# Study of RFI Filtering Algorithms for the GSB

Prasanna R

Indian Institute of Technology, Madras

Student Training Programme
May-July 2009

Under the guidance of

| | |
|---|---|
| **Prof. Yashwant Gupta** | **Shri. Jayanta Roy** |
| Chief Scientist | Research Associate |

# Certificate

This is to certify that Prasanna R, a student of Indian Institute of Technology Madras, Chennai, has done a project titled ”Study of RFI filtering algorithms for the GSB” at the GMRT observatory of NCRA-TIFR, Pune under the guidance of Prof. Yashwant Gupta and Mr. Jayanta Roy as part of the Student Training Programme (May 10th - July 10th, 2009).

**Prof. Yashwant Gupta**
Chief Scientist
GMRT

**Mr. Jayanta Roy**
Research Associate
NCRA

# Acknowledgement

I thank my guides Prof. Yashwant Gupta and Shri Jayanta Roy, and Prof. Rajaram Nityananda for giving me a chance to work on this project. I would also like to thank my guides for finding time out of their busy schedule, for guiding me. I had a wonderful experience of working in a free and stimulating environment over the summer. I would also like to thank all the people I met at GMRT  the staff, my fellow project students and astronomers.

# Contents

# 1 Introduction

## 1.1 GMRT Software Backend

The GMRT Software Backend uses a linux cluster consisting of 48 Intel Xeon nodes. This backend can play the dual role of a correlator and incoherent or phased array pulsar receiver.

In addition to a real-time processing pipeline, the backend also offers a baseband recording mode, where raw voltage data from all the antennas is written directly to disk, followed by 32 bit floating point offline processing to produce fringe stopped, integrated visibilities.

This software approach to the backend also opens up the scope for implementing specialized algorithms like **RFI rejection** and calibration schemes, that could greatly enhance the capability of the GMRT.

## 1.2 RFI Filtering Algorithms

Signals from astronomical sources follow gaussian probability distribution function [1] (Figure 1). Most of the Radio Frequency Interference (RFI) that the GMRT sees is non-gaussian in nature. The RFI filtering algorithms considered here are essentially flagging algorithms that indicate the presence of non-gaussian RFI. These algorithms have a common workflow. First, a statistical quantity is calculated from the data. This quantity is compared with the corresponding quantity for the gaussian distribution. If there is a large difference in these two quantities, the block of data considered is probably affected by non-gaussian RFI, and thus can be flagged.

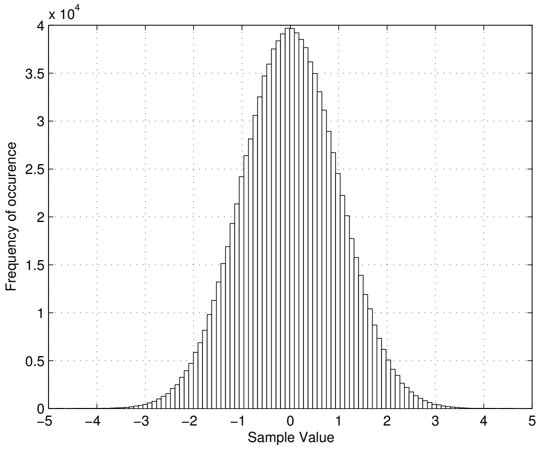

Figure 1: Gaussian Probability Distibution

This report describes the theory behind setting the thresholds for filtering and presents results from tests on simulated data as well as real data from the antennas.

# 2 Theory

## 2.1 Kurtosis based filtering

The astronomical signals detected by a Radio Telescope are Gaussian in nature[1]. The probability distribution function (pdf) of such a signal is given by

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \tag{2.1}$$

The central moments of this distribution are given by

$$m_n = \langle (x - \langle x \rangle^n \rangle = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (n-1)\sigma^n \tag{2.2}$$

where $\sigma$ is the standard deviation of $x$ and $n$ is even. The odd central moments are zero. The kurtosis of $x$ is defined as

$$k = \frac{m_4}{m_2^2} \tag{2.3}$$

From equations 2.2 and 2.3 it is evident that the kurtosis of a Gaussian distribution is always 3. If the signal is corrupted by RFI, it's pdf may deviate from that of a Gaussian. Consequently, the kurtosis may deviate from 3. In the kurtosis based approach of filtering, the detection of RFI boils down to calculating the kurtosis estimator and deciding whether this corresponds to that of a Gaussian random variable.

Since the kurtosis is estimated from a finite sample set, it is itself a random variable. Let it's pdf be denoted by $p(k; N)$ where $N$ is the number of samples. Closed form expressions for $p(k; N)$ exist only for $N = 4$. In the limit of large $N$, $p(k; N)$ tends towards a normal distribution with [3]

$$\langle k \rangle = 3\frac{N-1}{N+1} \tag{2.4}$$

$$\sigma_k^2 = \frac{24}{N} \tag{2.5}$$

Shown in Figure 2 is the behaviour of the kurtosis estimator with sample size $N$. The dashed lines are the $3\sigma$ limits as defined by equation 2.5.

The error in the kurtosis estimator as defined by equation 2.5 can be used as a detection threshold for non-gaussian signal blocks which have been corrupted by RFI.

## 2.2 Second Moment based filtering

For a Gaussian distributed variable, $(\sqrt{N-1})s/\sigma$ follows a $\chi$ - distribution with $N-1$ degrees of freedom, where $s$ is the sample standard deviation. Consequently,

$$E[s] = c_4\sigma \tag{2.6}$$

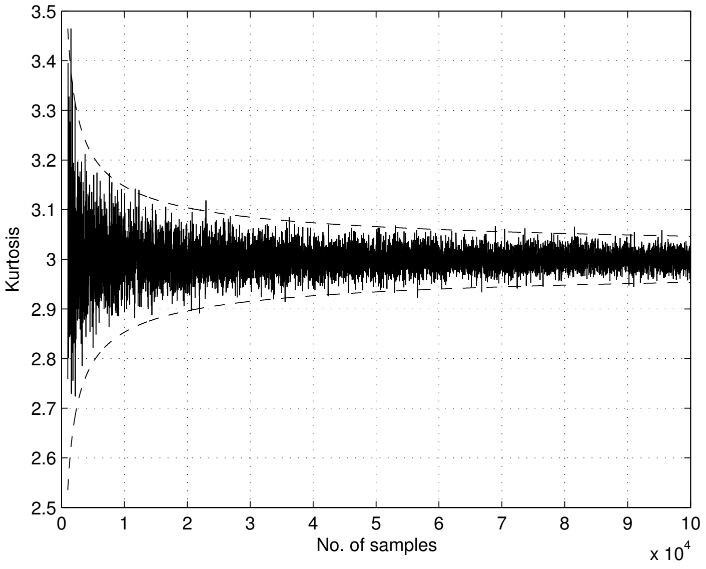

Figure 2: Behaviour of the kurtosis estimator with sample size N

where $c_4$ is a constant that depends on the sample size $N$ as follows.

