RAS Techniques and Instruments

RASTAI **3,** 535–547 (2024) https://doi.org/10.1093/rasti/rzae035
Advance Access publication 2024 August 20

Interference detection in radio astronomy: applying Shapiro–Wilks normality test, spectral entropy, and spectral relative entropy

Zhicheng Cao [1], Natalia A. Schmid [2,3]★ Kevin Bandura,[2,3]★ Duncan R. Lorimer,[3,4]★
Morgan Dameron,[2] Katelyn Crockett,[2] Clayton Grubick,[2] Andreas Schmid[3,5] and Shaonan Zheng[1]

[1]*Xidian University, School of Life Science and Technology, Shaanxi 710126, China*
[2]*Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, 26506, USA*
[3]*Center for Gravitational Waves and Cosmology, West Virginia University, Chestnut Ridge Research Building, Morgantown, WV, 26506, USA*
[4]*Department of Physics and Astronomy, West Virginia University, Morgantown, WV, 26506, USA*
[5]*Department of Electrical and Computer Engineering, University of Illinois in Urbana-Champaign, Urbana, IL, 61801, USA*

Accepted 2024 August 5. Received 2024 June 16; in original form 2024 January 6

ABSTRACT

Radio-frequency interference (RFI) is becoming an increasingly significant problem for most radio telescopes. Working with Green Bank Telescope data from PSR J1730+0747 in the form of complex-valued channelized voltages and their respective high-resolution power spectral densities, we evaluate a variety of statistical measures to characterize RFI. As a baseline for performance comparison, we use median absolute deviation (MAD) in complex channelized voltage data and spectral kurtosis (SK) in power spectral density data to characterize and filter out RFI. From a new perspective, we implement the Shapiro–Wilks (SW) test for normality and two information theoretical measures, spectral entropy (SE) and spectral relative entropy (SRE), and apply them to mitigate RFI. The baseline RFI mitigation algorithms are compared against our novel RFI detection algorithms to determine how effective and robust the performance is. Except for MAD, we find significant improvements in signal-to-noise ratio through the application of SE, symmetrical SRE, asymmetrical SRE, SK, and SW. These algorithms also do a good job of characterizing broad-band RFI. Time- and frequency-variable RFI signals are best detected by SK and SW tests.

Key words: Machine learning – algorithms – normality tests – spectral relative entropy – pulsars – PSR J1713+0747.

1 INTRODUCTION

Radio-frequency interference (RFI) are electromagnetic signals negatively impacting radio astronomical measurements. Both natural phenomena such as lightning strikes or the northern and southern lights and man-made devices such as radars, radio, television, cell phones, and satellites generate sources of RFI. Most of them are caused by using commodities as simple as a wireless telephone, an automotive radar installed on a car, or an aerial device flying close to the observatory. RFI may also be caused by failing electronics or by an open microwave located somewhere in a zone surrounding the telescope and leaking a radio signal. The amount of man-made RFI continues to increase as technology advances. For a recent review, see Saroff (2023). As an illustration, Fig. 1 shows several types of RFI typical for radio astronomy data. Currently, methods of RFI detection and removal are limited to the type of RFI, the position in which the excision algorithm is applied during the processing pipeline, and a radio telescope's hardware set-up (see e.g. Ford & Buch 2014). As the raw data are often averaged before any astronomical analysis, RFI becomes more capable of easily suppressing astronomical signals of interest and making them harder to study (Ramey et al. 2019).

In this work, we examine the excision of RFI from astronomical observations of transient phenomena in the radio sky. Examples of these sources are pulsars (Lorimer & Kramer 2005), Rotating Radio Transients (RRATs; McLaughlin et al. 2006), and Fast Radio Bursts (FRBs; Lorimer et al. 2007; Thornton et al. 2013). Observations of these sources are most commonly done by collecting the so-called filterbank data [power spectral density (PSD) of astronomical data displayed as a function of observing time and sky frequency]. An example pulse is shown for the first FRB in Fig.2.

