HERA Memo 106
Received 27 Oct 2021

Non-Gaussian Effects and Robust Location Estimates of Aggregated Calibrated Visibilities

Matyas Molnar and Bojan Nikolic

Astrophysics Group, Cavendish Laboratory, University of Cambridge

26th October 2021

Abstract

In the Hydrogen Epoch of Reionization Array (HERA) analysis pipeline, following the full calibration of individual datasets, complex visibilities undergo a round of median absolute deviation (MAD)-clipping at the 4σ level and are then mean averaged in Local Sidereal Time (LST). This procedure assumes an underlying Gaussian distribution for the visibility data, takes a cartesian approach when clipping and averaging (about the $\mathfrak{Re}$ and $\mathfrak{Im}$ components separately), and employs a clipping threshold that is somewhat ad hoc. In this memorandum, we visualize these visibilities as they are aggregated in LST and across redundant baselines, and test for the normality of these distributions. We further present and use the geometric median, a robust multivariate technique for location estimation, to better represent the data. We compare this robust estimate to the established clipping + mean approach in both the visibility and power spectrum (PS) domains.

1 Introduction

Complex visibility data from HERA is reduced by averaging both in LST across Julian dates (JDs) and again across redundant baselines, with the latter done after PS computation for further flexibility in rejecting dubious baselines and baselines paired with themselves, amongst other reasons. The methods used in the HERA analysis pipeline assume Gaussian distributed visibilities; several rounds of outlier rejection are performed before mean averaging with the assumption that the central distribution of visibilities is (near) normal such that mean averaging in LST can be done.

In light of the anomalous Band 2 Field 2 results in [1], which sees a systematic excess in power compared to thermal noise expectation, likely due to low-level radio-frequency interference (RFI), we wanted to apply robust location estimates instead of the established averaging to see if we could further remove RFI and improve results.

The data looked at in this memorandum is taken to be a subset of the fully calibrated[1] H1C_IDR2.2 dataset [2]. We consider all JDs except for JD 2458109, which is known to be mostly faulty. We look at channels $500 - 700$ (corresponding to frequencies $1.4883 - 1.6826$ MHz, coinciding with Band 2 from the results analysis in [1]), LSTs $5.2812 - 5.4577$ (a subset of Field 2), EE polarization and only look at the 19 short 14 m baselines redundant to $(1, 12)$. The resulting data array therefore has dimensions (JDs, frequencies, times, baselines) = (17, 200, 60, 19). When aggregating data in LST and/or across redundant baselines for location estimates, we further bin data from two consecutive time integrations such that two data points are considered for every input night, as is done in the LST-binning pipeline.

We note that the visibility data has also been rephased such that each row in time has its phase centre shifted to the mean for that row across JDs. Furthermore, when aggregating data into larger time

[1] Where the `.smooth_abs.calfits` calibration solutions have been applied.

bins, the data is rephased again to the centre of that bin.

In this memorandum, we first visualize visibility data that is aggregated both/either in LST and/or across redundant baselines in §2, and test for the normality of their distributions. We then review the steps used to average visibilities in LST in the HERA analysis pipeline in §3. We introduce robust multivariate estimators in §4, and in §5 we apply the geometric median estimator on these data to get their locations and show sample visibility results. Simple PS estimates are then also computed.

2 Data visualization

We employ kernel density estimate (KDE) plots to visualize the distribution of redundantly grouped and LST-binned HERA visibilities. We show a scatter plot in Fig. 1 and its corresponding KDE plot in Fig. 2, with various location estimators also marked (see §§4.2 and 4.3 for descriptions of the geometric and Tukey median estimators), for some sample data that is quite typical of the H1C_IDR2.2 dataset. As a reminder, two time integrations (spanning a cadence of 21.4 s) per JD are considered for each new LST bin throughout this memorandum, as to increase the sample count.

Four further examples of KDE plots are shown in Fig. 3. A video of the distribution of visibilities evolving as we sweep through frequency channels at fixed LST 5.2824 can be found here.

While Figs. 1 to 3 consider data aggregated both across JDs and redundant baselines, we also split up this aggregation and look at data distributed across JDs for each baseline redundant to (1, 12, EE). A video of this as we again sweep through frequency channels at fixed LST 5.4379 can be found here, while a snapshot at channel 527 is shown in Fig. 4.

Additionally, we examine the distributions of visibilities for each JD separately for all baselines in redundant group (1, 12, EE) at fixed LST 5.4379 across frequencies in a video here. We show a frame from this video at channel 544 in Fig. 5.

These KDE plots and videos reveal a considerable amount of structure in the distribution of visibilities that are aggregated together and separately across JDs and across redundant baselines. The spread of the visibilities is quite large and many distributions are multi-modal, with few appearing to be Gaussian. There also appears to be some possible leakage from neighbouring frequency channels. It is difficult to interpret these plots and even harder to narrow down where these effects would come from. We, however, notice that the data across JDs is more consistent than it is across redundant baselines, with the latter showing higher discrepancies between baselines, which potentially indicates that the calibration pipeline is not able to fully reconcile redundant visibilities.

2.1 Multivariate normality

We present the Henze-Zirkler statistic [3] to test and quantify the normality of multivariate data, and apply it to complex redundant HERA visibilities across JDs. We cautiously advise that this test, while commonly used in the literature, should not be solely relied on to check the normality of data; instead, multiple statistics should be used and the data itself should be visualized, if possible.

The Henze-Zirkler is a statistic based on a non-negative functional that measures the distance between the hypothesized function (which is the multivariate normal) and the observed function. It is given by

$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}} - 2(1+\beta^2)^{-\frac{p}{2}}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}D_i} + n(1+2\beta^2)^{-\frac{p}{2}} \tag{1}$$

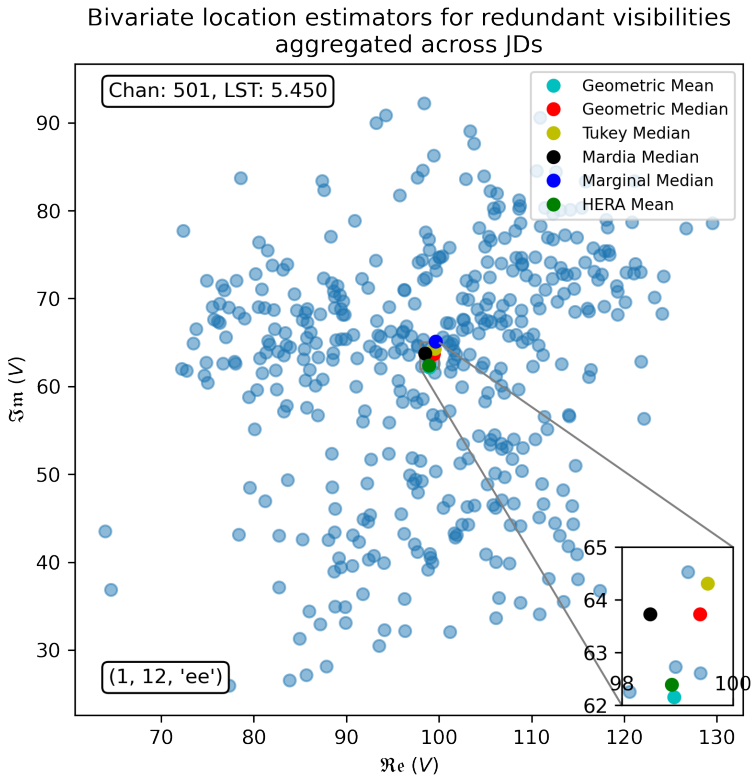


Figure 1: Scatter plot of aggregated visibility data for baseline group (1, 12, EE) across JDs at frequency channel 501 and LST 5.4533.