$$c_4 = \sqrt{\frac{2}{N-1}}\frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})} = 1 - \frac{1}{4N} - \frac{7}{32N^2} - O(N^{-3}) \tag{2.7}$$

Using this, we find out that the error in this Standard deviation estimator is $\sigma\sqrt{1-c_4^2}$

From a Gaussian distributed data set, samples of different sizes were drawn and their standard deviation was calculated. This is shown in Figure 3 and is overlayed with 3 times the error in its esimation, that is, $1 \pm 3\sigma\sqrt{1-c_4^2}$. This error can be used as a practical threshold for detecting non-gaussian blocks.

## 2.3 MAD Filtering

MAD stands for Median Absolute Deviation. MAD is a robust measure of the variability of quantitative data. For a data set $X_1, X_2, X_3, \ldots X_n$

$$MAD = median_i|X_i - median_j(X_j)| \tag{2.8}$$

MAD is more resilient to outliers than the standard deviation. In the standard deviation, the distances from the mean are squared and hence larger deviations are weighted more heavily, thus outliers can heavily influence it.

To use MAD as a consistent estimator for the estimation of standard deviation $\sigma$, we take,

$$\hat{\sigma} = K \cdot MAD \tag{2.9}$$

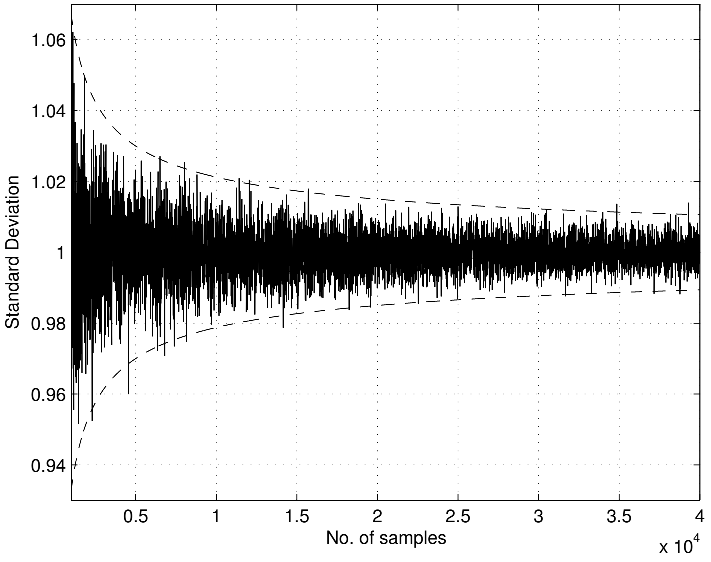

Figure 3: Behaviour of the standard deviation estimator with sample size N

For gaussian distributed data, $K$ is given by $1/\Phi^{-1}(3/4) \approx 1.4826$ where $\Phi$ is the cumulative distribution function for the normal distribution. This is because

$$\frac{1}{2} = P(|X-\mu| \leq \mathrm{MAD}) = P\left(\left|\frac{X-\mu}{\sigma}\right| \leq \frac{\mathrm{MAD}}{\sigma}\right) = P\left(|Z| \leq \frac{\mathrm{MAD}}{\sigma}\right) \tag{2.10}$$

Hence,

$$\sigma \approx 1.4826\ \ \mathrm{MAD}\,. \tag{2.11}$$

The MAD approach works on individual samples, as opposed to the kurtosis and second moment approaches that work on blocks of data. One could look at the value $|X_i - median(X_j)|$ for each sample, and based on a $3\sigma$ threshold given by equation 2.11, filter out data points.

## 2.4 Spectral Kurtosis

Assume that a time-domain signal $\{x(t_n)\} \equiv \{x_n\}$ is characterised by the following property: the time-domain samples are drawn from the same parent Gaussian population with zero mean, $\langle x \rangle = 0$, and a constant variance $\sigma_x^2 = \langle x_n^2 \rangle$.

To obtain a PSD estimate of a time-domain signal, we first calculate the DFT coefficients by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi ikn/N} \qquad\qquad k = 0, \ldots, N-1 \tag{2.12}$$

The actual expressions for the spectral powers corresponding to each of the $N/2+1$ discrete frequency bins,

$$f_k = \frac{2k}{N} f_c \qquad\qquad k = 0, 1, \ldots, \frac{N}{2} \tag{2.13}$$

of the PSD estimate are:

$$\hat{P}_k = \frac{2}{N}|X_k|^2 \tag{2.14}$$

The mean and variance of the PSD esitmates are

$$\mu_k = \langle \hat{P}_k \rangle \tag{2.15}$$

$$\sigma_k^2 = \langle \hat{P}_k^2 \rangle - \langle \hat{P}_k \rangle^2 \tag{2.16}$$

We define a term called the Spectral Variability

$$V_k^2 = \frac{\sigma_k^2}{\mu_k^2} \tag{2.17}$$

Spectral Kurtosis is defined as,

$$SK(f_k) = \frac{\langle |X_k|^4 \rangle - 2\langle |X_k|^2 \rangle^2}{\langle |X_k|^2 \rangle^2} \tag{2.18}$$

which is expected to be exactly 0 for a Gaussian time-domain signal. Taking into account that $|X_k|^2 \sim \hat{P}_k$, it follows that

$$V_k^2 = SK(f_k) + 1 \tag{2.19}$$

Thus, we see that the spectral variability is equivalent to the spectral kurtosis upto an additive constant.

