HERA Memo #110
Received 13 Feb 2022

# Multivariate Outlier Detection Using Robust Mahalanobis Distances

Matyas Molnar and Bojan Nikolic

Astrophysics Group, Cavendish Laboratory, University of Cambridge

31st January 2022

**Abstract**

Robust outlier rejection is required for the reduction of radio interferometric data. Following the Hydrogen Epoch of Reionization Array (HERA) analysis pipeline, visibilities across Julian dates (JDs) are aligned in Local Sidereal Time (LST), rephased and placed into bins of 21.4 s cadence. At this stage, before any averaging is performed, a round of outlier detection that uses median absolute deviation (MAD)-clipping is conducted. We present an improved outlier rejection routine that considers robust Mahalanobis distances calculated with Minimum Covariance Determinant (MCD) location and covariance estimates. The flagging capabilities of these two methods are compared and preliminary averaged visibility and power spectrum (PS) results are shown.

## 1 Introduction

In the HERA LST-binning pipeline, a further round of outlier rejection is performed before the fully calibrated visibilities are averaged across JDs. This is to ensure that any missed radio-frequency interference (RFI) and/or other problems with the data that do not repeat night-to-night at the same LST are flagged.

In each aggregated time bin, MAD-clipping is used to reject samples for every frequency/time/baseline slice that has a modified $Z$-score $|Z_i^{\text{mod}}| > 5$, as defined below:

$$Z_i^{\text{mod}} = \frac{x_i - \text{med}(x)}{\sigma^{\text{mad}}} \tag{1}$$

$$\sigma^{\text{mad}} = 1.482 \times \text{med}\,\big|x - \text{med}(x)\big| \tag{2}$$

where $\sigma^{\text{mad}}$ is the MAD, a robust measure of variability that isn't skewed by outliers. The factor of 1.482 in Eq. (2) is a consistency correction that is required to reproduce the standard deviation in the case of white Gaussian noise.

This MAD-clipping is performed on the $\mathfrak{Re}$ and $\mathfrak{Im}$ components of the visibilities separately, and data points are flagged if either component is clipped. Through this method, it is implied that the location of the visibility distribution is given by the marginal median. As discussed in [1], despite having a high breakdown value of 1/2, the marginal median does not necessarily represent the central tendency of the distribution as it is not affine equivariant: for a location estimator $\Lambda$ to be affine equivariant, it must transform properly under rotation of the data, as well as changes in location and scale. That is to say, for a $p$-by-$p$ nonsingular matrix $\mathbf{A}$ and vector $\mathbf{b}$ with length $p$:

$$\Lambda(X_1\mathbf{A} + \mathbf{b}, \dots, X_n\mathbf{A} + \mathbf{b}) = \Lambda(X_1, \dots, X_n)\mathbf{A} + \mathbf{b} \tag{3}$$

where $X_1, \dots, X_n$ is a sample from a $p$-variate distribution and each $X_i$ is a vector of length $p$.

Similarly, $Z_{\mathfrak{Re}}^{\text{mod}}$ and $Z_{\mathfrak{Im}}^{\text{mod}}$ only represent the spread of the data along the $\mathfrak{Re}$ and $\mathfrak{Im}$ axes, resulting in a rectangular boundary that does not account for the covariance of the visibility distribution.

In this memorandum, we present another clipping method that uses robust Mahalanobis distances, which overcomes some of the shortfalls of MAD-clipping. The aim is not only to present a theoretically better method of outlier rejection, but to also potentially address issues in the results that may be due to low-level RFI that has not been picked up by the current pipeline (cf. the Band 2 Field 2 results in [2]). This work is separate but concurrent to the robust location estimates for LST-binning and averaging that has been presented in [1].

# 2 Outlier detection with robust Mahalanobis distances

The Mahalanobis distance [3] is a multivariate distance metric that measures the distance between a point and a distribution. It is given by

$$\text{MD}(x_i) = \Big((x_i - \hat{\mu})\hat{\Sigma}^{-1}(x_i - \hat{\mu})^{\intercal}\Big)^{1/2} \tag{4}$$

where $\hat{\mu}$ is the sample multivariate mean and $\hat{\Sigma}$ is the sample covariance matrix. Unlike Euclidean distances, it accounts for any correlation between variables. It is commonly used to find outliers in multivariate sets.

Naturally, the mean and covariance will be heavily influenced by the presence of outliers; obtaining good robust estimators of $\hat{\mu}$ and $\hat{\Sigma}$ are necessary to measure the *outlyingness* of data points and to have a proper distance-based outlier detection procedure. Therefore, we modify Eq. (4) to get robust Mahalanobis distances:

$$\text{RMD}(x_i) = \Big((x_i - \hat{\mu}_r)\hat{\Sigma}_r^{-1}(x_i - \hat{\mu}_r)^{\intercal}\Big)^{1/2} \tag{5}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ have been replaced with $\hat{\mu}_r$ and $\hat{\Sigma}_r$, which are robust estimators of centrality and covariance matrix.

In practice, the most frequently used covariance estimator is the MCD estimator [4], which is based on the computation of the ellipsoid with the smallest volume or with the smallest covariance determinant that would encompass at least half of the data points.

## 2.1 Minimum Covariance Determinant estimator

This MCD estimator is a high-breakdown and affine equivariant estimator of both location and scatter. It consists of determining the subset $J$ of observations of size $h$ that minimizes the determinant of the sample covariance matrix, computed from only these $h$ *good* observations, which are not considered to be outliers. The choice of $h$ (also called the tuning constant) determines the robustness of the estimator; it is a compromise between robustness and efficiency. Once this subset of size $h$ is found, it is possible to estimate the centrality and the covariance matrix based only upon that subset.