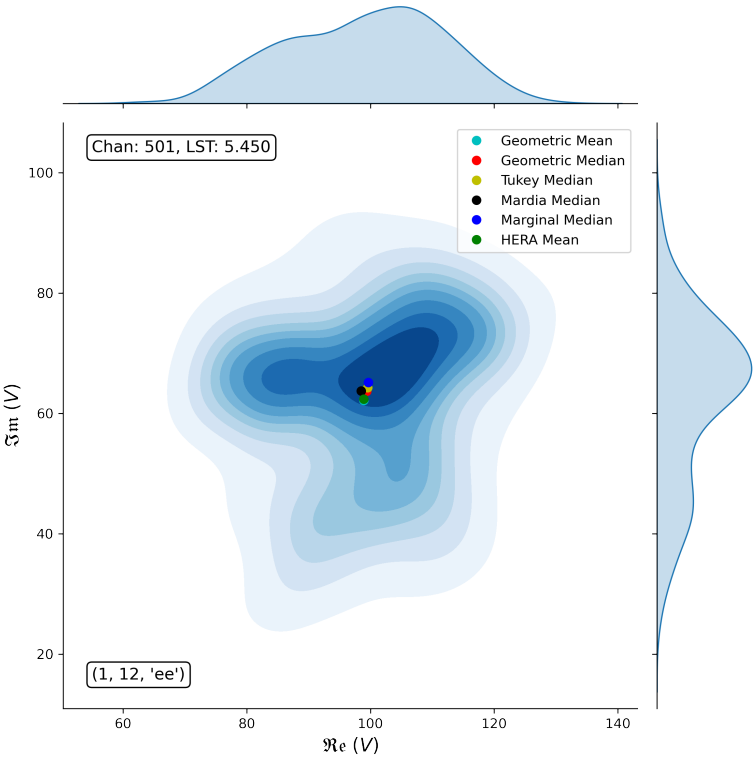
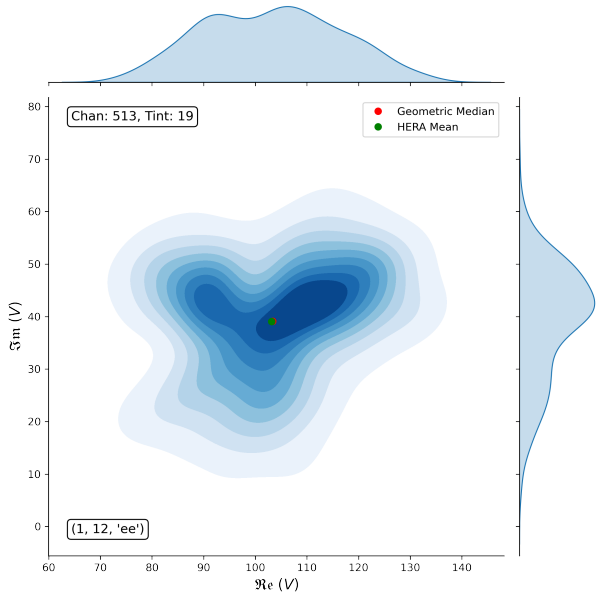
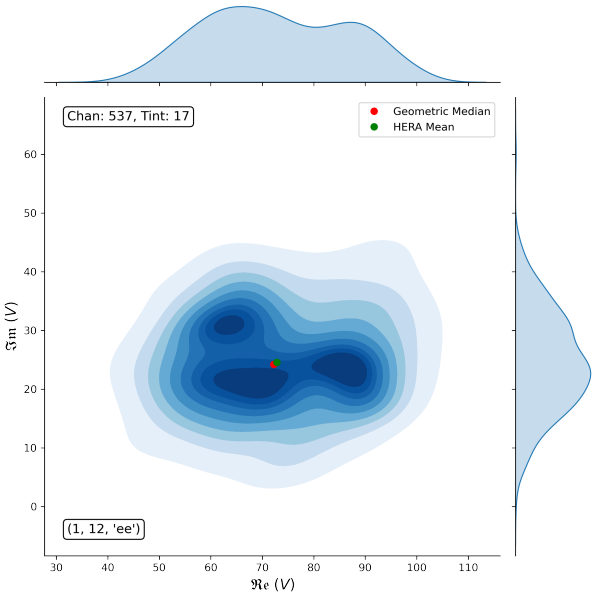


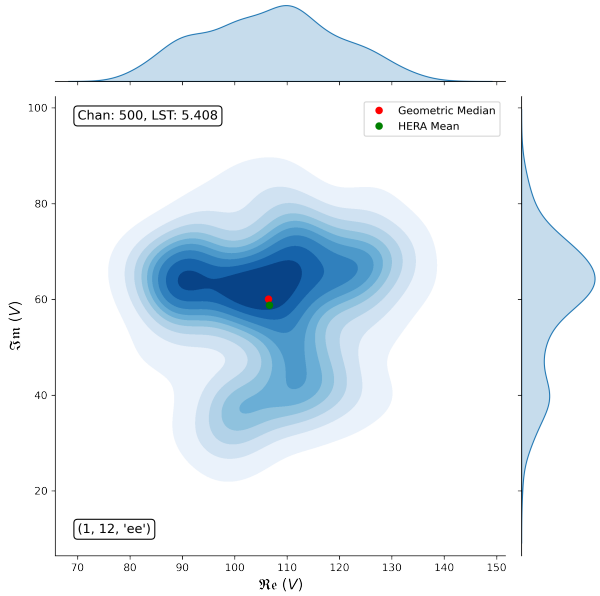
Figure 2: KDE of aggregated visibility data for baselines redundant to (1, 12, EE) across JDs at frequency channel 501 and LST 5.4533. The central distribution of visibilities does not appear to be Gaussian, with the geometric and Tukey medians appearing to best represent the location of the distribution.



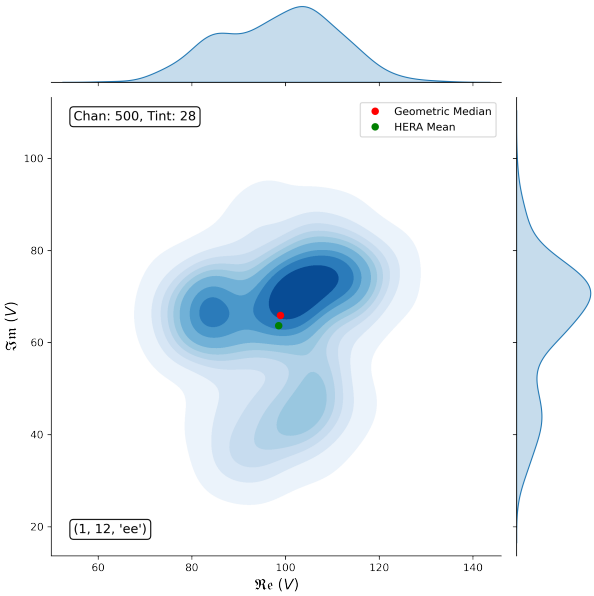
(a) Randomly selected redundant visibility distribution for baselines redundant to (1, 12, EE) at frequency channel 513 and LST 5.3965.



(b) Randomly selected redundant visibility distribution for baselines redundant to (1, 12, EE) at frequency channel 537 and LST 5.3843.



(c) Visibility distribution with the lowest HZ test p value (see §2.1) for baselines redundant to (1, 12, EE), which occurs at frequency channel 500 and LST 5.4084.



(d) Visibility distribution with the largest distance between its geometric median and HERA mean for baselines redundant to (1, 12, EE), which occurs at frequency channel 500 and LST 5.4504.

Figure 3: KDE plots for redundant and LST aggregated visibilities at various frequencies and LSTs, to illustrate the distribution of typical HERA data.