### Implementation

We define a Spectral Kurtosis Estimator as,

$$\hat{V}_k^2 = \frac{\hat{\sigma}_k^2}{\hat{\mu}_k^2} \tag{2.20}$$

where $\hat{\sigma}_k^2$ and $\hat{\mu}_k^2$ are unbiased estimators of $\sigma_k^2$ and $\mu_k^2$ respectively. In practice, these two unbiased estimators will be computed from a number of M adjacent blocks of N time-domain samples used to obtain M spectral estimates, $\hat{P}_{ki}(k = 0, \ldots, N/2; i = 1, \ldots, M)$.
Using $S_1$ and $S_2$ as notation for the two sums,

$$S_1 = \sum_{i=1}^{M} \hat{P}_{ki} \qquad S_2 = \sum_{i=1}^{M} (\hat{P}_{ki})^2 \tag{2.21}$$

we get,

$$\hat{\mu}_k = \frac{1}{M} S_1 \qquad \hat{\sigma}_k^2 = \frac{MS_2 - S_1^2}{M(M-1)} \tag{2.22}$$

Now, the Spectral Kurtosis estimator becomes,

$$\hat{V}_k^2 = \frac{M}{M-1}(M\frac{S_2}{S_1^2} - 1) \tag{2.23}$$

Variance of this estimator is given by[3]

$$Var(\hat{V}_k^2) = \frac{4M}{(M-1)^2} \qquad k = 1, ..., (N/2-1) \tag{2.24}$$

For large enough number of blocks $M$,

$$Var(\hat{V}_k^2) = \frac{4}{M} \qquad k = 1, ..., (N/2-1) \tag{2.25}$$

Equation 2.25 can be used used to obtain a detection threshold for practical purposes.

# 3 Simulations

To understand the various statistical parameters (described above) and their responses to different kinds of additive noise, a number of simulatios were carried out. A random gaussian signal was first generated and to that two types of additive noise were added:

- Uniform random noise
- Pulsed Sinusoidal noise

## 3.1 Using Uniform Random Noise

The most trivial form of non-gaussian noise is the uniform random noise. Shown below in Figure 4, is the histogram of symmetric unifrom noise having a range (amplitude) -1 to 1.

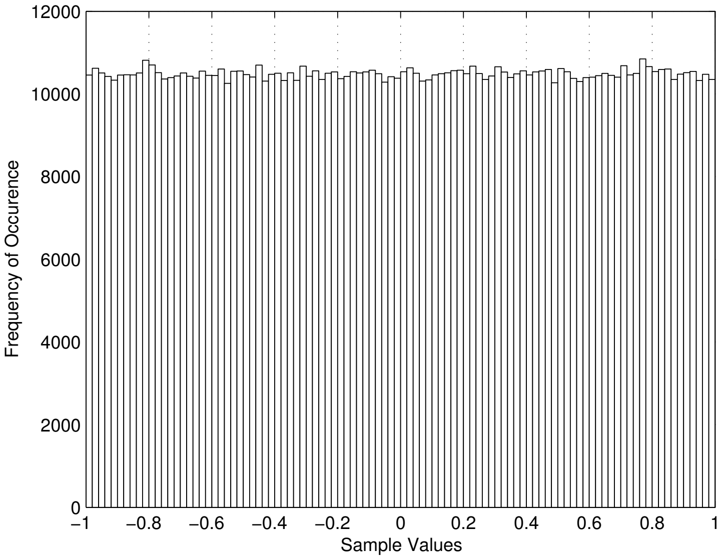

Figure 4: Uniform Random noise having amplitude 1

### 3.1.1 Kurtosis based filtering

A uniform probability distribution is flatter than a gaussian. This leads to decrease in the value of kurtosis with the increase in the amplitude of the uniform noise added.

A random gaussian data (length: $2^{20} = 1048576$, mean $= 0$ and variance $= 1$) was generated. To this, uniform noise of different amplitudes were added (uniform random noise of amplitude x denotes that the random variable is uniformly distributed in -x to x). This data is split into blocks of size N. For each block, the sample kurtosis is calculated. Using the thresholds (described in Section 2.1, the block was marked as either bad or good. Similar flagging procedure was carried out for different block sizes. This is shown in Figure 5 where the percentage of bad blocks is plotted against amplitude of the unifrom noise for various blocking factors.

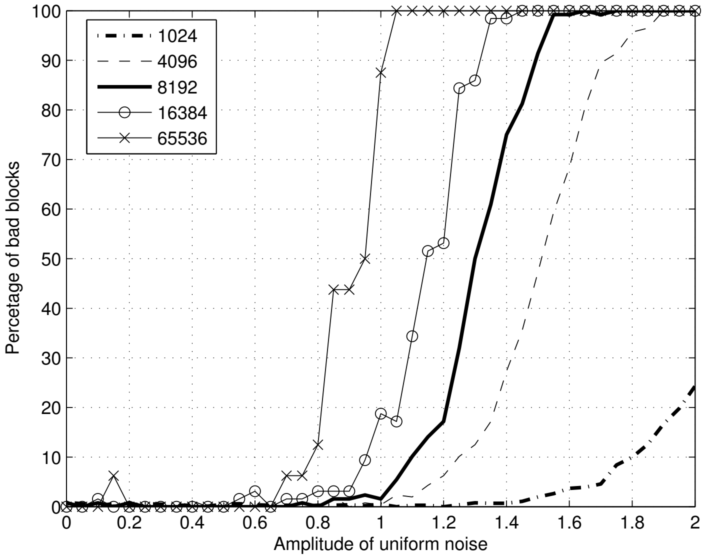

Figure 5: Percentage of bad blocks versus amplitude of uniform random noise

We see that the amount of data marked as bad decreases with decrease with block size. this could be attributed to the fact that as the block size decreases, the limits on the kurtosis estimator become wider. This could cause very few blocks to have kurtosis outside this large window.

### 3.1.2 Second Moment based filtering

Exactly as in the case of the kurtosis based filtering, the standard deviation was calculated for each block of size N and depending upon the thresholds on the standard deviation estimator (Section 2.2) the blocks were marked as either good or bad. Show below in Figure 6 is the percentage of bad blocks versus the amplitude of uniform noise for different block sizes.