More formally, $J$ is defined as

$$J = \Big\{ h : |\hat{\Sigma}_J| < |\hat{\Sigma}_K| \; \forall K \text{ s.t. } \#K = h \Big\} \tag{6}$$

with $(n + d + 1)/2 \leq h \leq n$ for an $n \times d$ data matrix, and where $||$ denotes the determinant of the matrix, and $\#K$ denotes the cardinality of the subset $K$. The location and scatter are then estimated to be

$$\hat{\mu}_{\text{MCD}} = \frac{1}{h} \sum_{i \in J} x_i \tag{7}$$

$$\hat{\Sigma}_{\text{MCD}} = \frac{1}{h} \sum_{i \in J} (x_i - \hat{\mu}_{\text{MCD}})(x_i - \hat{\mu}_{\text{MCD}})^{\intercal} \tag{8}$$

The tuning constant $h$ is generally taken to be its minimum value of $(n + d + 1)/2$ to maximize the robustness of the MCD estimator.

The computation of the exact MCD estimator is very demanding, as it requires the evaluation of $\binom{n}{h}$ subsets of size $h$. The FAST-MCD algorithm [5] is computationally efficient and allows the MCD estimator to be applied to large datasets; it involves the key *C-step*, which considers a selected set of $h$-subsets, starting from random subsets of size $p + 1$.

If the outlier detection algorithm is applied to data for the same baselines but on different JDs (as is done in the LST-binning pipeline), then the data should mostly be near-normal enough for MCD to work adequately. However, MCD should not be used for multimodal distributions or those that deviate too far from Gaussianity; as we showed in [1], there is evidence that distinct but redundant HERA baselines may fall under this category.

### 2.2 Robust Mahalanobis distance clipping

For robust outlier detection, we require $\text{RMD}(x_i) > c_d$ for some $c_d$ threshold. We note that $\text{RMD}(x_i)$ approximately follows a $\chi^2$ distribution [6], hence we can use

$$c_d = \sqrt{\chi^2_{d;\,1-\alpha}} \tag{9}$$

that corresponds to the square root of the upper $\alpha$-quantile of the $\chi^2$ distribution with $d$ degrees of freedom.

We therefore outline steps for finding outliers through these MCD Mahalanobis distances, which we call robust Mahalanobis distance (RMD)-clipping:

1. Compute $\text{RMD}(x_i)$ using FAST-MCD with $h = (n + d + 1)/2$
2. Compute the $p$% quantile $Q$ of the chi-square distribution $\chi^2_{d;\,p}$ ($p$ usually taken to be 0.975, 0.99, 0.999 etc.)
3. Declare $\text{RMD}(x_i) > Q$ as possible outliers

We note that the threshold $Q$ for this method can be modified such that it is adjusted to the sample size; an adjusted quantile $AQ$ can be used instead (see e.g. [7]) in step 2. Proceeding with the $AQ$ option generally improves the false classification rates, while maintaining the same correct classification rates [8]. The adjusted threshold is computed by comparing the theoretical cumulative $\chi^2_d$ distribution function and the empirical cumulative distribution function of the squared robust distance samples, and finding the supremum of the difference between the two tails of these distributions.

### 2.3 Relating quantiles between the $\mathcal{N}$ and $\chi^2$ distributions

When dealing with outlier detection procedures, we commonly set the rejection threshold in terms of multiples of $\sigma$, the standard deviation of the normal distribution. From Eq. (9), the threshold for outlier rejection in RMD-clipping is given in terms of $\chi^2$ quantiles. We therefore need to relate the quantiles of the $\chi^2$ distribution $\chi^2_{d;q}$ to those of a Gaussian distribution $z_q$.

Given a number of standard deviations $n$, the probability $p$ that a normal deviate lies in the range between $\mu - n\sigma$ and $\mu + n\sigma$ is given by

$$p = F_{\mathcal{N}}(\mu + n\sigma) = \Phi(n) - \Phi(-n) = \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{10}$$

where $F_{\mathcal{N}}$ is the cumulative distribution function (CDF) for a generic normal distribution, $\Phi$ the CDF for the standard normal distribution and erf is the error function, which are all related through

$$F_{\mathcal{N}}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{11}$$

To retrieve $\chi^2_{d;\,p}$ we apply the quantile function $F^{-1}_{\chi^2}$ (i.e. the inverse of the CDF) to $p$:

$$\chi^2_{d;q} = F^{-1}_{\chi^2}(p; d) \tag{12}$$

$F^{-1}_{\chi^2}$ does not have a simple, closed-form representation. Its CDF can, however, be given in terms of complete $\Gamma$ and lower incomplete $\gamma$ gamma functions:

$$F_{\chi^2}(x; d) = \frac{\gamma\left(\frac{d}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{13}$$

Using Eq. (12), we numerically compute the $\chi^2_{d;q}$ quantile equivalents for $z_q = 4$ and $z_q = 5$ (corresponding to a $4\sigma$ and $5\sigma$ threshold) to be 19.334 and 28.744, respectively.