Electromagnetic radiation from pulsars, RRATs, and FRBs arrive on Earth as extremely weak broad-band signals. As an extraterrestrial signal propagates through space, it passes through an environment, called the interstellar medium (ISM), which is full of free electrons. This causes the signal to become dispersed. As shown in Fig.2, the result of dispersion is that the lower frequency components of the signal get delayed from the higher frequency components. The time delay observed,

$$\Delta t = 4149\ \mathrm{s} \times \left(\frac{\mathrm{DM}}{\mathrm{cm^{-3}\ pc}}\right) \times \left[\left(\frac{f_{\mathrm{low}}}{\mathrm{GHz}}\right)^{-2} - \left(\frac{f_{\mathrm{high}}}{\mathrm{GHz}}\right)^{-2}\right], \quad (1)$$

where the dispersion measure, DM, is the integrated column of free electrons over the line of sight and f_{low} and f_{high} are, respectively, the low and high frequencies of the received band. This unique dispersion property separates celestial signals from other signals.

★ E-mails: natalia.schmid@mail.wvu.edu (NAS);
kevin.bandura@mail.wvu.edu (KB) duncan.lorimer@mail.wvu.edu (DRL)

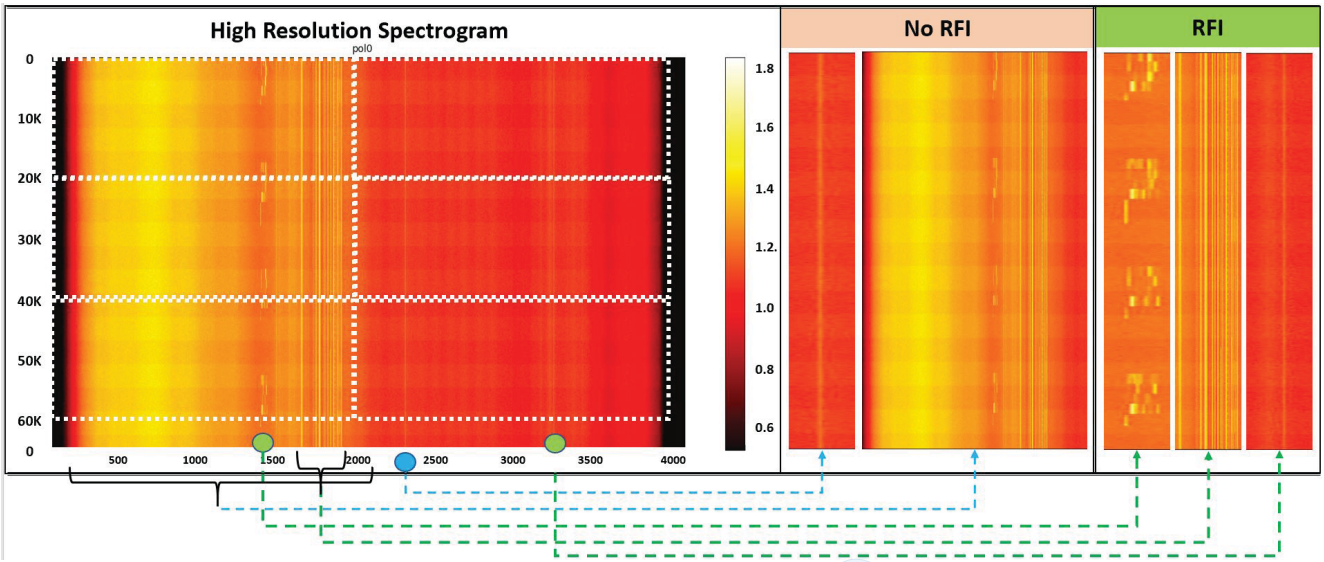

Figure 1. Snapshot of a sky-view high-resolution spectrogram (left), with the frequency channel along the x-axis (0 = 1900 MHz, 4096 = 1100 MHz), and time in y-axis (0–65 024 time samples or 0–0.33 s), increasing in the downwards direction. On the right, we see, from left to right, two cases of no RFI (should not be flagged), namely 'Milky Way Galaxy' around the discrete frequency channel 2300, a 'representation of a band-pass shape of a long bandwidth' from the frequency channel 200 to the frequency channel 2100, and three cases of RFI (should be flagged), namely, the 'Iridium SatCom signals' in the frequency channels 1402–1433 demonstrating periodicity, the unknown RFI in the range of discrete frequency channels from 1700 to 1900, and the 'Bedford Radar' in the frequency channel 3300.