A similar type of variation in the percentage of bad blocks with block size is seen as in the case of Kurtosis based filtering. Again, this could be attributed to the same fact that the limits on the standard deviation estimator become wider with the decrease in block size.

### 3.1.3 MAD Filtering

MAD filtering is different from the above two algorithms because of the fact that this works on each individual sample rather than on blocks of data.

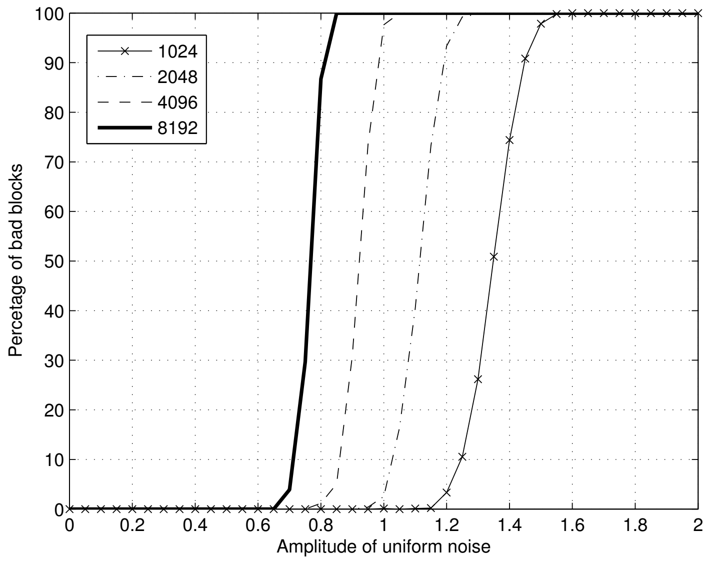

Figure 6: Percentage of bad blocks versus amplitude of uniform random noise

As described in Section 2.3, MAD for a symmetric distribution about the mean is the $75^{th}$ percentile. As the amplitude of the uniform noise increases, the distribution function becomes flatter and MAD also increases. For a large enough amplitude of uniform noise, say x, MAD will be approximately $\frac{3}{4}x$.

To filter out data, one has to look at the absolute deviation of individual samples from the median and then compare it with the MAD. If MAD is close to $\frac{3}{4}x$. and the highest individual sample value is x, it is not possible to find samples with large enough deviations that could get filtered out. Hence, with the increase in the amplitude of uniform noise, MAD ends up throwing out lesser data.

### 3.1.4 Conclusions

For a fixed blocking factor (N = 8192), the percentage of bad blocks marked by Kurtosis and second moment are plotted in Figure 8. Second moment seems to respond to the uniform noise more sensitively than the kurtosis. As expected, MAD performs poorly, hardly flagging any data.

After flagging the bad blocks, the standard deviation of the remaining good blocks was calculated. The reciprocal of this standard deviation is regarded as a measure of the SNR of the signal. The percentage improvement in SNR after filtering is shown in Figure 9.