We note the following approximation [9] to Eq. (12), which bypasses the complexities of working with $F^{-1}_{\chi^2}$ and works well for large $d$:

$$\chi^2_{d;q} \approx \frac{1}{2}\left(z_q + \sqrt{2d-1}\right)^2 \tag{14}$$

# 3 LST-binning results & comparison to MAD-clipping

## 3.1 Illustrative example

To better visualize how MAD and RMD-clipping work and differ, we look at sample visibilities across the 18 nights of H1C_IDR2.2 for the 14 m baseline (55, 71, EE) at frequency channel 514 that fall into the LST bin centred at 5.5902 with cadence 21.4 s (meaning 2 data points for each JD). We then perform outlier rejection with both the MAD and RMD-clipping procedures, with clipping threshold at the $5\sigma$ threshold. We also only perform the clipping on each slice if there are at least 5 unflagged data points.

In Fig. 1, we show the scatter of selected data points as well as the MAD-clipping boundary and outlier region. We show the same for RMD-clipping in Fig. 2 with concentric Mahalanobis distance contours also marked.

We compare the MAD and RMD boundaries in Fig. 3, and also show an indicative ellipse with width $2 \times 5\sigma_{\mathfrak{Re}}^{\mathrm{mad}}$ and height $2 \times 5\sigma_{\mathfrak{Im}}^{\mathrm{mad}}$ to show a somewhat halfway house between the two methods. As seen in Fig. 3 and typical of other data slices, the RMD boundary is tighter and more closely confines the distribution of the data compared to MAD-clipping. The former method is also seen to generally reject more data points.

The estimated MCD covariance computed for the RMD-clipping in this example is given by the following matrix:

$$C = \begin{bmatrix} 6.5727 & -2.0061 \\ -2.0061 & 3.5462 \end{bmatrix} \tag{15}$$

From the eigendecomposition of Eq. (15), we find the eigenvectors and eigenvalues to be

$$\vec{v}_1 = \begin{bmatrix} 0.8950 \\ 0.4460 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} -0.4460 \\ 0.8950 \end{bmatrix} \tag{16}$$

$$\lambda_1 = 7.5723, \quad \lambda_2 = 2.5466 \tag{17}$$

These can be used to draw the covariance error ellipse that sets the boundary for outliers. The width and height of the ellipse are given by $w = \chi_q \sqrt{\lambda_1}$ and $h = \chi_q \sqrt{\lambda_2}$, with the orientation given by $\alpha = \mathrm{arctan2}(\lambda_1 - C_{0,0}, C_{0,1})$, where $C_{i,j}$ denotes the entry of the covariance matrix at the $i^{\mathrm{th}}$ row and $j^{\mathrm{th}}$ column.

In Fig. 4, we show a situation where RMD-clipping rejects a significant number of data points that would otherwise be well-within the MAD-clipping boundary. This data slice is for baseline (124, 143, EE) at channel 201 and LST 5.6321. Even by eye, it is difficult to judge if the points on the outside of the RMD boundary are outliers. As RMD-clipping already rejects a much higher proportion of data points compared to MAD-clipping (for equivalent quantile thresholds), the threshold for RMD-clipping could be lowered to reduce the average number of outliers it picks out. As mentioned in §2.1, the MCD estimator is likely to perform poorly for non-Gaussian data, which will affect a small proportion of data slices. However, even in such circumstances, it seems that the resulting location of the clipped distribution still represents the central tendency of the data.

The MCD-clipping procedure implies a covariance to the distribution of data; for perfect white noise, we do not expect any off-diagonal elements to the covariance matrix. However, certain calibration

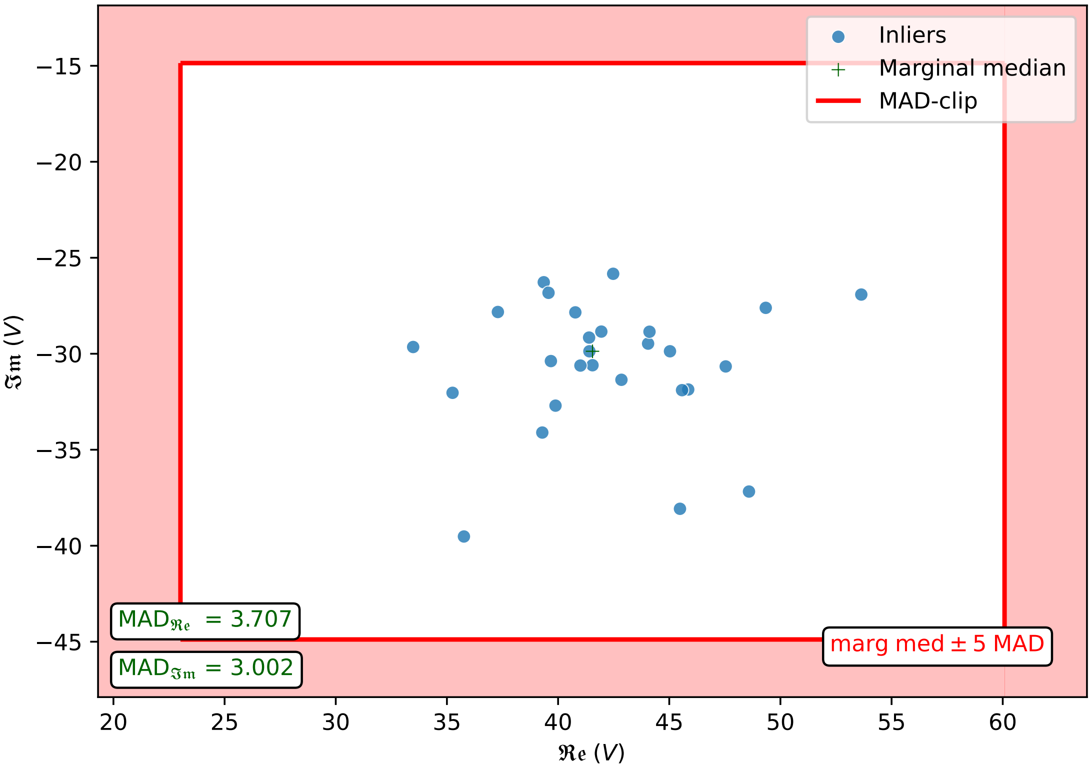

Figure 1: Outlier detection with MAD-clipping, where the red rectangle shows the boundary with width $2 \times 5\sigma^{\text{mad}}_{\mathfrak{Re}}$ and height $2 \times 5\sigma^{\text{mad}}_{\mathfrak{Im}}$. No data points are flagged as outliers. The marginal median is also shown, which represents the location of the distribution according to this particular method.