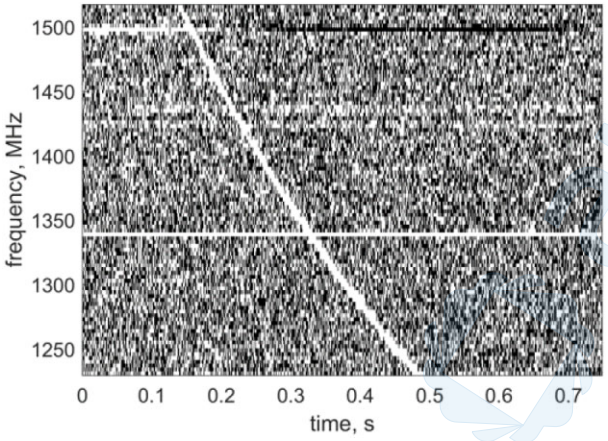

Figure 2. Example data set showing the 'Lorimer burst' (FRB 010724) in a frequency versus time plot. The white line sweeping left to right is the signal. The pixelated black and white background is noise being received at the same time as the pulse. Note the pulse is dispersed: the higher frequency components arrive earlier than their lower frequency counterparts.

By the time the signal is received on Earth, its signal strength, x_{source}, has decreased dramatically (typical power densities in the range –150 to –220 dBWm^{-2}, see e.g. Ford & Buch 2014) and it must compete with noisy signals produced from the instrumentation and thermal background of the receiver, x_{system}. In addition, any RFI that was transmitted across the same radio spectrum is also received as x_{RFI}. The resulting amplitude of the signal is a sum

$$x(t) = x_{\text{source}}(t) + x_{\text{system}}(t) + x_{\text{RFI}}(t), \quad (2)$$

where each component is a function of time, t. Even in very remote sites, terrestrial and orbital RFI signals can dominate the astronomical signal. New and improved RFI detection and characterization approaches will fully utilize the sensitivity of radio telescopes.

The goal of this research is to develop novel, high-level, and efficient, real-time RFI detection and flagging algorithms using inferential statistics and information theoretical measures in application to raw channelized voltages. In this paper, we will be working primarily with output from the Green Bank Telescope (GBT). On raw complex-valued channelized data, we explore the applications of symmetric and asymmetric spectral relative entropy (SRE; Ferrante, Masiero & Pavon 2011), spectral entropy (SE; Shen, Hung & Lee 1998), and Shapiro–Wilks (SW) test for normality (Shapiro & Wilks 1965). We will generate the resultant masks for each test. Since masks are generated through a thresholding procedure, different values of the threshold will be involved in testing to aid in determining the most effective value of the threshold. As a baseline for comparison with our algorithms, we will use two well-known RFI detection algorithms, spectral kurtosis (SK; Dwyer 1983) and Median Absolute Deviation (MAD; Buch et al. 2016).

The main contributions from our work are fourfold: (i) we propose using SE, SRE, and the SW test for normality as new methods of RFI detection in raw complex-valued channelized voltage data; (ii) the main constraint of our design is that channels must be processed independently for the benefit of parallel implementation in FPGA or GPU; (iii) we compare the performance of the proposed methods to that of MAD and SK and illustrate the benefit of applying each method to the channelized voltage data of PSR J1713+0747; (iv) we analyse the performance of each method by generating folded pulse profiles of PSR J1713+0747 from the data and evaluating its signal-to-noise ratio (S/N).

The remainder of this paper is organized as follows. Section 2 discusses the current state-of-the-art techniques for RFI detection and mitigation. In addition, it explains the basic foundations of inferential statistics and information theory techniques that were researched and developed for efficient, real-time RFI detection. Section 3 presents characteristics of the test data. Section 4 compares the observational and qualitative results of the various algorithms explored. It presents the results of different threshold values for RFI mask generation

and analyses the S/N of each method tested. Finally, in Section 5 we summarize the main findings of our work and also provide suggestions for future research.