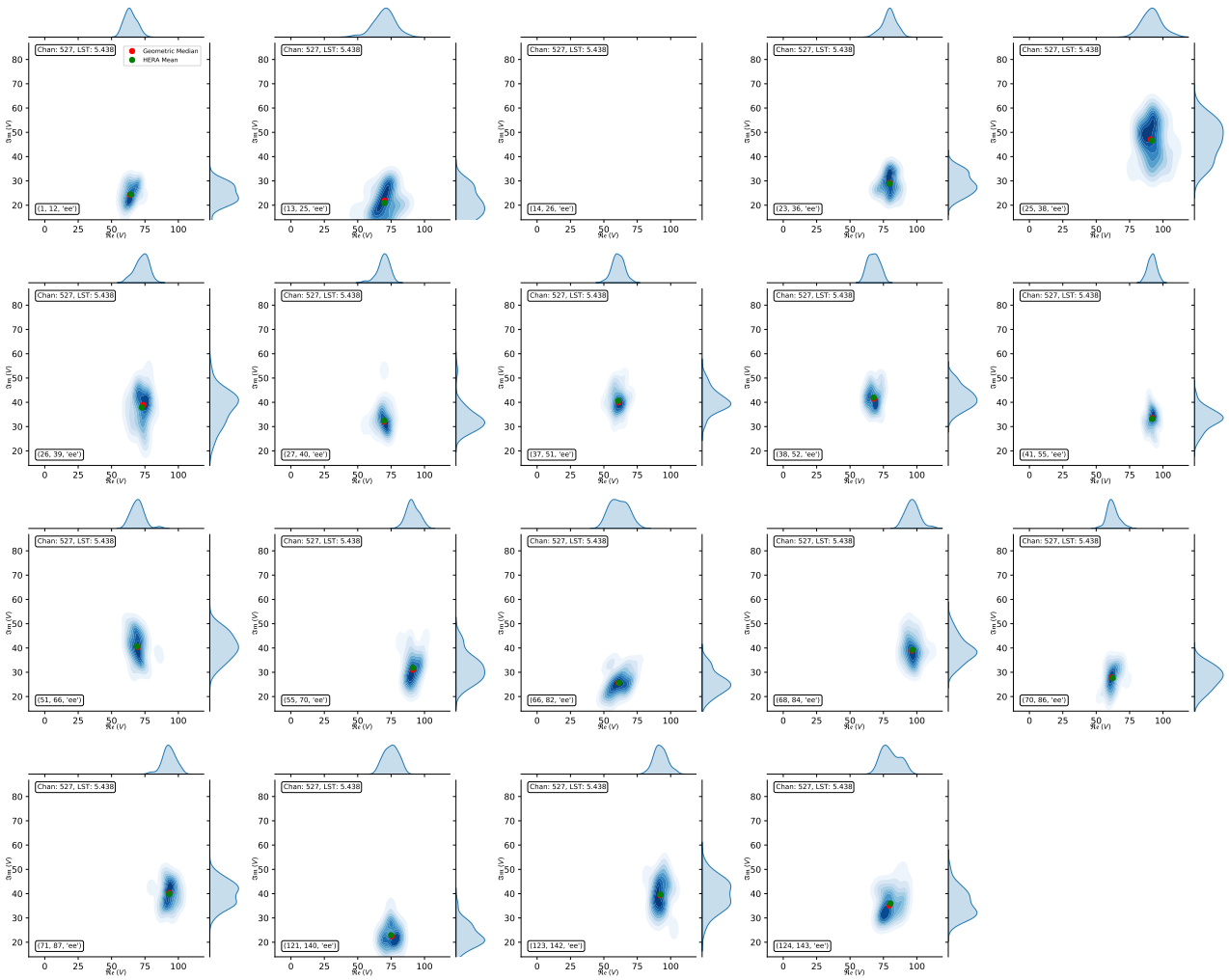


Figure 4: KDE plots for the visibility distribution for each baseline in redundant group (1, 12, EE) aggregated across all JDs at frequency channel 527 and LST 5.4379.

where

$$p = \#\text{ variables} \tag{2}$$

$$\beta = \frac{1}{\sqrt{2}} \left(\frac{n(2p+1)}{4} \right)^{\frac{1}{p+4}} \tag{3}$$

$$D_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j) \tag{4}$$

$$D_i = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \tag{5}$$

with D_i giving the squared Mahalanobis distance of i^{th} observation to the centroid and D_{ij} the Mahalanobis distance between the i^{th} and j^{th} observations, as S is the covariance matrix. If the data is multivariate normal, HZ is approximately log-normally distributed.

We show the HZ statistic, its corresponding p value and whether the distribution is deemed Gaussian or not for our selected HERA visibilities in Fig. 6.

In Fig. 7, we repeat the HZ test but this time using MAD-clipped data. Even with this further outlier rejection step, most of the data still does not appear to be Gaussian.

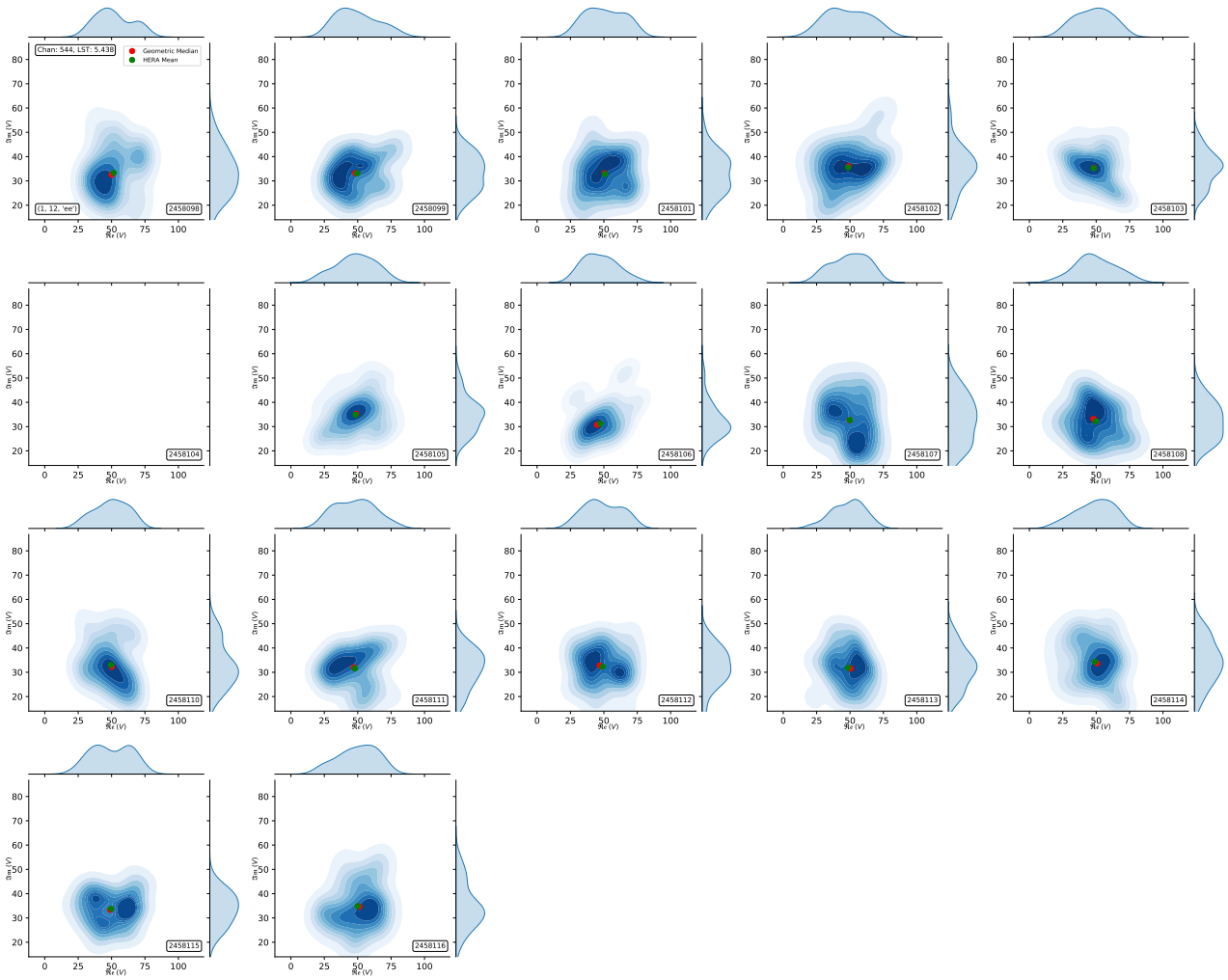


Figure 5: KDE plots for the visibility distribution for each JD of all baselines in redundant group (1, 12, EE) at frequency channel 544 and LST 5.4379.

We note that we also looked at the Shapiro-Wilk test [4] (used for univariate data) across the $\mathfrak{Re}$ and $\mathfrak{Im}$ components and found similar results to those above, with the majority of visibility slices testing non-Gaussian.

