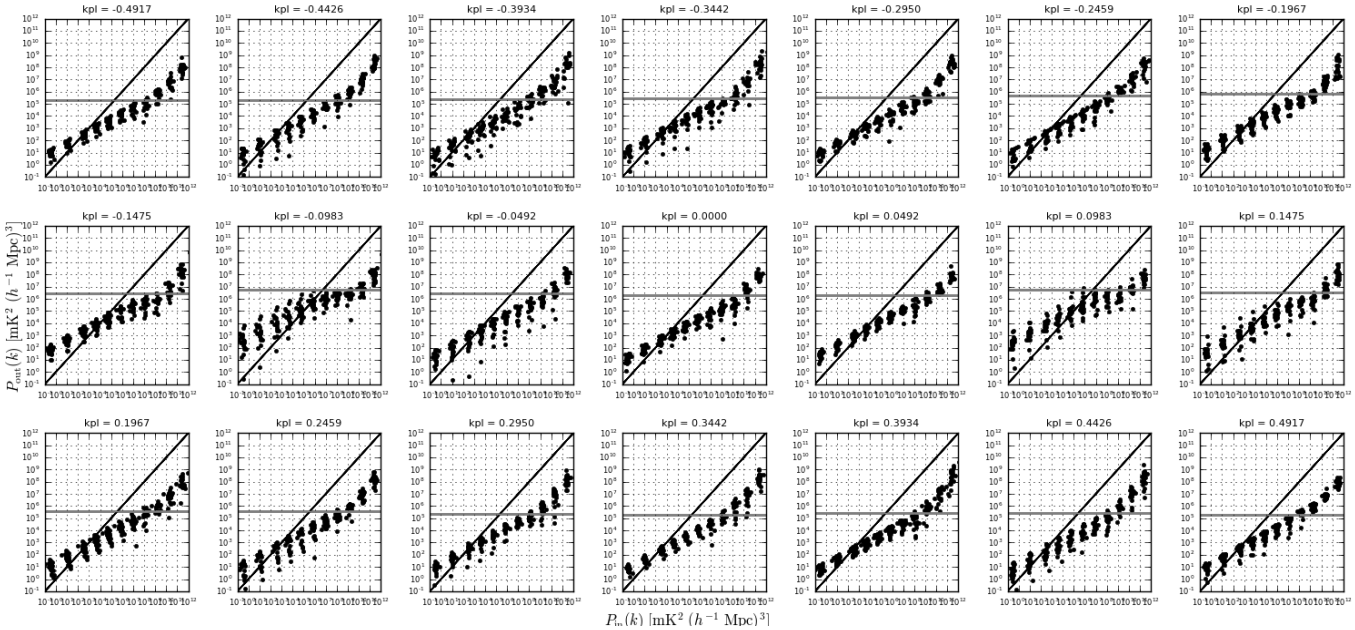

Figure 6. P_{in} vs. P_{out} (black) for 15 injection levels, 20 bootstraps, and 21 k's. **[Plot a semi-transparent grey range of pCv values instead of just the max]**

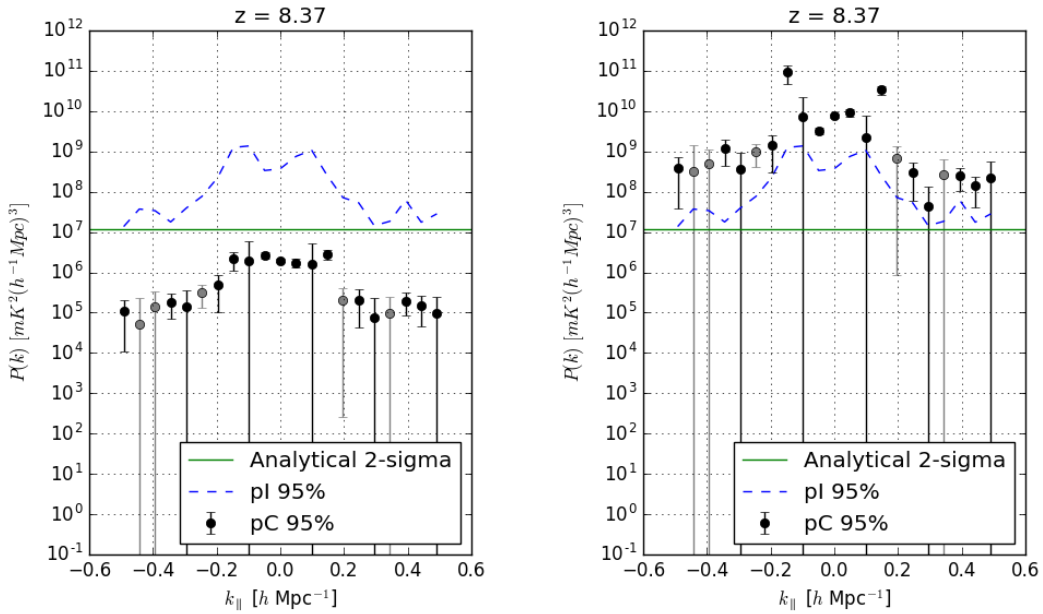

Figure 7. Full inverse covariance weighted power spectrum of PAPER-64 data (black and grey points, with 2σ error bars) before signal loss correction (left) and after (right). Black points correspond to positive values, while grey points correspond to originally negative values that have been made positive for plotting. The dashed blue line is the unweighted power spectrum (2σ upper limit). The solid green line is the theoretical noise level prediction based on observational parameters.

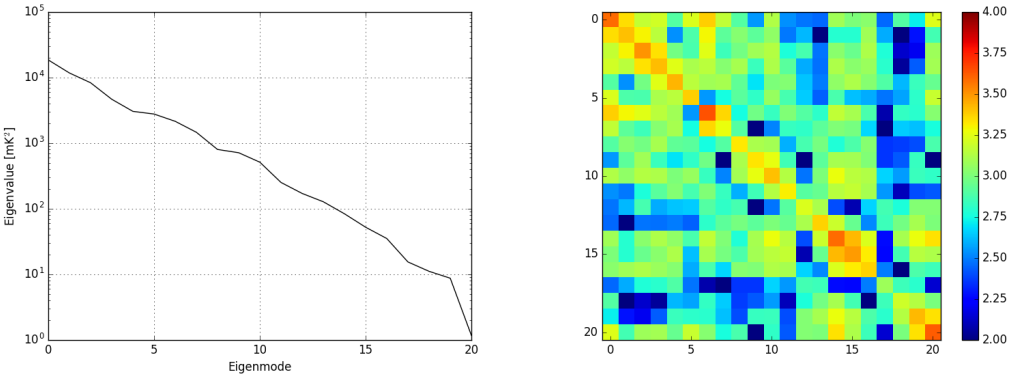

Figure 8. Eigenspectrum for **C** for baseline (1,4) for the 21 channels ofs interest (left) and covariance matrix **C** for the same baseline (right).