2 RFI DETECTION AND MITIGATION TECHNIQUES

There are many different RFI detection and mitigation methods currently implemented for radio telescopes. The detection or mitigation technique used varies greatly depending on the type of interference, hardware implementation, and the processing pipeline step in which the excision method is applied (Ford & Buch 2014). However, not all excision methods are published. Moreover, they are often specific to the application the radio telescope is being used for, i.e. pulsar searches, FRBs searches, galaxy mapping, etc.

2.1 Processing pipeline

A radio telescope's receiver outputs data in time series, complex-valued voltages. These voltages are then converted to complex-valued channelized voltages where they are broken down into K frequency channels and N time samples (also called time bins). This is done by performing a short-time Fourier Transform (FT) over the time-series data. Here, each frequency channel represents a small portion of the receiver bandpass. Next, the pixel-wise power of the complex-valued channelized voltages is computed to create high-resolution PSD data known in radio astronomy as filter bank data (Lorimer & Kramer 2005). In computer science, it is known as a spectrogram (Flanagan 1972). Post-processing algorithms are applied to the complex-valued channelized voltages and spectrograms to sort the data and find astronomical signals of importance.

2.2 Current state-of-the-art RFI mitigation techniques

RFI can be mitigated in a variety of locations in the observatory pipeline. This includes regulatory methods which are applied before a signal is received at a radio telescope and technical processing methods which are applied at various locations throughout the receiver's pipeline (Ford & Buch 2014). A breakdown of each of these categories is shown in the block diagram in Fig.3.

2.2.1 Attenuation of terrestrial RFI

Before technical mitigation methods are applied, observatories take regulatory methods to negate the effects of RFI. These efforts start with the locations radio observatories are built. They are strategically placed in sparse population density areas so that the narrowband RFI produced by man-made devices can be minimized. In the United States, for example, a National Radio Quiet Zone (NRQZ) exists for this purpose. First established on 1958 November 19, by the Federal Communications Commission and the Inter-department Radio Advisory Committee on 1958 March 26, the NRQZ was formed to minimize possible harmful interference with the NRAO and the United States Navy. The NRQZ covers roughly 13 000 square miles of land[1] across West Virginia and Virginia, encompassing Green Bank Observatory (GBO) in Green Bank, West Virginia. For additional attenuation, electromagnetic shields, such as Faraday cages, are placed on-site around equipment and enclosures that

[1] https://greenbankobservatory.org/about/national-radio-quiet-zone

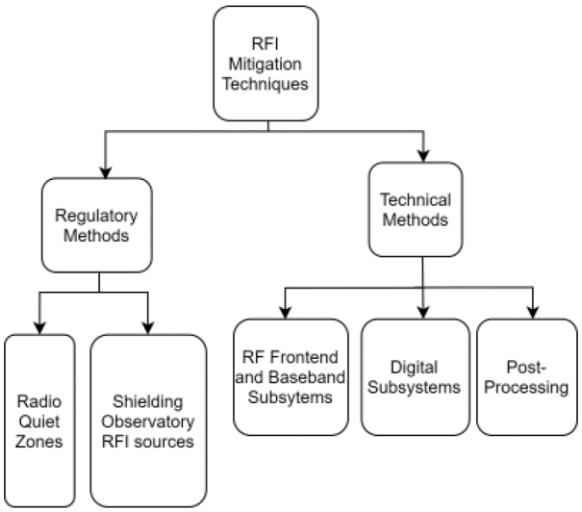

Figure 3. A breakdown of various places RFI mitigation can be performed as described by Ford & Buch (2014).

emit electromagnetic leakage (Ford & Buch 2014). However, the control over terrestrial RFI diminishes as more equipment is used at observatories. This increases the importance of RFI mitigation from other positions in the radio telescope pipeline. Analogue RFI excision is performed in the receiving system of the telescope. It is at this point that signal processing and learning excision methods can be applied (Ford & Buch 2014).

2.2.2 Edge-thresholding

This refers to a method to flag RFI against FRBs (Boyle & Sclocco 2019). It uses two unique characterization differences to flag regions of non-smooth, narrow, high-intensity data. First, it takes into account that FRBs are wider than RFI. Secondly, FRBs are pseudo-normally distributed. During edge-thresholding, data are processed iteratively across increasing window sizes. RFI becomes flagged when the difference between the window boundary and sample point is above a threshold, T, typically based on standard deviation or median absolute deviation. The algorithm is summarized by the decision rule

$$f(x_i) = \min(|x_i - x_0|, |x_i - x_w|) > T, \quad (3)$$

where the data window $w = (x_0, ..., x_w)$ and a point $x_i \in (x_1, x_{w-1})$ are flagged as RFI if they are greater than the set threshold T (Boyle & Sclocco 2019).