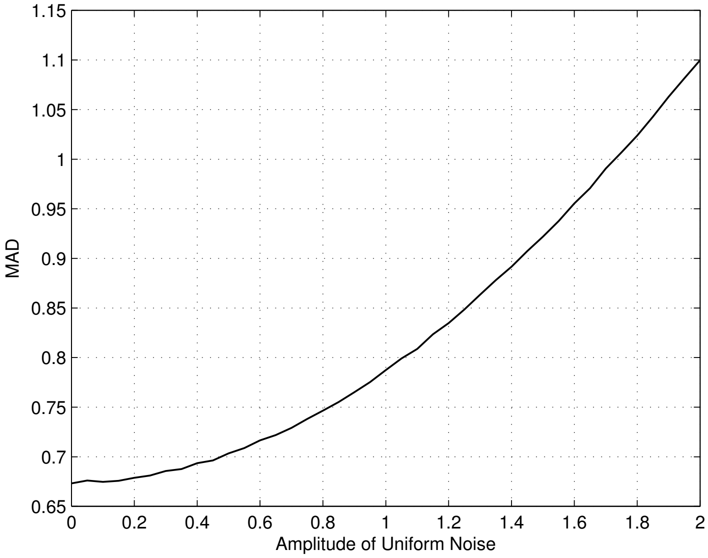

Figure 7: Variation of MAD with amplitude of uniform noise

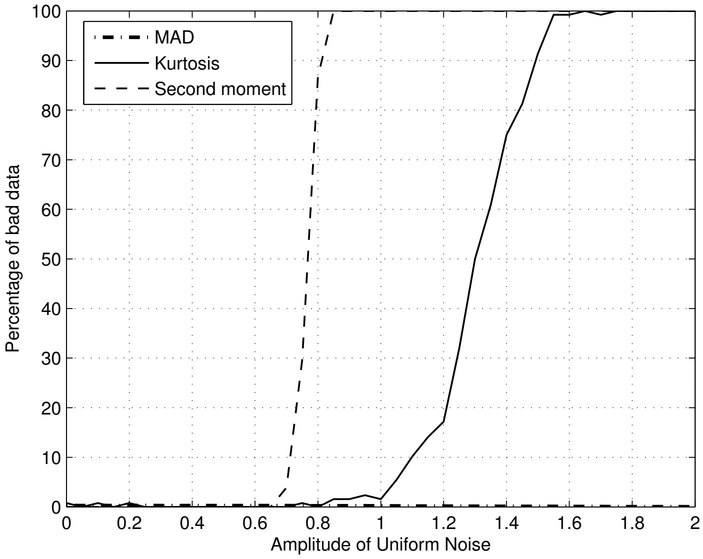

Figure 8: Comprison: Percentage of flagged data

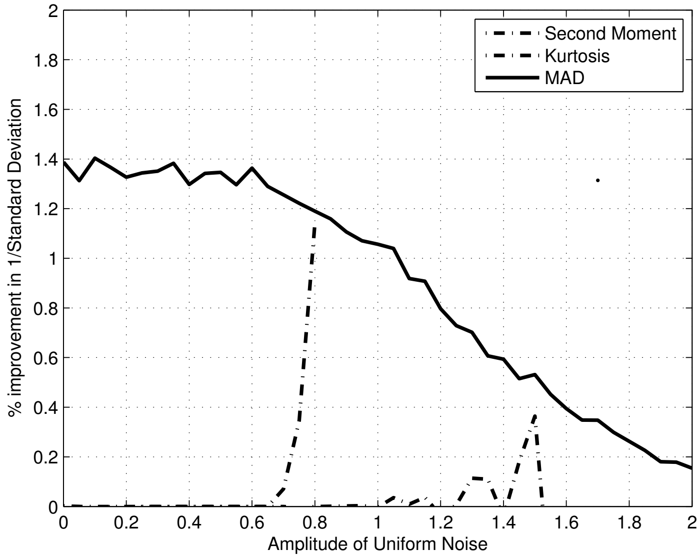

Figure 9: Comprison: Percentage improvement in SNR

## 3.2 Using Pulsed Sinusoids

Pulsed sinusoidal noise is a more realistic non-gaussian RFI than the uniform random noise. A pulsed sinusoid is generated by by first generating a sine wave and multiplying it with gating pulses. The duty cycle of the gating pulse can be varied to get different duty cycle pulsed sinusoids. Shown in Figure 10 is the histogram of a 90% duty cycle pulsed sinusoid.

### 3.2.1 Kurtosis based filtering

A random gaussian data (length: $2^{20} = 1048576$, mean $= 0$ and variance $= 1$) was generated. To this, pulsed sinusoids of different amplitudes and different duty cycles were added. The variation of the kurtosis estimator with amplitude of the pulsed sinusoid for different duty cycles is shown in Figure 11. The X-axis marked SNR is the ratio (in dB) of the variance of the pulsed sinusoid and the variance of the gaussian.

Wee see that for fairly intermittent type of RFI (low duty cycle pulsed sinusoids) the kurtosis estimator is very sensitive. As the duty cycle increases, kurtosis becomes less sensitive and has a blind spot at 50% duty cycle.

### 3.2.2 Second Moment based filtering

Using the same approach as that employed with kurtosis, to a random gaussian signal, pulsed sinusoids of different amplitudes and different duty cycles were added. The variation of the standard deviation estimator with amplitude of the pulsed sinusoid for different duty cycles is shown in Figure 12. The X-axis marked SNR is the ratio (in dB) of the variance of the pulsed sinusoid and the variance of the gaussian.

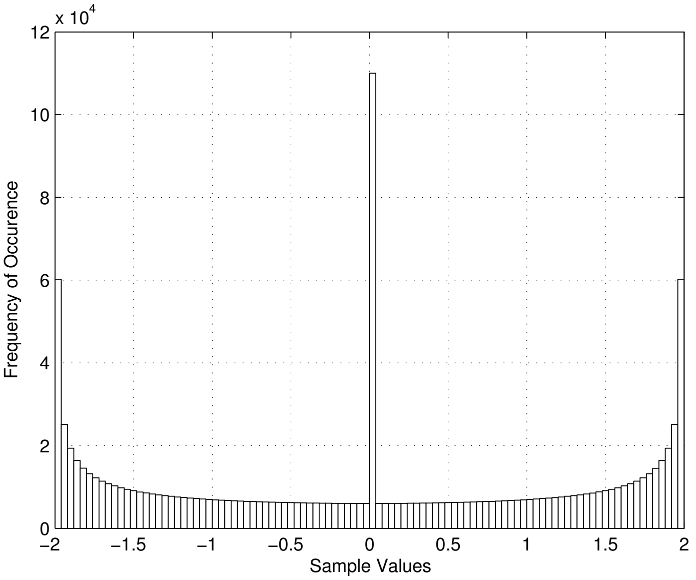

Figure 10: Histogram of a 90% duty cycle pulsed sinusoid

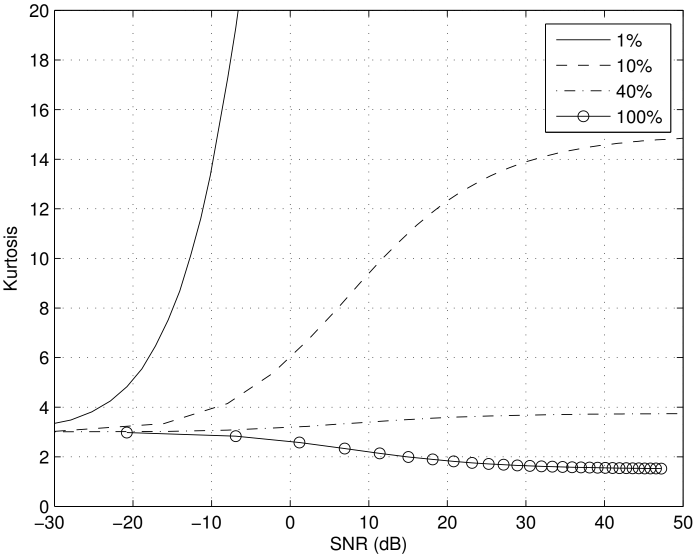

Figure 11: Variation of Kurtosis with amplitude of pulsed sinusoid

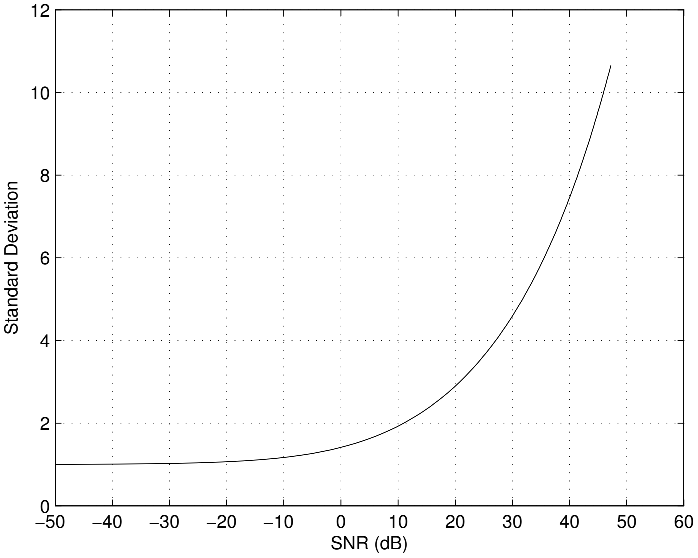

Figure 12: Variation of Standard deviation with amplitude of pulsed sinusoid

We notice that the standard deviation, unlike the kurtosis, varies in the exact same way with amplitude of pulsed sinusoid for all the duty cycles.