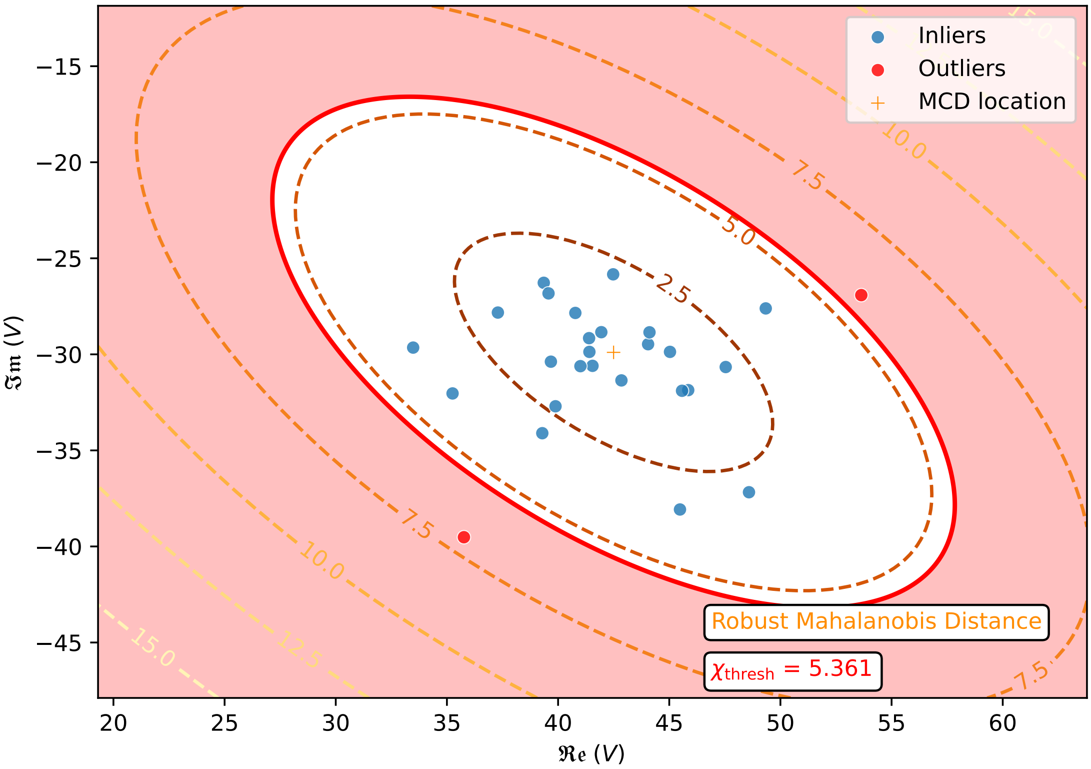

Figure 2: Outlier detection with RMD-clipping, with Mahalanobis distance contours shown and the red ellipse demarcating the inlier/outlier boundary, corresponding to the contour with $\text{RMD}(x) = \chi_{\text{thresh}} = 5.361$. Two points are flagged in this case. The location returned from the MCD estimator is also marked.