2.2.3 Spectral kurtosis (SK)

This is a well-known method for the analysis of non-stationary non-Gaussian signals. Its initial development (Dwyer 1983) was applied to improve the detection of distorted underwater acoustic signals. It was later applied to radio astronomical data by Nita et al. (2007) and is being increasingly used in this field to mitigate RFI. In terms of the principle of its operation, SK is based on the estimation of the fourth central moment known in probability theory as kurtosis in application to the data in the form of PSD. Nita et al. (2007) showed that SK is a robust estimator to distinguish Gaussian noise from non-Gaussian RFI using PSD data. It is based on a selection of M channelized power values P_k for each channel k from the spectrometer. Values that deviate from unity beyond analytically determined thresholds

are flagged. This is denoted with SK_k and done by constructing two sums

$$S_{1,k} = \sum_{m=1}^{M} P_k(m) \text{ and } S_{2,k} = \sum_{m=1}^{M} P_k^2(m).$$

The SK detection statistic is given as,

$$\mathrm{SK}_k = \frac{M+1}{M-1}\left(\frac{MS_{2,k}}{S_{1,k}^2} - 1\right). \quad (4)$$

Any data flagged outside of a threshold which is often chosen to be $\pm 3\sigma \approx \pm 6/\sqrt{M}$ on the SK_k is considered RFI with the threshold optimized for a given situation. Nita et al. (2007) have continued to improve upon this algorithm by generalizing it so the spectral averages may be taken before the SK estimator is calculated and using it in the two-bit digitized time domain (Gary, Liu & Nita 2010; Nita & Gary 2010; Nita et al. 2016; Nita, Keimpema & Paragi 2019; Taylor et al. 2019).

2.2.4 *Median absolute deviation (MAD)*

The MAD statistic for RFI detection in radio astronomy was proposed by Buch et al. (2016). Its FPGA prototype was later developed by Ramey et al. (2019) and by Buch et al. (2019). MAD uses the first-order statistic of the median to develop a decision rule to flag RFI. Its mathematical formulation is as follows.

Let the median of data set X be represented by

$$M_X = \text{median}(X), \quad (5)$$

and the median of the absolute deviation of the data set X from M_X be denoted by

$$\upsilon = \text{median}(|X - M_X|). \quad (6)$$

Given the two statistics M_X and υ, the MAD decision rule is formed by comparing the absolute deviation of any given point within the set X from the median M_X with the threshold $A\sigma_r$, where A is often chosen to be 3 but is optimized for particular situations. In general, we have

$$|x_i - M_X| \lessgtr A\sigma_r, \quad (7)$$

where x_i is the i-th sample point of the data set and the robust standard deviation is

$$\sigma_r = 1.4826 \times \upsilon. \quad (8)$$

Any sample outside the chosen deviation range is considered RFI.

2.3 Exploring statistical goodness-of-fit tests

Since the Gaussian nature of RFI-free complex channelized voltage data is the main feature for distinguishing between the RFI-free data and the data containing RFI, involving normality tests developed to differentiate between Gaussian and non-Gaussian statistics would be a natural approach to the problem of RFI detection. One of the most popular tests for normality in statistics is the Shapiro–Wilks (SW) test (Shapiro & Wilks 1965). In addition to its mathematical simplicity, it has the benefit of being easily parallelizable when implemented by hardware. It is for these practical reasons that SW is chosen over other similar tests such as the Anderson–Darling test.