3 LST-binning

Following the final calibration and flagging of individual datasets in the HERA analysis pipeline (see e.g. here for that of H1C_IDR2.2), complex visibilities are aggregated and coherently averaged across JDs with the LST-binning function `lst_bin`, in the `lstbin.py` module from the hera_cal package. A LST grid with cadence of 21.4 s (double the integration time) is established to account for the sidereal drift between consecutive JDs. In LST-binning, each integration from individual datasets is assigned to the nearest LST bin, but is also rephased to account for the slight LST difference between its centre and the bin's centre; every LST bin gets two data points for each input night.

In each time bin, a further round of outlier rejection is then performed using MAD-clipping, which rejects samples for every frequency/time/baseline slice that has a modified Z-score $Z_i^{\text{mod}} > 4$, as defined below:

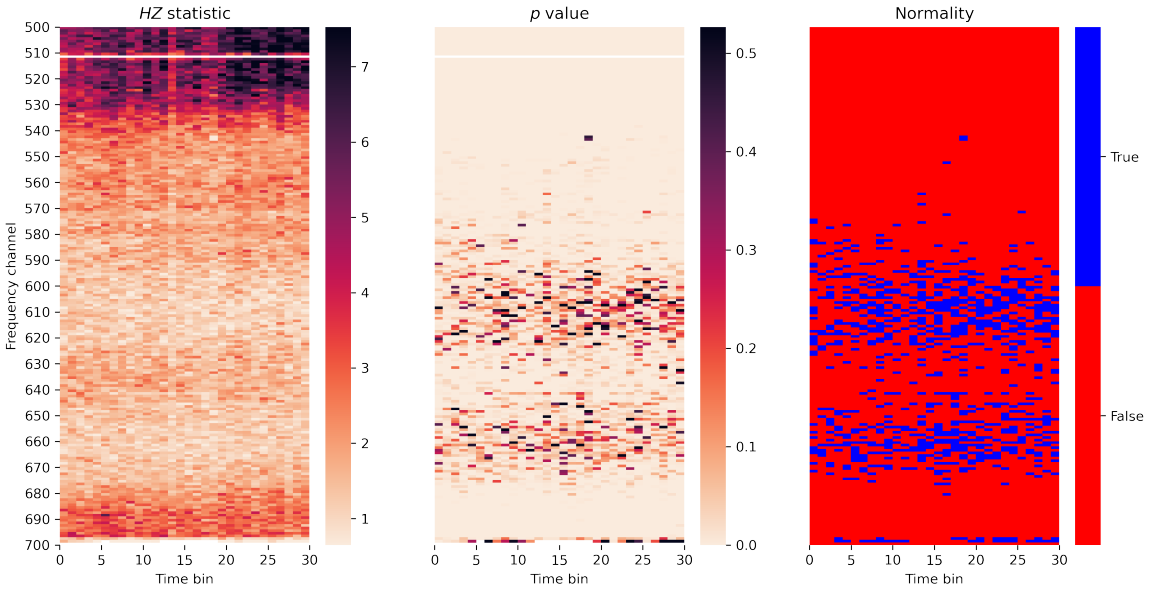


Figure 6: HZ statistic, p value and test outcome for LST-binned and redundantly aggregated HERA visibility data for baseline group (1, 12, EE).

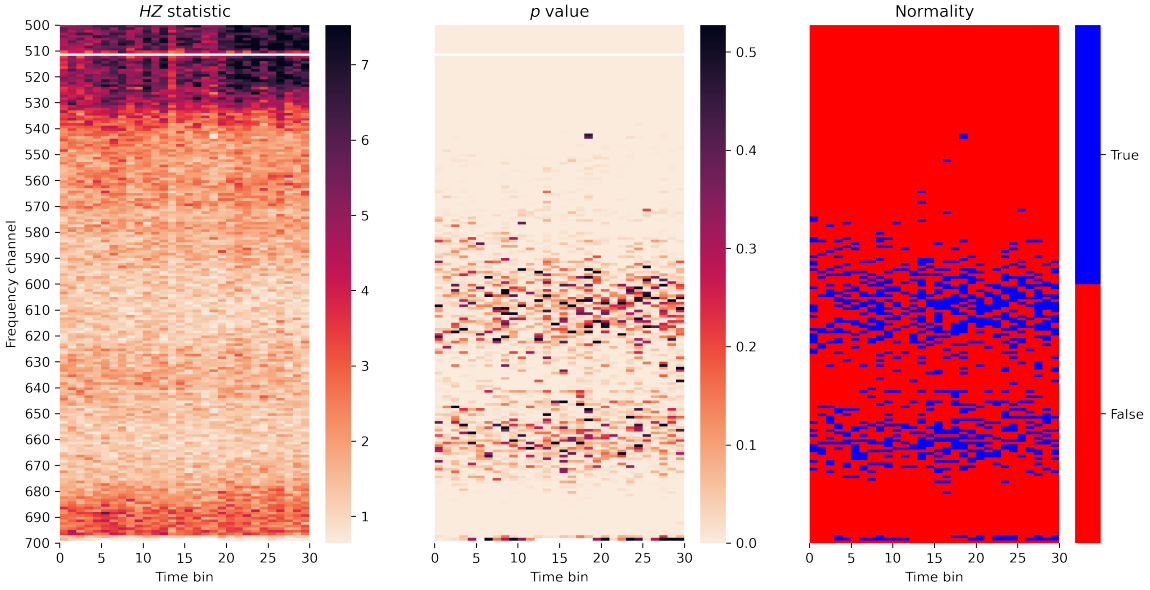


Figure 7: HZ statistic, p value and test outcome for LST-binned and redundantly aggregated HERA visibility data for baseline group (1, 12, EE), post MAD-clipping.

$$Z_i^{\text{mod}} = \frac{x_i - \text{med}(x)}{\sigma^{\text{mad}}} \tag{6}$$

$$\sigma^{\text{mad}} = 1.482 \times \text{med}\left|x - \text{med}(x)\right| \tag{7}$$

where σ^{mad} is the MAD, which is a robust measure of variability that isn't skewed by outliers. The factor of 1.482 in Eq. (7) is a consistency correction that is required to reproduce the standard deviation in the case of white Gaussian noise. This MAD-clipping is computed in `sigma_clip` and is invoked here in LST-binning.

The binned complex visibilities are then mean averaged about their $\mathfrak{Re}$ and $\mathfrak{Im}$ components separately here. We note that there is also the option of computing the marginal median instead too by turning on `median=True` in `lst_bin` here.

This approach of MAD-clipping + mean averaging works well for Gaussian distributed data with low levels of contamination. We have, however, seen in §2 that this is not the case, with the complex visibilities not following a bivariate Gaussian distribution, even after MAD-clipping. We further note that the MAD-clipping threshold is rather arbitrary and difficult to justify.

A more robust statistical approach here could improve results. Furthermore, wholly considering the bivariate data instead of dealing with the $\mathfrak{Re}$ and $\mathfrak{Im}$ components separately would lead to a better location estimate of the complex distribution.

4 Robust multivariate median estimators

4.1 Medians in bivariate and higher dimensional data

The median of a univariate dataset is a natural robust estimator for the centre of the distribution. It can be defined as the order statistic of rank $(n+1)/2$ when n is odd and as the mean of the order statistics of rank $n/2$ and $(n+1)/2$ when n is even (other definitions also exist, see e.g. [5]). It has different characteristics to the mean, chiefly through its breakdown properties: a single infinite point contaminating a data set will send the mean to infinity. By comparison, at least 50% of the data must be moved to infinity to reproduced the same effect in the median; it is said to have a breakdown point of 1/2. The median is thus robust to outliers and is commonly used in nonparametric problems.

Generalizing the median to higher dimensions is not straightforward. While it can be tempting to take the marginal median of multivariate data, that is to take the univariate median across each coordinate separately, this leads to a result that is not necessarily representative of the central tendency of the distribution. The concept of the median relies on ordering. In higher dimensions, there is however no natural concept of *rank* and a marginal median estimate will heavily depend on the choice of axes. For these reasons, alternative methods that wholly consider all coordinate information together are required to produce a better embodiment of the median to higher dimensions.