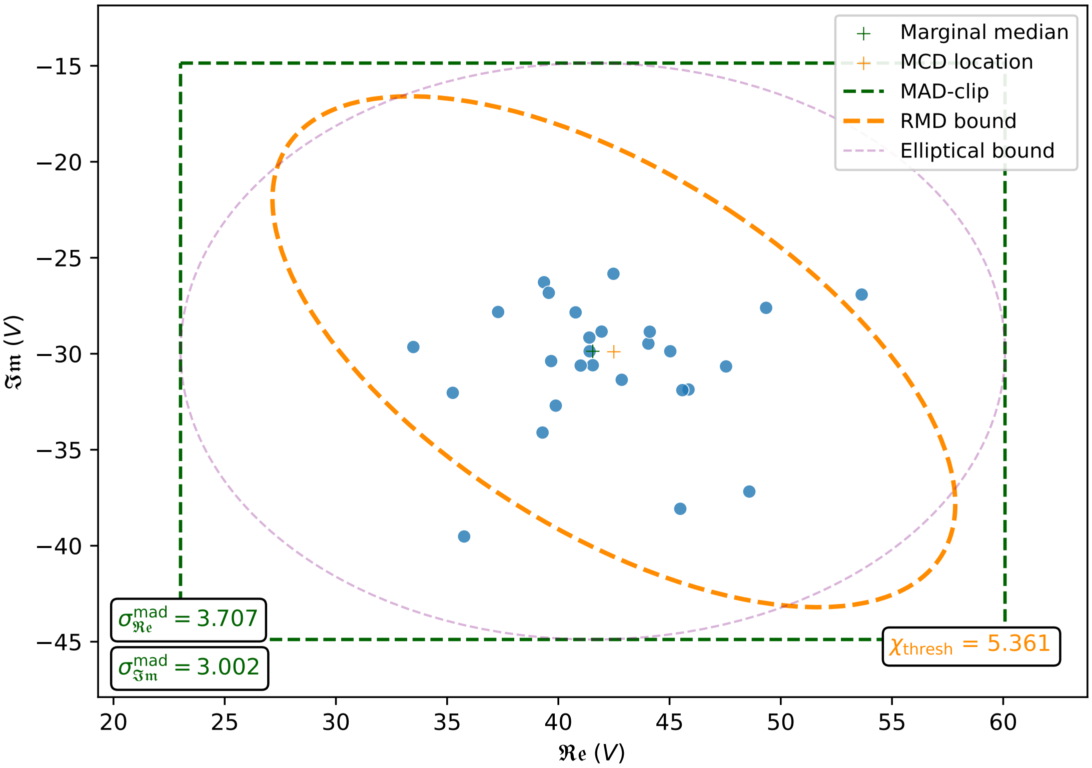

Figure 3: Comparison of the outlier boundaries and locations set by the MAD (green) and RMD (orange) anomaly detection methods. The ellipse with equation $\left(\frac{x-\text{mmed}_{\mathfrak{Re}}}{5\sigma_{\mathfrak{Re}}}\right)^2 + \left(\frac{y-\text{mmed}_{\mathfrak{Im}}}{5\sigma_{\mathfrak{Im}}}\right)^2 = 1$ is drawn as potentially smoother version of MAD-clipping.

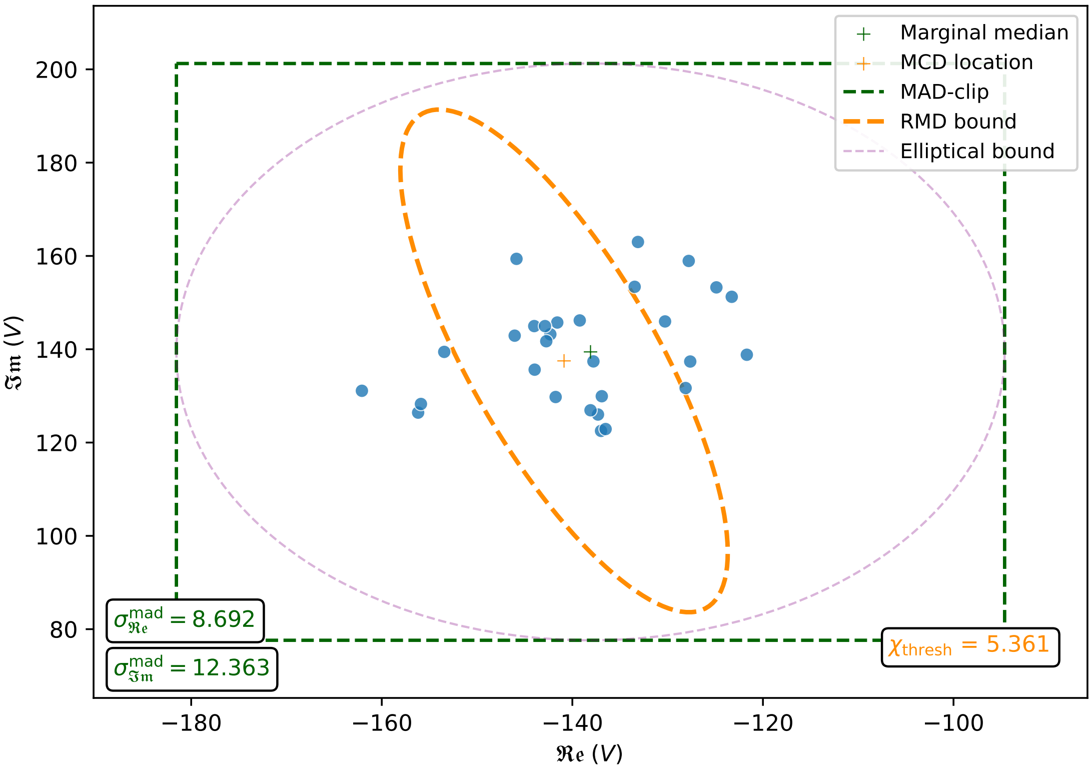

Figure 4: Comparison of the outlier boundaries and locations set by the MAD (green) and RMD (orange) anomaly detection methods for a different example data slice. In this case, the RMD algorithm discounts 13 points that would otherwise not come close to being clipped with the MAD method.

steps from the analysis pipeline may stretch the data along particular axes, which would explain the tilted covariance ellipses seen in the data. This asymmetry and tilt could also be caused if sky emission contributes to the noise - this is particularly relevant at lower frequencies.

## 3.2 Clip flags for a 20 min LST-binned file

We test the clipping routines on the LST-binning of 20 min of H1C_IDR2.2 visibility data between LSTs 5.3868–5.7398 for all the unflagged 14 m baselines. We look at the EE polarization only. Again, we use a $5\sigma$ threshold and only flag data slices with more than 5 data points. We summarize the results in Table 1.

| Clipping routine | Compute time | Number of additional flags | % of total flags |
|---|---|---|---|
| MAD | $\approx 30\,\mathrm{s}$ | 146,141 | 0.192% |
| RMD | $\approx 3\,\mathrm{h}$ | 1,250,164 | 1.619% |

Table 1: Sample flagging capabilities for the MAD and RMD-clipping routines. Compute times are from using 1 node, 8 cores at NRAO. For the equivalent quantile thresholds, RMD-clipping flags 8.55 times more data point than MAD-clipping; this is far more than would be expected if the data were Gaussian distributed. We note that the `sklearn.covariance.MinCovDet` MCD estimator used for these results has a (pseudo) randomness element that is used for shuffling the data; the random state should be fixed for reproducibility. RMD-clipping is computationally expensive, being over 360 times slower than MAD-clipping; this is because the MCD estimator and subsequent Mahalanobis distance calculations need to be run separately for each frequency/time/baseline slice, while for MAD-clipping the operations are simple and vectorized in NumPy.