The idea behind the SW normality is simple and elegant. Given a set of samples from a standard Gaussian distribution and a query set of samples, each sorted in the order of increasing values and then plotted in pairs, if a straight line can be fitted to the pairs of sorted samples, then the query set is Gaussian in its nature. Otherwise, the Gaussian hypothesis is rejected. To describe the test mathematically, Shapiro and Wilks developed a dimensionless statistic by solving a generalized least-squares problem. The developed statistic is described as

$$W = \frac{\sum_{i=1}^{n} a_i x_{(i)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (9)$$

where x_i is the original unsorted i-th sample, $x_{(i)}$ is the i-th sorted sample, $\bar{x}$ is the sample mean. The coefficients a_i form a vector

$$(a_1, \ldots, a_n) = \frac{\mathbf{m}^T\mathbf{V}^{-1}}{(\mathbf{m}^T\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}, \quad (10)$$

where $\mathbf{m}$ is the vector of sorted mean values of the samples from a standard Gaussian distribution and $\mathbf{V}$ is the covariance matrix of the sorted samples from the same standard Gaussian distribution. Based on the statistical analysis performed by Shapiro and Wilks, the hypothesis that the set of query samples is Gaussian is accepted if the test's p-value (the probability that the Gaussian distribution occurred by chance) is larger than the α-level (a preset conditional probability of error) of the test. The Gaussian hypothesis is rejected otherwise.

2.4 Exploring information theoretical performance metrics

As a subject, information theory (IT) characterizes the achievable limits in designing efficient, high-performance communication systems (Cover & Thomas 2006). Attributed to Shannon (1948), in the past 70 yr, IT grew into a large discipline, overlapping with and in part encompassing both statistics and physics. As a result, the concept of entropy has a strong presence in both physics and IT. Entropy in physics was developed as a precise mathematical way of testing if the second law of thermodynamics holds in a particular process. Entropy in IT was developed to quantify the uncertainty (average self-information) in a random variable and the limit of lossless compression (Shannon 1948). Relative entropy also called the Kullback–Leibler divergence was proposed as a metric to quantify the penalty for using the wrong probability distribution in lossless data encoding. It was later realized that it can be treated as a distance between two probability distributions (Moulin & Veeravalli 2019). Although relative entropy is not a real distance, since it does not satisfy the triangular inequality, it is a popular means to differentiate between two probability distributions in communication theory.

Before introducing new IT-based statistics for testing the Gaussianity of channelized complex voltages, we formally define entropy and relative entropy. Given a probability mass function (pmf) $p(x)$ of a random variable X, its entropy is

$$H(X) = -\sum_{x} p(x)\log p(x), \quad (11)$$

i.e. the average negative logarithm of the probability. The relative entropy, $D(p, q)$, between two pmfs $p(x)$ and $q(x)$, defined on the same set of outcomes, is the average of the log-likelihood ratio of $p(x)$ to $q(x)$, where the average is evaluated with respect to $p(x)$. The relative entropy is therefore

$$D(p, q) = \sum_{x} p(x)\log\frac{p(x)}{q(x)}. \quad (12)$$

Spectral entropy (SE) is a spectral tool developed for speech signal processing (Shen et al. 1998). Unlike SK, which relies on the estimates of the mean and variance of pixel intensities in a spectral

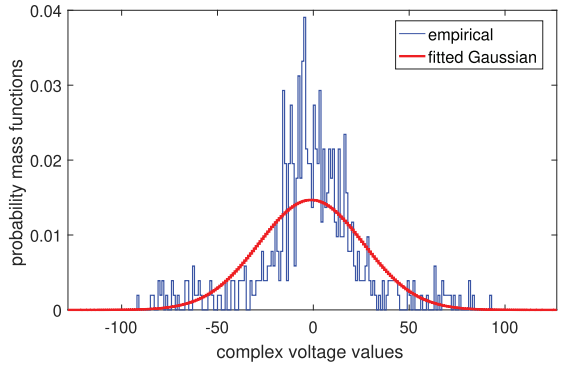

Figure 4. The bar plot shows the empirical probability mass function of the real part of complex voltage values in channel 1830. The smooth line is the fitted Gaussian distribution with the mean and variance of the empirical data. The substantial deviation of the shape of the empirical probability mass function from the fitted Gaussian pdf is due to the presence of strong RFI in channel 1830.

channel and on their ratio, SE is evaluated using the estimate of the probability mass function for each possible digitized voltage level, e.g. $2^8 = 256$ possible values of x for 8-bit data. Similar to the SW test, SE relies on the assumption that the RFI statistic is not Gaussian. First, SE per channel is evaluated (following equation 11) together with the sample estimate of the digitized voltage variance per channel. Then the entropy of a Gaussian random variable is evaluated

$$H_{\rm base}(X) = \frac{1}{2}\left[1 + \log\left(2\pi\sigma^2\right)\right], \tag{13}$$

where σ^2 is the estimated variance in an analyzed channel, and the entropy has units of nats, if we use the natural logarithm. The rest of the analysis relies on the fact that RFI-free channels have entropy close to the Gaussian entropy with the variance of the channel, and thus, the absolute difference in entropy values $|H(X) - H_{\rm base}(X)|$ is almost zero.