### 3.2.3 MAD Filtering

Using the same approach as that employed with kurtosis and standard deviation, to a random gaussian signal, pulsed sinusoids of different amplitudes and different duty cycles were added. The variation of MAD with amplitude of the pulsed sinusoid for different duty cycles is shown in Figure 13. The X-axis marked SNR is the ratio (in dB) of the variance of the pulsed sinusoid and the variance of the gaussian.

In the previous two cases, we were looking directly at the parameter used to filter data. Shown in Figure 13 is MAD which is not the direct parameter used to filter out the data. Rather, data is filetered on the basis of their deviations from the median compared with MAD. From Figure 13 we can also see that MAD hardly shows any variation for a 1% duty cycle sinusoid, which means that MAD is not heavily influenced by a few outliers.

### 3.2.4 Spectral Kurtosis

A random gaussian data (length: $2^{20} = 1048576$, mean = 0 and variance = 1) was generated. To this, pulsed sinusoids of different amplitudes and different duty cycles were added. Using a block size of 1024, 1024 point FFT was calculated for each block. Looking at the time series plot of each frequeny bin (taking 1024 such blocks into account), the Spectral Variability was calculated. (M=1024, N=1024, Refer section 2.4)

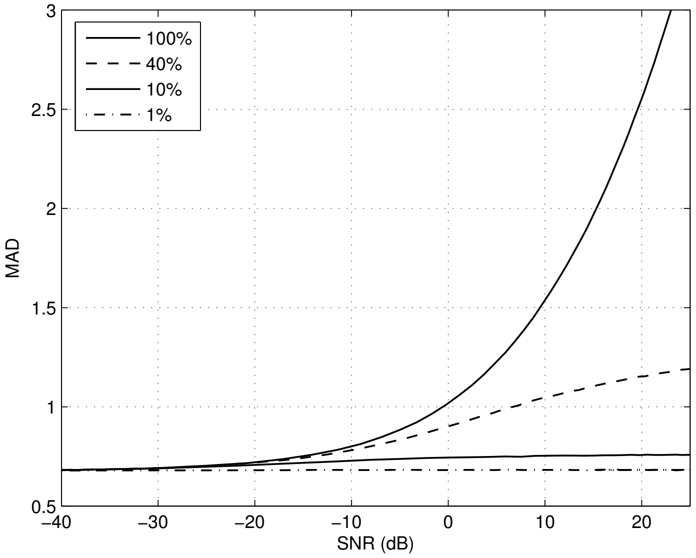

Figure 13: Variation of MAD with amplitude of pulsed sinusoid

Show below in figure 14 is the Spectral Variability (a measure of spectral kurtosis) as a function of the frequency bins for a gaussian signal without any additional noise. Figure 15, figure 16 and figure 17 show the same plot after adding -50 dB of 1%, 10 and 100% duty cycle pulsed sinusoid respectively.

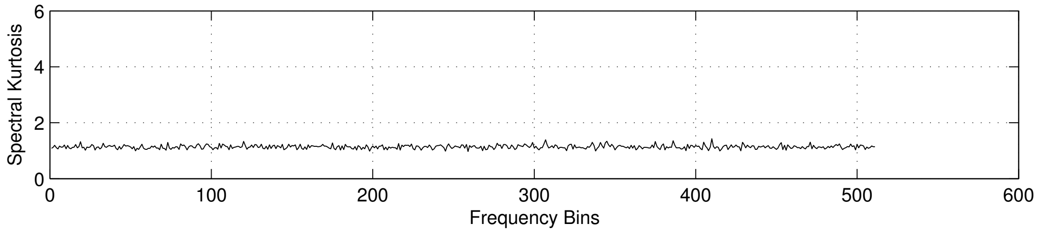

Figure 14: Gaussian signal

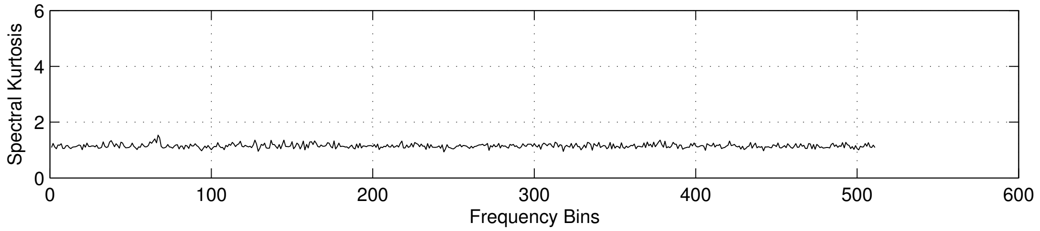

Figure 15: 1% Duty cycle

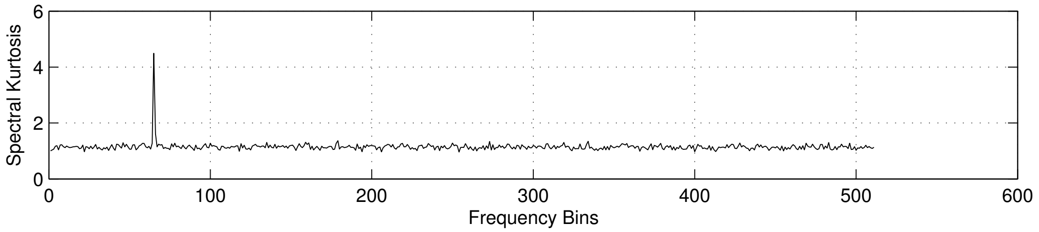

Figure 16: 10% duty cycle

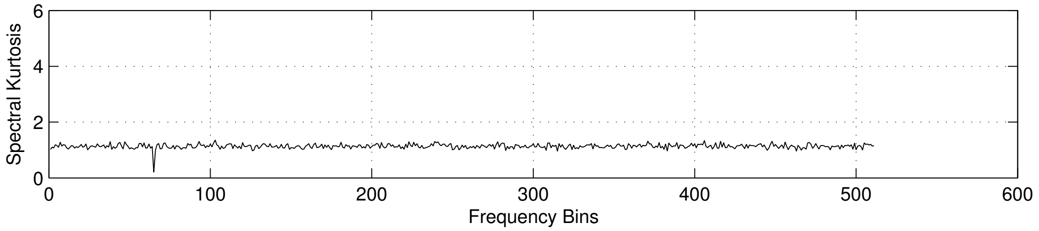

Figure 17: 100% duty cycle

### 3.2.5 Conclusions

For a fixed blocking factor (N = 8192) and duty cycle (10%) the percentage of bad blocks marked by Kurtosis and second moment are plotted in Figure 18.