In this section, we present higher-dimensional analogues of the median that are used in robust and nonparametric data analysis and inference. Such robust statistical techniques, while less influenced by abnormal observations, have higher computational complexity, making them historically less appealing in practice. With the increased computational power now available, together with the dawn of machine learning (ML) libraries, these methods can be efficiently implemented and their use more tractable. Such robust location estimators are particularly relevant in the averaging of large datasets riddled with outliers; this is especially relevant in radio interferometry where RFI and other non-Gaussian effects are prevalent - with limited data, *good* estimates of visibilities are required that are not skewed by

anomalous effects.

4.2 Geometric median

The geometric median, also known as the L_1-median, is defined as the value of the argument y that minimizes the sum of Euclidian distances between y and all points x_i:

$$\mathrm{GM} = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} \|x_i - y\| \tag{8}$$

The geometric median can also be weighted:

$$\mathrm{GM_w} = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} \eta_i \|x_i - y\| \tag{9}$$

with this latter minimization also known as the famous *Weber problem* [6] in location theory, which was initially posed as a transportation cost minimization problem where the best location for a warehouse is to be found that services n customers, with different customers being associated with different transportation costs η_i.

When x_i are not collinear, the cost function $\sum_{i=1}^{n} \|x_i - y\|$ is positive and strictly convex in $\mathbb{R}^d$, and hence the minimum is unique.

An iterative solution technique to Eq. (9) was first proposed by Weiszfeld in 1937 [7], which is given by:

$$y \to y^{t+1} = T(y^t) \tag{10}$$

where

$$T(y) = \left\{ \sum_{x_i \neq y}^{n} \frac{\eta_i}{\|x_i - y\|} \right\}^{-1} \sum_{x_i \neq y}^{n} \frac{\eta_i x_i}{\|x_i - y\|} \tag{11}$$

for steps $t = 0, 1, \ldots$ with $t_0 \in \mathbb{R}^d \backslash \mathcal{X}$, where $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ denotes the set of data points, until a termination criterion is reached.

Since then, various other algorithms have been proposed that improved on computational speed (see e.g. [8]). Even without such methods, we can now easily solve Eq. (9) numerically with common software, which can also be accelerated with ML libraries such as JAX (see e.g. [9] for some worked examples).

4.3 Tukey median

Tukey [10] proposed the *halfspace* depth as a tool to visually describe multivariate datasets. He suggested this depth could be used to define multivariate analogues of univariate rank and order statistics via depth-induced *contours*.

For a finite set of data points $\mathcal{X}_n = \{x_1, ..., x_n\}$ in $\mathbb{R}^d$, the Tukey depth, or halfspace depth of any point $y \in \mathbb{R}^d$ determines how central the point is inside the data cloud; it is defined as the minimal number of data points in any closed halfspace determined by a hyperplane through y:

$$d_T(y; \mathcal{X}_n) = \min_{\|u\|=1} \#\{i \in \{1, \ldots, n\} : u^\intercal x_i \geq u^\intercal y\} \tag{12}$$

where $\|u\| = 1$ denotes the unit sphere in $\mathbb{R}^d$.

The set of all points with depth $\geq \kappa$ is called the Tukey κ^{th} depth region $\mathcal{D}(\kappa)$:

$$\mathcal{D}(\kappa) = \{x \in \mathbb{R} : d_T(x) \geq \kappa\} \tag{13}$$

The halfspace depth regions form a sequence of closed convex polyhedra, with each polyhedron included in $\text{conv}(\mathcal{X}_n)$, making it compact. Tukey regions are also nested: they shrink with increasing κ. An empirical distribution is fully characterized by its Tukey regions.

Donoho & Gasko [11] noted that this notion of depth could be the basis of a definition for a multivariate median: the Tukey median $\bar{\mathcal{T}}$ is defined as the gravity centre of the points that maximize the Tukey depth, known as the Tukey median set $\mathcal{T}$ (the innermost Tukey region). The latter is given by:

$$\mathcal{T} = \arg\max_{y \in \mathbb{R}^d} d_T(y; \mathcal{X}_n) \tag{14}$$

The depth d_T decreases when y moves on a ray originating from the Tukey median $\bar{\mathcal{T}}$; it also vanishes outside the convex hull, $\text{conv}(\mathcal{X}_n)$, of the data. The depth, therefore, provides a centre-outward ordering of points in $\mathbb{R}^d$, thus extending univariate rank to multivariate data.

In Fig. 8, we show the Tukey regions for increasing depth for simulated data using the TukeyRegion R package [12].

While the Tukey median is an intuitive and geometrically appealing way of robustly estimating the location of a distribution, a library that computes it has not yet been implemented in Python. From the plots in §2, it has very similar location estimates to the geometric median; we therefore employ the latter in the subsequent analysis.

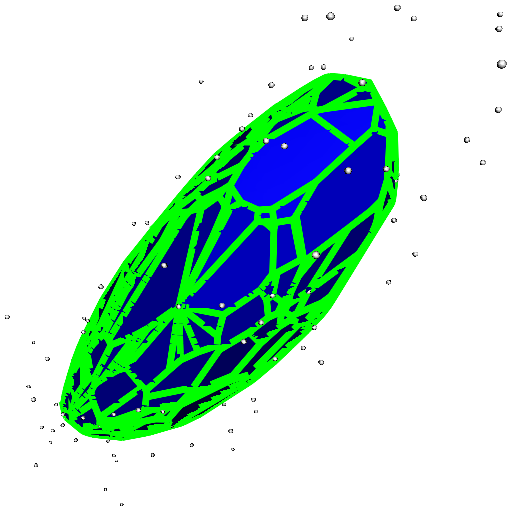
Many other robust multivariate estimators exist; see e.g. [13] for descriptions of other robust multivariate estimators based on the notion of statistical depth.

4.4 Implementation

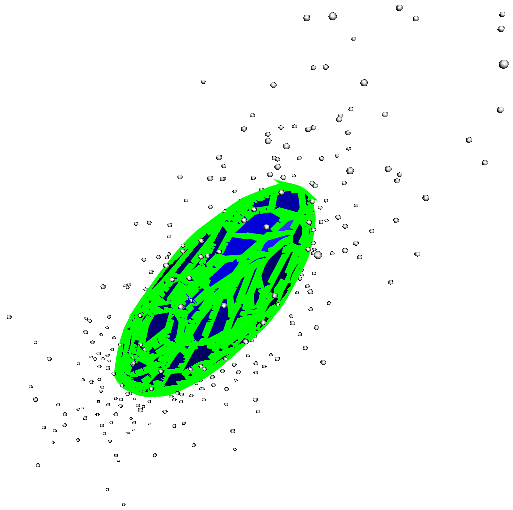
We introduce the robstat Python package, which contains various robust statistical functions to estimate the locations of directional and multivariate data. The package requires an installation of R, the rpy2 Python-R bridge, as well as other standard packages common in data science. R is called upon as it has packages available for the evaluation of depth measures and their associated medians; these are not yet found in Python.

The geometric and Tukey medians can be computed by calling `geometric_median` and `tuky_median` from robstat, respectively. The former is a SciPy minimization of the L_1 cost function (see Eq. (8)) using the BFGS algorithm and the latter calls from the TukeyRegion R package.