We then apply these new flags to the dataset, and, in line with the LST-binning pipeline, we take the mean of each LST bin.

Looking more closely at the visibilities redundant (both in length and orientation) to (1, 12, EE), we further take the mean over baselines to get the visibility estimates shown in Fig. 5 (amplitude) and Fig. 6 (phase). These look seemingly identical, as the fraction of flags due to clipping is minute (see Table 1), and the difference in the mean between unflagged MAD and RMD-clipped points is also very small; any difference is further suppressed from averaging along the baselines axis.

In Fig. 7, we show the number of flags for each time/frequency slice with the MAD and RMD outlier rejection algorithms (with clipping done across JDs only and flags summed over the baseline axis). Both processes seem to pick out the same problematic time integrations. The RMD-clipping routine does, however, seem to be able to better pick out anomalous features confined to particular frequencies (i.e. it has some light vertical lines not seen in the plot for MAD-clipping, even if the number of flags is normalized), with some of these channels known to be bad. This suggests that RMD-clipping is more effective at finding genuine problematic issues with the data.

For a simple PS comparison of these clipping results, we compute the cross-PS across all baseline permutations, where we are still only looking at baselines redundant to (1, 12, EE). We then mean average these cross-PS in time, producing Fig. 8. The resulting PS are very similar and achieve the same noise-floor, although the mean residual between the two PS for delays $|\tau| \geq 1.25\,\mu\mathrm{s}$ (chosen to avoid the $\pm 1\,\mu\mathrm{s}$ bump) is marginally (but insignificantly) lower, meaning that the PS from RMD-clipped visibilities very slightly edges out that from MAD-clipped ones. A deeper analysis would offer more definitive results.

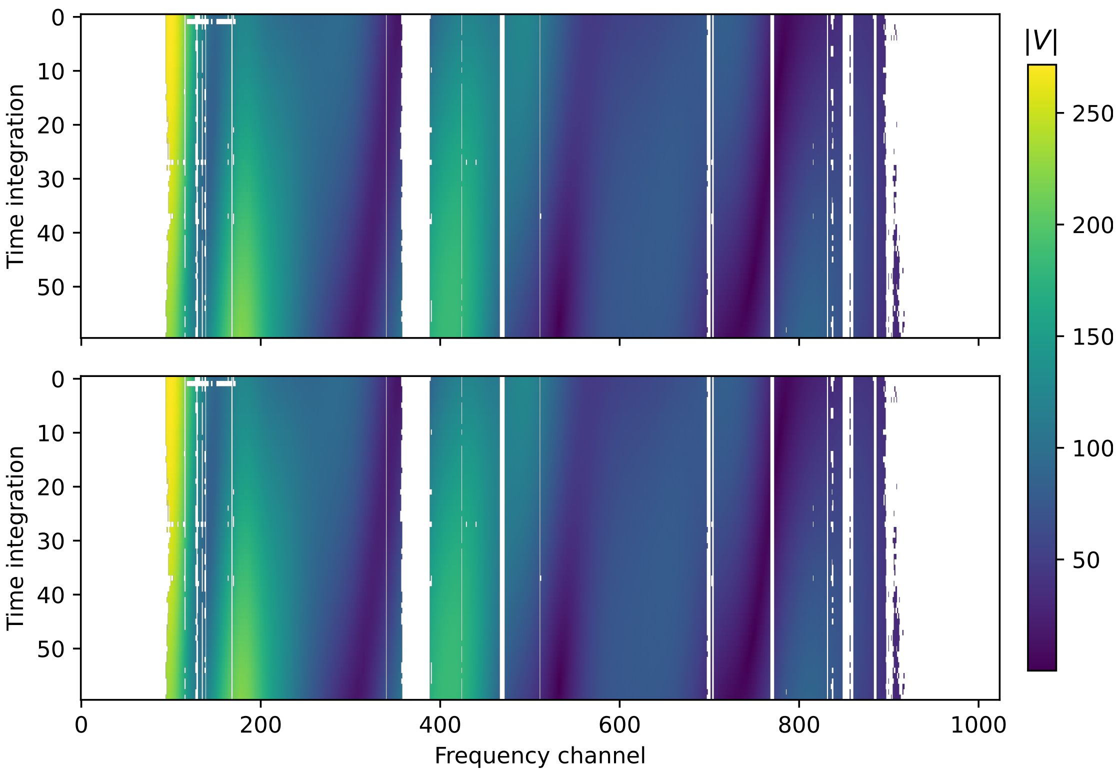

Figure 5: Visibility amplitudes after mean averaging across JDs and baselines post MAD (top) and RMD (bottom) clipping.