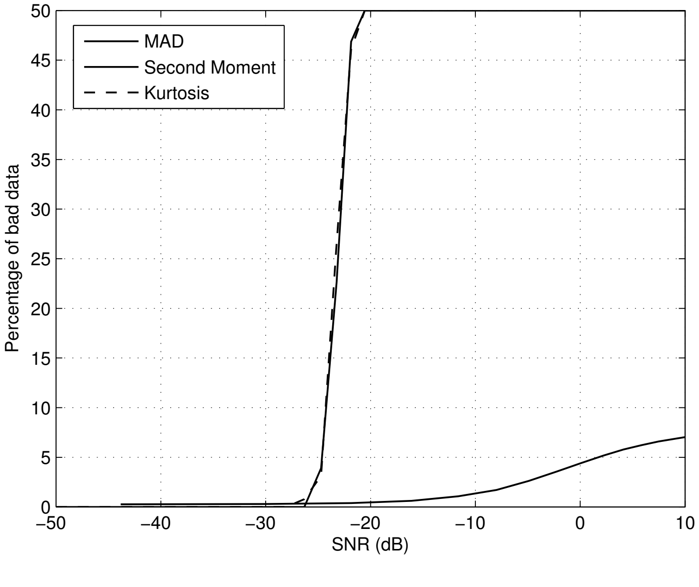

Figure 18: Comparison: Percentage of flagged data

As noted earlier, second moment filtering does equally well for all duty cycles Kurtosis is more sensitive at lower duty cycles. Both of these methods seem to do equally well at a duty cycle of 10%. Below 10% Kurtosis is more sensitive.

Kurtosis-based and second moment based filtering work on blocks of data. Hence, from a probabalistic point of view, they would throw more data compared to MAD which works on each individual sample. This can be observed in Figure 18 where the block-based algorithms flag 50 % of the data whereas MAD flags only ≈ 10% of the data. Since the noise added here is a 10% duty cycle pulsed sinusoid, MAD possibly flags only the exact samples of the data that have been affected by RFI. This is also reflected in the SNR (1/st. dev.) plot in Figure 19.

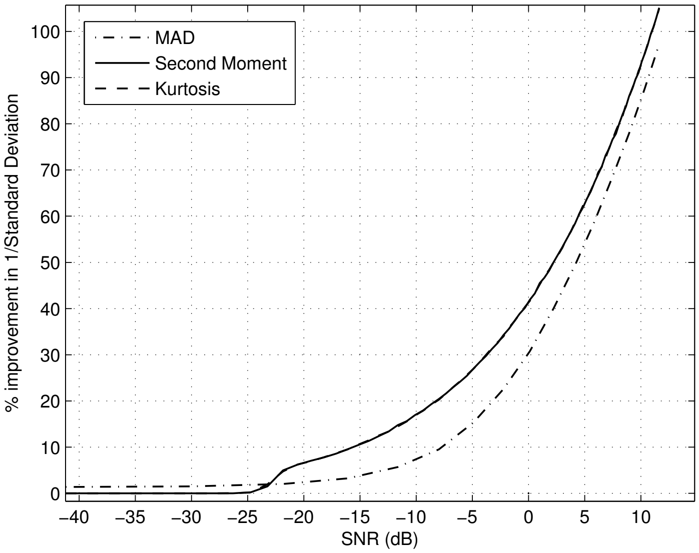

Figure 19: Comparison: Percentage improvement in SNR

MAD seems to be doing almost equally well here as the other two methods, even though it flags very less amount of data. This leads one to conclude that MAD indeed flags only the data samples affected with RFI, this doing a better job. Spectral kurtosis seems to be picking up RFI in particular frequency bin very well. This needs to be examined further.

# 4 Tests on Real Data

## 4.1 Kurtosis based filtering

Raw voltage data at 150 MHz from 4 antennas were taken as test cases. Kurtosis of blocks of data (8192 samples) were calculated and plotted. This is shown in Figure 20. We see that 3 of the antennas show kurtosis highly deviated from 3. This shows the presence of heavy non-gaussian RFI at 150 MHz. Shown in Figure 21 is the Fourier Spectrum of the data from antenna 2 (The antenna having kurtosis centered around 3). Clear presence of powerline and its harmonics can be seen.

The antenna with kurtosis close to 3 was chosen for further analysis. This was passed through the kurtosis flagging algorithm. Shown in Figure 23 is the intensity plot after the bad blocks had been thrown out. Compared to the original intensity plot (Figure 22), this does not seem to be much different. Looking at the kurtosis plot of this antenna in Figure 20, we see that the kurtosis variation is not symmetric about 3, but more biased towards the lower side. But the thresholds set on the kurtosis estimator are symmetric about 3. This causes less data to be flagged with kurtosis above 3, even though most of the RFI seems to be present in this data.