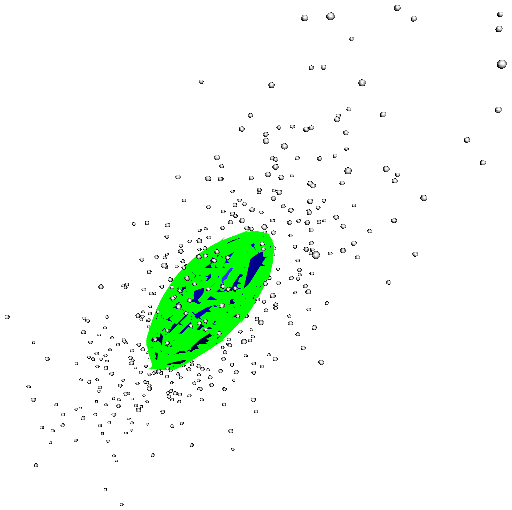
Both of these median estimates are computed in the example below:



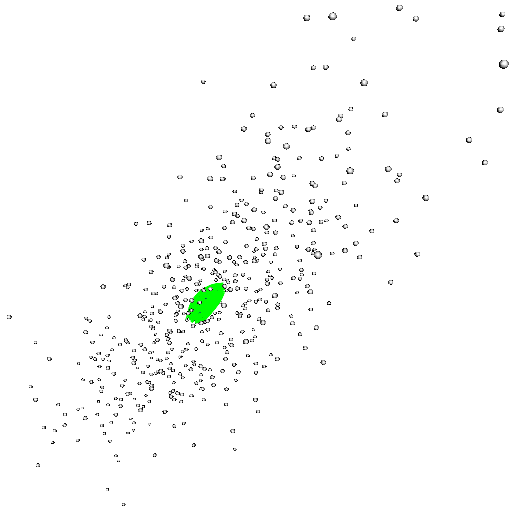
(a) $\mathcal{D}(10)$.



(b) $\mathcal{D}(50)$.



(c) $\mathcal{D}(100)$.



(d) $\mathcal{D}(200)$.

Figure 8: Tukey regions $\mathcal{D}(\kappa)$ for $\kappa = 10, 50, 100, 200$ (left to right, top to bottom) for 3 dimensional Gaussian simulated data with $\mu = [0, 0, 0]$ and covariance matrix $[[1, 1, 1], [1, 2, 2], [1, 2, 4]]$. The Tukey median $\mathcal{T} = [-0.0032, 0.01057, -0.03960]$ is reached at $\kappa = 234$ and is the barycenter of a region with 8 vertices and 12 hypertriangles defining facets.

```
import numpy as np
from robstat.robstat import geometric_median, tukey_median

# generate noisy data with some outliers
np.random.seed(1)
points = np.random.random(500).reshape(-1, 2)*2
points = np.concatenate((points+2, np.random.random(50).reshape(-1, 2)+5))

# compute location estimates
sample_mean = np.mean(points, axis=0)
sample_gmed = geometric_median(points, weights=None)
sample_tmed = tukey_median(points)['barycenter']

med_ests = list(zip(['Mean', 'Geometric median', 'Tukey Median'], \
                    [sample_mean, sample_gmed, sample_tmed]))
for med_est in med_ests:
    print('{:16s}: {}'.format(med_est[0], med_est[1]))

# Mean            : [3.21828555 3.26175702]
# Geometric median: [3.06956999 3.15071265]
# Tukey Median    : [3.09257657 3.16488129]
```

These results are also plotted in Fig. 9.

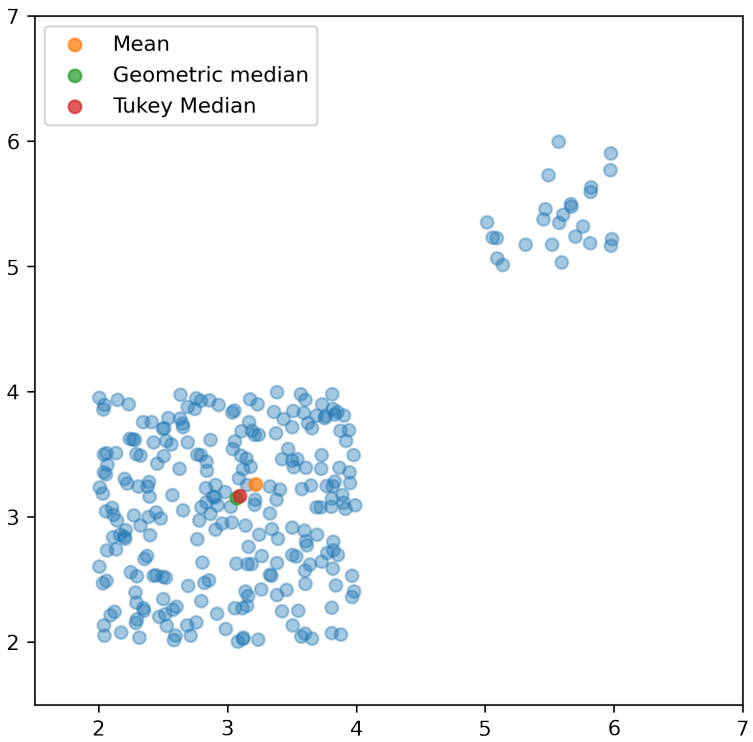


Figure 9: Location estimates for noisy data with outliers. The geometric and Tukey median are not skewed by outliers, unlike the mean, and are at roughly the same location.

5 Results

In this results section, we apply the geometric median estimator to HERA visibilities to find robust location estimates of the complex data. We also compute simple PS of the resulting location estimates.

We reproduce the averaging in the LST-binning pipeline by first MAD-clipping the data with a 4σ cut-off followed by taking the mean (see [1] for further details). This can be done with `rsc_mean` in the `stdstat.py` robstat module. We call this method *HERA mean* in below plots, and use this as the benchmark for comparison with other location estimates.

5.1 LST + redundant averaging

The visibility estimates from averaging both in LST and across redundant baselines with various robust and standard estimators are shown in Fig. 10. By eye, all these estimates seem consistent with each other and have the same overall shape - the statistics of these visibility estimates will enable us to better compare them.

As an aside, we note that the estimates for the marginal median are considerably worse when the data is not rephased, with some adjacent time or frequency channels becoming disjointed for the amplitude, phase and $\mathfrak{Im}$ plots, whereas the other estimators perform adequately (not shown here).

5.1.1 Location estimate smoothness

We look at the smoothness for the location estimates shown in Fig. 10 by calculating the standard deviation of the distances between successive points in either frequency or time:

$$\mathcal{S} = \sqrt{\frac{\sum_i^D \left(d_i - \bar{d}\right)^2}{N_D}} \tag{15}$$

where

$$d_i = x_{i+1} - x_i \tag{16}$$

We then take the mean of Eq. (15) over the other dimension to obtain a single quantity.

These smoothness proxies are shown in Table 1. The HERA mean provides the smoothest visibility estimates, with the geometric median in second place.

5.2 Power spectrum estimation

Without resorting to a full PS analysis using the HERA PS pipeline, as an intermediate result, we compute the simple PS using a periodogram at each time bin for the geometric median and HERA mean location estimates of the LST + redundantly grouped visibilities shown in §5.1. In order to compute the PS, we estimate the location for channel 511, which is flagged throughout, using 2D cubic interpolation. We also discard channel 699 at this stage, as it is also flagged and sits at the end of the data array. We plot the respective PS, as well as the comparison of their means (in time), in Fig. 11.