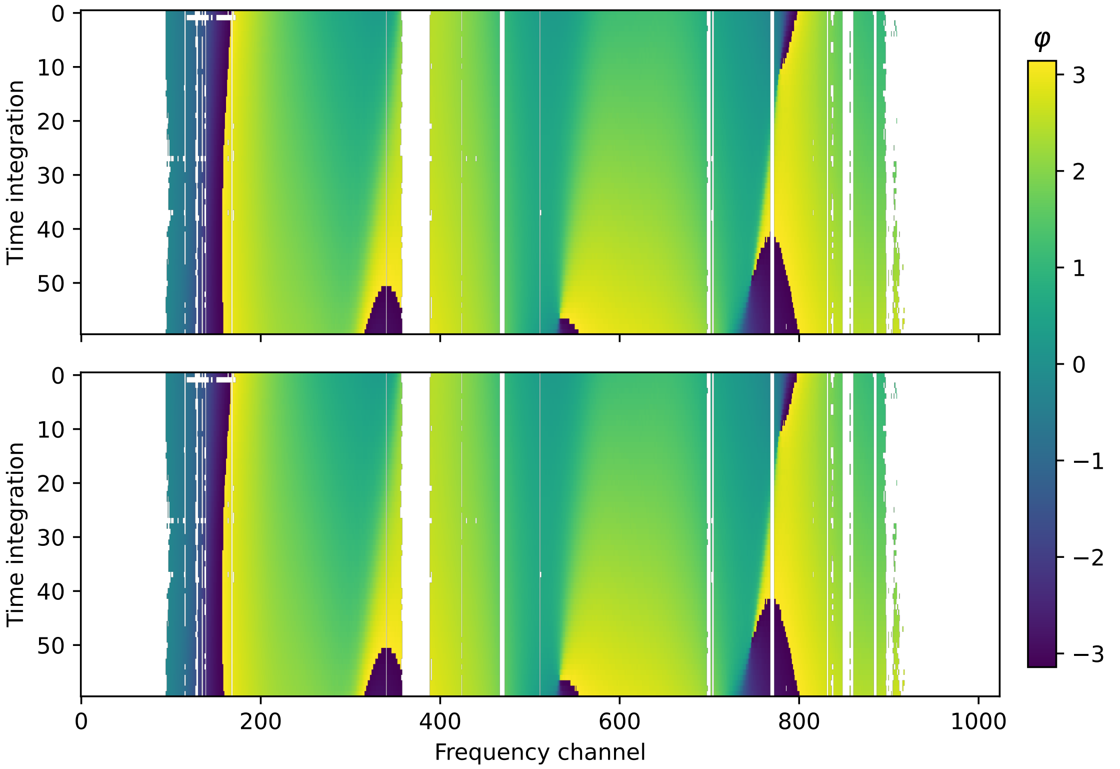

Figure 6: Visibility phases after mean averaging across JDs and baselines post MAD (top) and RMD (bottom) clipping.

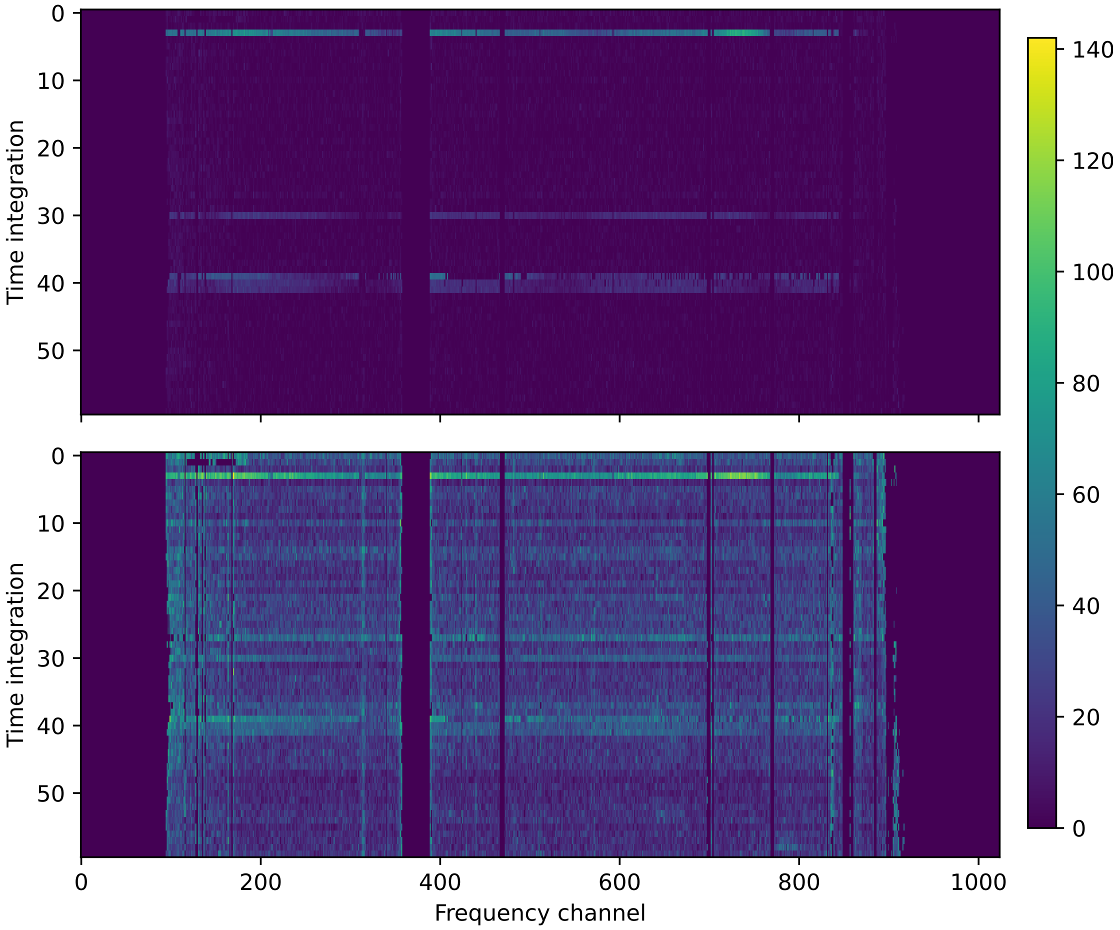

Figure 7: Number of flags across JDs and baselines returned by the MAD (top) and RMD (bottom) clipping routines.