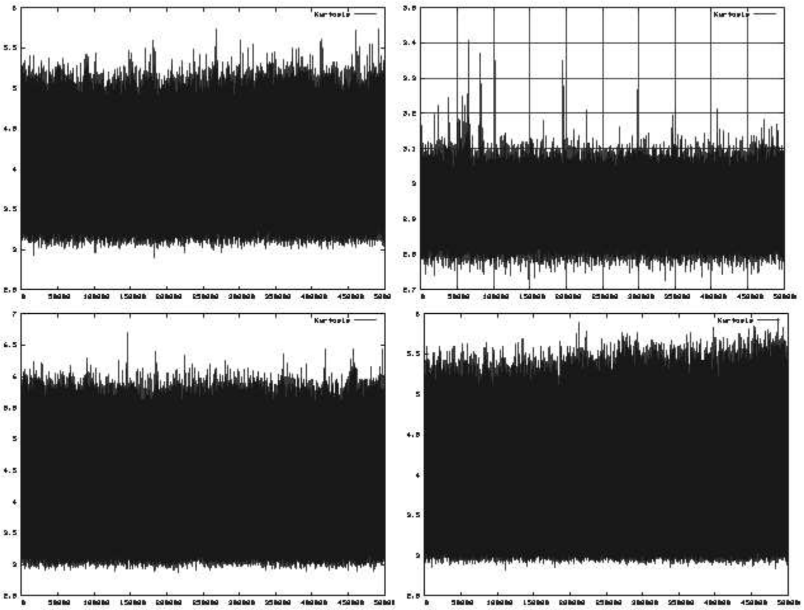

Figure 20: Kurtosis plots for data from four antennas

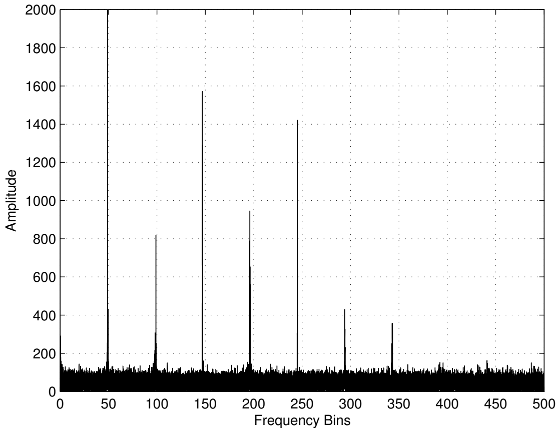

Figure 21: Fourier Spectrum of Antenna 2

## 4.2 Second Moment filtering

The raw voltage data considered here is 4-bit data. This contains 16 levels: -8 to 7. The Analog to Digital converter used to obtain this data is assumed to span $10\sigma$ ($5\sigma$ on both sides of zero). This means that the standard variation of the data should be 1.6. Indeed the antenna that was considered for kurtosis filtering shows standard deviation centered around 1.6. This is shown in Figure 22.

Data was passed through a second-moment based filter. The spikes in the standard deviation of the data essentially get chopped off. This is shown in Figure 24.

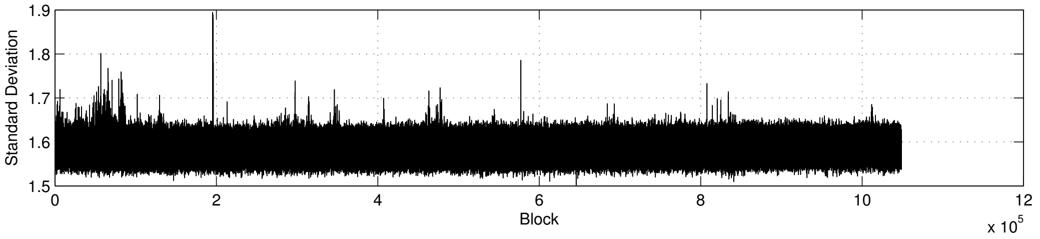

Figure 22: Intensity plot of raw data

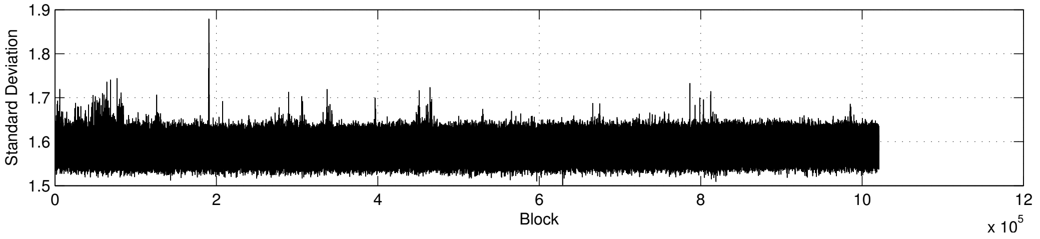

Figure 23: Intensity plot of data passed through Kurtosis filter

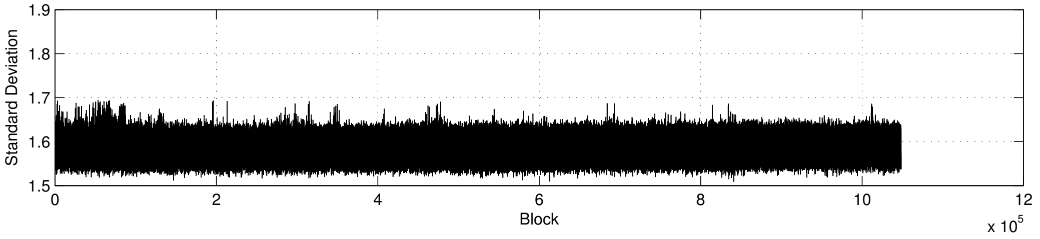

Figure 24: Intensity plot of data passed through second moment filter

# 5 Future Work

Real data from the antennas should be subjected to more analyis to verify conclusions drawn from the simulations. Filtered data should be passed through the correlator pipeline and improvements in the output parameters such as the visibilty, phase stability etc. must be checked to see if the RFI has been filtered out.

Depending upon the results, a combination of time-domain and spectral-domain filtering techniques can be adopted and implemented into the software backend.

# References

[1] J.N. Chengalur, Y. Gupta, K.S. Dwarakanath, *Low Frequency Radio Astronomy*. NCRA, 2003.

[2] C.S. Ruf, S.M. Gross, S. Mishra *RFI Detection and Mitigation for Microwave Radiometry with an Agile Digital Detector*. IEEE Transactions on Geoscience and Remote Sensing, Vol. 44 No. 3, March 2006.

[3] R.D. De Roo, S. Mishra, C.S. Ruf, *Sensitivity of the Kurtosis Statistic as a Detector of Pulsed Sinusoidal RFI*. IEEE Transactions on Geoscience and Remote Sensing, Vol. 45 No. 7, July 2007.

[4] G.M. Nita, D.E. Gary, Z. Liu *Radio Frequency Interference Excision using Spectral-Domain Statistics*. Publications of the Astronomical Society of the Pacific, 119: 805-827, July 2007.