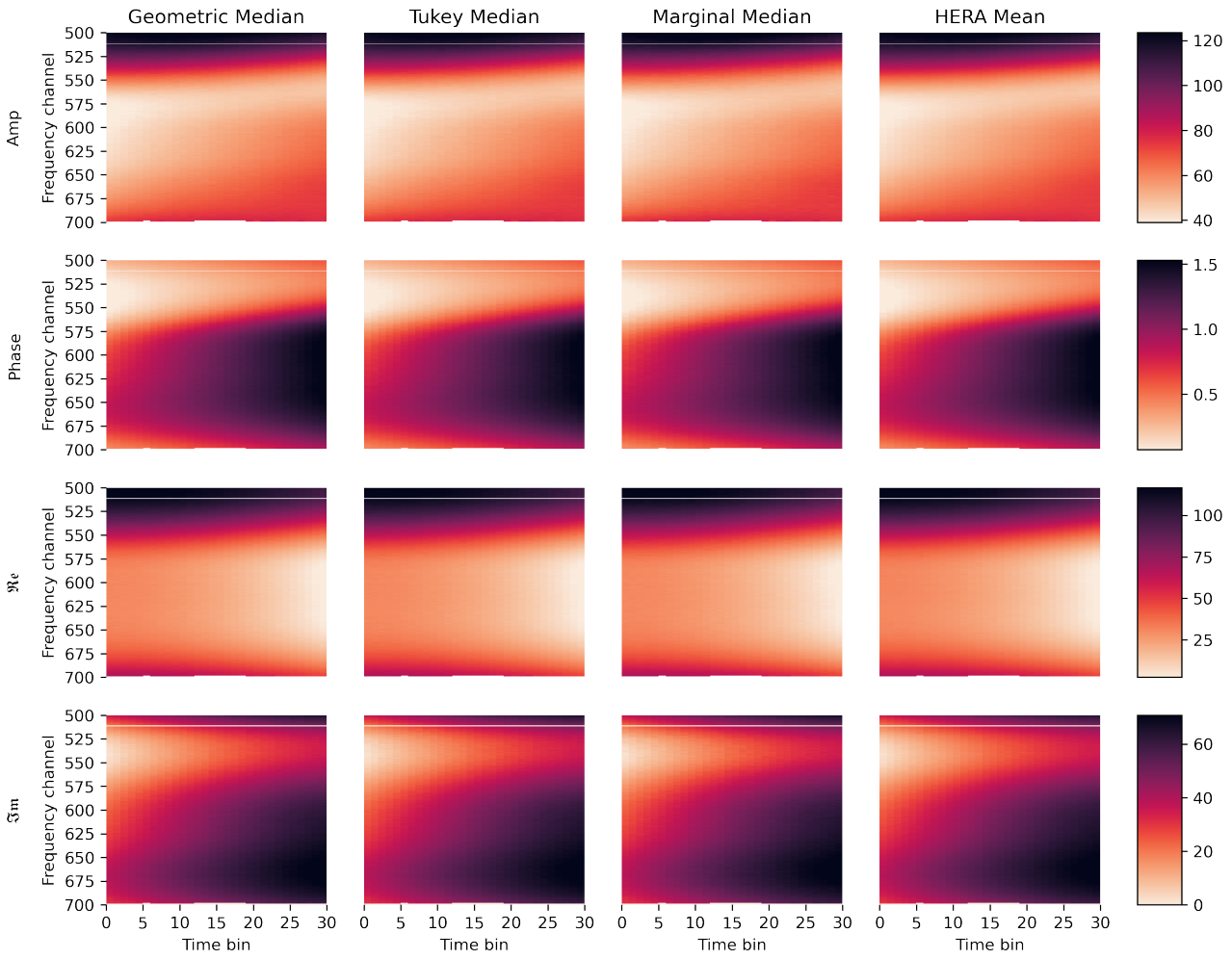


Figure 10: Location estimates for the complex visibility, with amplitude, phase and $\mathfrak{Re}$ and $\mathfrak{Im}$ components shown.

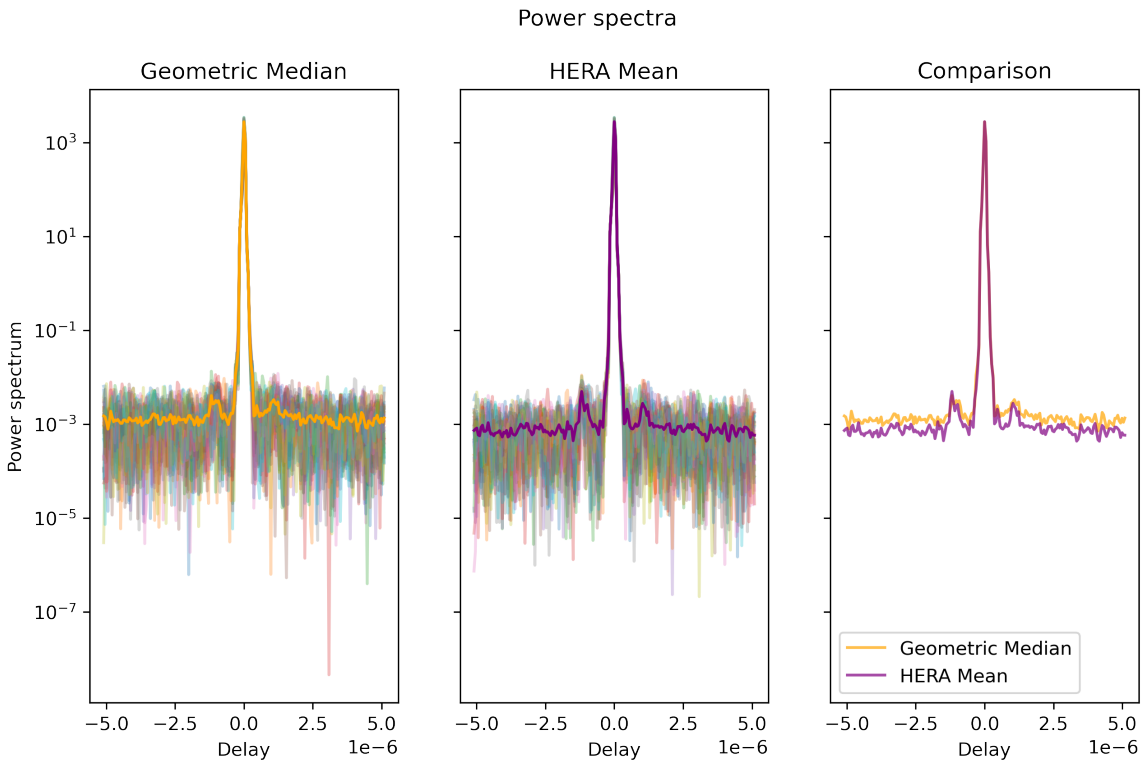


Figure 11: PS of geometric median and HERA mean estimates of LST + redundantly grouped visibilities for the 30 time bins between LSTs 5.2824 to 5.4559. The mean PS over time are also shown and compared, with the HERA mean PS having a lower noise floor than its geometric median counterpart.

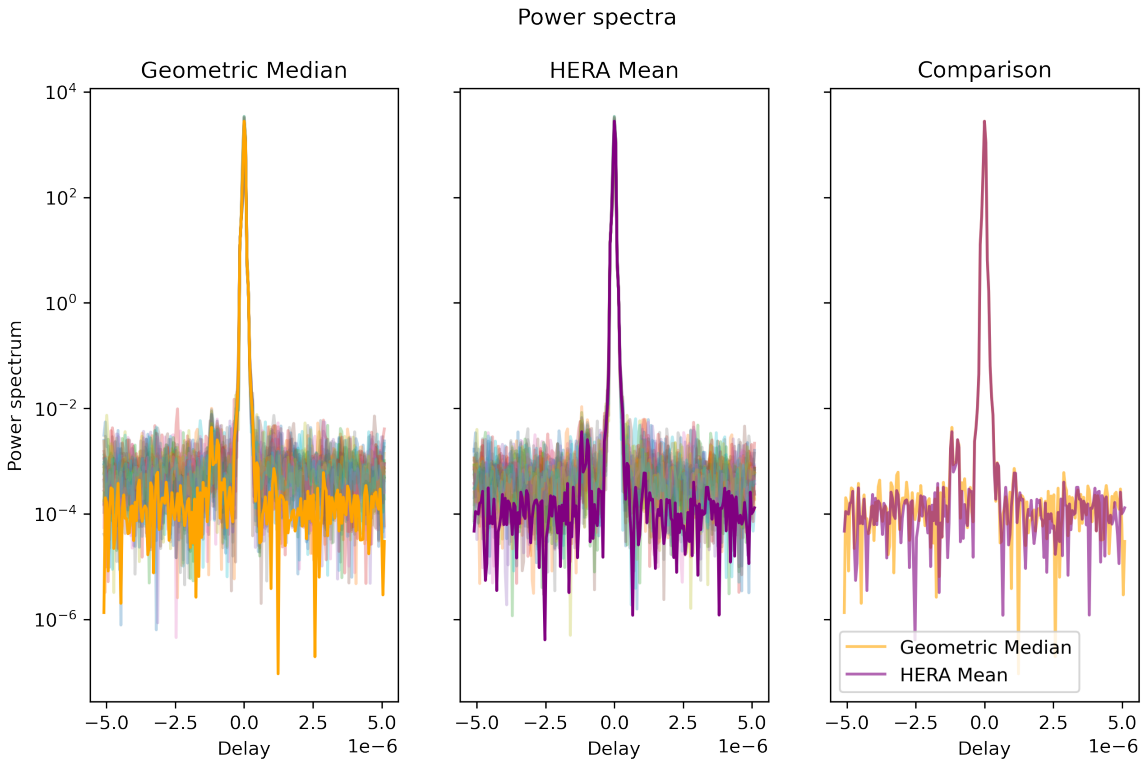


Figure 12: Absolute value of the cross-PS averaged across all baseline permutations for the geometric median and HERA mean estimates of LST-binned visibilities. The absolute value of the mean of all cross-PS over time are also shown and compared.