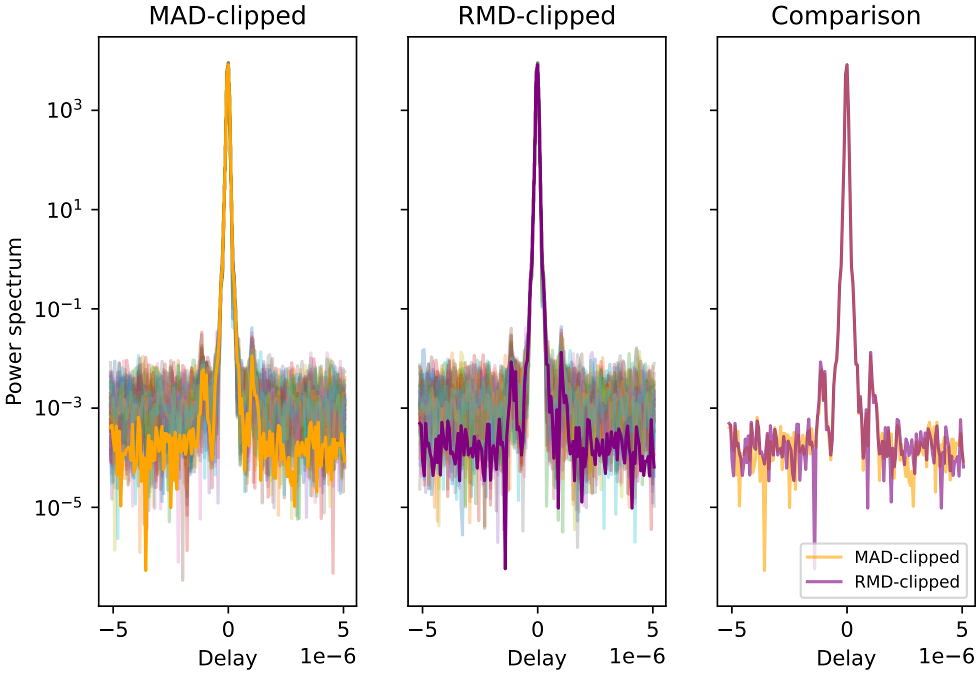

Figure 8: Absolute value of the cross-PS averaged across all baseline permutations for the MAD and RMD-clipped visibilities that have been mean averaged across JDs. The absolute value of the mean of all cross-PS over time are also shown and compared.

# 4 Conclusion

We have presented a robust method of outlier rejection that employs MCD-based Mahalanobis distances, which we call RMD-clipping.

We compare RMD-clipping to the established MAD-clipping that is used in the LST-binning pipeline: statistically speaking, RMD-clipping should be more sound for data contaminated with non-Gaussian noise such as RFI as it considers the robust location and covariance of the data and delineates reasonable elliptical outlier boundaries. Comparatively, MAD-clipping is rather rudimentary as it draws a boundary box around the data and treats the $\mathfrak{Re}$ and $\mathfrak{Im}$ components completely separately.

Empirically, through the research presented in this memorandum and in that of [1], it is found that the $\mathfrak{Re}$ and $\mathfrak{Im}$ components of LST-binned visibilities not only have unequal variance, but they have non-zero covariance. The RMD-clipping method should thus perform better than MAD-clipping, with preliminary results showing slight improvement in the PS; extending this procedure to the full H1C_IDR2.2 analysis could further improve the limits set in [2].

The clipping observed for equivalent quantiles is more aggressive for RMD-clipping (by a factor of around 8). With existing concerns of overflagging, the threshold could be increased if RMD-clipping is to be used in production.

While computationally expensive, we recommend the use of RMD-clipping as the preferred method of outlier rejection. This technique could also be used at other stages of the analysis pipeline.

In [1], we explained how robust multivariate estimators such as the geometric median can be used to get location estimates that better represent contaminated/non-normal data; such robust location estimators could be used outright without having to conduct any prior outlier rejection.

# References

[1] M. Molnar and B. Nikolic, "HERA Memorandum #106: Non-Gaussian Effects and Robust Location Estimates of Aggregated Calibrated Visibilities", October 2021.

[2] 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.

[3] P. C. Mahalanobis, "On the Generalised Distance in Statistics", *Proceedings of the National Institute of Science of India*, vol. 12, pp. 49–55, 1936.

[4] P. J. Rousseeuw, "Least Median of Squares Regression", *Journal of the American Statistical Association*, vol. 79, pp. 871–880, 1984.

[5] P. J. Rousseeuw and K. V. Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, vol. 41, pp. 212–223, 1999.

[6] 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.

[7] P. Filzmoser, R. G. Garrett, and C. Reimann, "Multivariate Outlier Detection in Exploration Geochemistry", *Computers and Geosciences*, vol. 31, pp. 579–587, June 2005.

[8] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate Outlier Detection Based on a Robust Mahalanobis Distance with Shrinkage Estimators", *Statistical Papers*, vol. 62, pp. 1583–1609, August 2021.

[9] R. A. Fisher, *Statistical Methods for Research Workers.* New York, NY: Springer New York, 1992, pp. 66–70.