	Difference measure	Geometric Median	Tukey Median	Marginal Median	HERA Mean
$\bar{\mathcal{S}}_t$	$d_i = \lvert x_{i+1} - x_i \rvert \;\; d_i \in \mathbb{R}$	0.4400	0.4639	0.5190	0.3494
$\bar{\mathcal{S}}_\nu$		0.6155	0.6373	0.6722	0.5700
$\bar{\mathcal{S}}_t$	$d_i = x_{i+1} - x_i; \;\; d_i \in \mathbb{C}$	0.8268	0.8605	0.9477	0.7238
$\bar{\mathcal{S}}_\nu$		1.1738	1.2005	1.2686	1.0929

Table 1: The mean smoothness for consecutive location estimates of redundant visibilities in time $\bar{\mathcal{S}}_t$ and frequency $\bar{\mathcal{S}}_f$, with the difference measure being the absolute value of consecutive values in the first two rows, while the last two rows compute Eq. (15) while keeping the complex differences between consecutive values. A reminder that $\mathrm{Var}[z] = \mathrm{Var}[\mathfrak{Re}(z)] + \mathrm{Var}[\mathfrak{Im}(z)]; \;\; z \in \mathbb{C}$ - therefore, strictly speaking, the last two rows are the more *correct* smoothness values. It is still, however, interesting to compute the smoothness of the distances (absolute values) to see if they are consistent with the ordering in the last two rows.

From Fig. 11, the HERA mean approach performs better than its robust statistical counterpart. This may be because the HERA mean smoothes out the signal too much, especially considering that the visibility distribution is non-Gaussian; a clipping + mean estimator will tend to compute a location that is more smooth between frequencies (and other axes), whereas a robust estimator will *choose* between peaks of a multi-modal distribution that slightly vary in strength, thus resulting in a more amplified and varied signal.

As an alternative approach that follows the steps of the HERA PS pipeline, we run location estimates on the visibilities aggregated in LST and compute the (complex) cross-PS across all baseline permutations. We then mean average these cross-PS in time, producing Fig. 12. The PS in Fig. 12 have lower noise floors by an order of magnitude than those in Fig. 11, which is expected as the cross-PS suppresses noise that is specific to individual baselines. These results are more encouraging, and the geometric median performs just as well as the HERA mean. As more baselines and times are considered, it is hoped that the robust estimator would outperform the HERA mean.

We note that this entire analysis was repeated for frequency channels $175 - 333$ (Band 1) and LSTs $9.5880 - 9.7644$ (a subset of Field 3) as a consistency check, with very similar results obtained.

6 Conclusion & future work

We have visualized and shown that the aggregated underlying HERA H1C_IDR2.2 visibility data is mostly non-Gaussian, with various unknown effects often causing the distribution to be multi-modal. We introduced the geometric median, a robust multivariate location estimator, and applied it to the aggregated visibilities in order to improve on the location estimates for these non-normal data. We found that in general, the geometric median seems to better find the central location of the data compared to the clipping + mean approach currently used in the LST-binning pipeline.

In the visibility KDE plots and videos, we notice that the spread of data between redundant baselines is higher than that between JDs, possibly due to apparent miscalibration of data. Direction-dependent effects could be at play, here. It appears that even after the full analysis pipeline, different redundant baselines are still seeing different skies. Even robust estimators may not work quite as desired due to the multi-modality observed in this case. We refer to the robust redundant calibration detailed in [9] as a method of potentially better reconciling redundant baselines.

When looking at the PS of the visibility estimates, the current HERA mean approach appears *better* in the (auto-)PS of Fig. 11, while the geometric median performs just as well in the cross-PS of Fig. 12. Implementing the geometric median estimator in the LST-binning pipeline and pushing it through for the whole of H1C_IDR2.2 is required for a full, comprehensive analysis. Proper PS estimates with the PS pipeline also need to be conducted, which could see the geometric median improve results and/or could reveal parts of the data that are anomalous (i.e. if the PS from the geometric median estimates performs much better). As an ad hoc approach, the HERA mean could also be artificially suppressing noise and affecting any potential signal, which needs further investigating.

As a reminder, we only presented results for one redundant group over a small selected frequency and time range. While a full analysis of the normality of visibilities would be needed for all baselines, LSTs, frequencies, polarizations and for different time bin sizes, similar effects were also observed in other parts of the H1C_IDR2.2 dataset, consistent with the findings of this memorandum.

Further work includes using a high pass filter in delay space (such as `fourier_filter`) to remove the low order modes of the data, and then transforming back into the visibility domain (i.e. effectively working on the `.OCRSD.uvh5` files). This would make it easier to robustly average over redundant baselines that are otherwise inconsistent, and would also provide better insight into the higher-order effects seen across redundant visibilities. This is currently being looked into.

Should the collaboration decide to continue with the clipping + mean approach, we would suggest performing the outlier rejection with a robust multivariate method, such as one based on robust Mahalanobis distances (see e.g. [14], we also present an implementation here). Simulations of the sky signal (HI + foregrounds + systematics + Gaussian noise) should also be examined to see how MAD-clipping and other outlier rejection techniques influences signal recovery.

References

[1] The HERA Collaboration, Z. Abdurashidova, J. E. Aguirre, P. Alexander, Z. S. Ali, Y. Balfour, A. P. Beardsley, G. Bernardi, T. S. Billings, J. D. Bowman *et al.*, "First Results from HERA Phase I: Upper Limits on the Epoch of Reionization 21 cm Power Spectrum", *arXiv e-prints*, p. arXiv:2108.02263, Aug. 2021.

[2] J. S. Dillon, "HERA Memorandum #69: H1C IDR 2.2", July 2019.

[3] N. Henze and B. Zirkler, "A class of invariant consistent tests for multivariate normality", *Communications in Statistics - Theory and Methods*, vol. 19, pp. 3595–3617, 1990.

[4] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)", *Biometrika*, vol. 52, pp. 591–611, 1965.

[5] H. Oja, *Multivariate Median*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 3–15.

[6] A. Weber, *Über den Standort der Industrien*. University of Chicago Press, 1909.

[7] E. Weiszfeld, "Sur le Point pour Lequel la Somme des Distances de n Points Donnés Est Minimum", *Tohoku Mathematical Journal, First Series*, vol. 43, pp. 355–386, 1937.

[8] M. B. Cohen, Y. Tat Lee, G. Miller, J. Pachocki, and A. Sidford, "Geometric Median in Nearly Linear Time", *arXiv e-prints*, p. arXiv:1606.05225, Jun. 2016.

[9] M. Molnar and B. Nikolic, "HERA Memorandum #84: A Generalized Approach to Redundant Calibration with JAX", November 2020.

[10] J. W. Tukey, "Mathematics and the Picturing of Data", in *Proceedings of the International Congress of Mathematicians, Vancouver*, vol. 2, 1975, pp. 523–531.

[11] D. L. Donoho and M. Gasko, "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness", *The Annals of Statistics*, vol. 20, pp. 1803–1827, 1992.

[12] X. Liu, K. Mosler, and P. Mozharovskyi, "Fast Computation of Tukey Trimmed Regions and Median in Dimension $p > 2$", *Journal of Computational and Graphical Statistics*, vol. 28, pp. 682–697, 2019.

[13] P. Rousseeuw and A. Struyf, *Handbook of Discrete and Computational Geometry - Chapter 58 - Computation of Robust Statistics: Depth, Median, and Related Measures*, J. E. Goodman, J. O'Rourke, and C. D. Tóth, Eds. Chapman and Hall / CRC, 2017.

[14] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points", *Journal of the American Statistical Association*, vol. 85, pp. 633–639, 1990.