To demonstrate the difference between an actual histogram of the time samples per frequency channel and a normalized fitted Gaussian curve with the mean and variance of the empirical data, both are displayed in Fig. 4. The histogram has the number of bins based on the bit-resolution of the complex-valued channelized voltages. An 8-bit signed complex-valued channelized voltage would have $2^8 = 256$ bins ranging from –127 to 128.

To illustrate the potential of SE for the detection of RFI, Fig. 5 displays the absolute difference between the empirical and Gaussian SE values as a function of the number of spectral channels and time samples grouped by 512 original (high resolution) time samples. The original channelized voltage data block of size $65\,024 \times 4096$ is partitioned into 127 non-overlapping segments, each of size 512×4096. The 512 time samples per channel are used to calculate a single value of $|H(X) - H_{\rm base}(X)|$. Note that the absolute difference SE can easily detect several sources of RFI present in the data.

Finally, the RFI detection rule implements the modified Z-score, an efficient statistical method for detecting data outliers (Iglewicz & Hoaglin 1993), applied to the values of $|H(X) - H_{\rm base}(X)|$. The modified Z-score is the same outlier detection method as in the MAD algorithm [see (5) through (8)]. The only difference is that the data points in the MAD rule are replaced with the values of $|H(X) - H_{\rm base}(X)|$. This choice of the decision rule ensures that channels are treated independently, enabling a parallel implementation on GPU.

SRE yields a powerful test for Gaussianity, provided that the reference distribution $p(x)$ is selected to be normalized Gaussian with the mean and variance estimated per channel from empirical data. Unlike SE, which gains its power due to the subtraction of the Gaussian entropy, SRE relies not only on the difference in shapes of the two involved pmfs but also on the difference in terms of their higher order statistics.

Similar to the computation of SE, we first find the estimate of the pmf of the channelized quantized voltages (in our example, time segments are grouped in sets of 512 original samples). As a second step, we evaluate the sample mean and sample variance per channel per segment. Next, we fit a Gaussian probability density function with the sample mean and variance of the data in the channel minimizing the least-squares metric. Since the empirical pmf is based on an 8-bit representation, the Gaussian pdf is sampled at 256 locations of the empirical pmf bin centers. Finally, the relative entropy between the fitted Gaussian and the empirical pmf of channelized voltage levels (over a given time segment) is evaluated. The right panel in Fig. 5 displays the plot of SRE as a function of the number of spectral channels and time samples grouped by 512 original (high resolution) time samples. Note that the information in this plot is much more refined than the information provided in the plot of the normalized spectral entropy. This difference is attributed to the fact that the relative entropy measure contains information not only about the shape of two individual probability density functions but also about other high-order statistics describing the data.

Similar to the case of SE, the detection rule implements the modified Z-score but is applied to SRE values, with a Z-score magnitude greater than 3 often chosen as a threshold, but is optimized for a given environment. In addition to the relative entropy between theoretical and estimated pmfs, we also look at the symmetrical case of SRE formed by the summation of two asymmetrical SREs:

$$\begin{aligned} D_{\rm sym}(p,q) &= D(p,q) + D(q,p) \\ &= \sum_x p(x)\log\frac{p(x)}{q(x)} + \sum_x q(x)\log\frac{q(x)}{p(x)}. \end{aligned} \tag{14}$$

To differentiate between the two cases of SRE, we name the symmetrical case as $\rm SRE_s$ and the asymmetrical case as $\rm SRE_a$.

3 DATA

In this section, we describe the characteristics of the data that RFI detection tests were performed on. All data are illustrated with the highest frequency channel as the lowest frequency in the bandwidth and the lowest frequency channel as the highest frequency in the bandwidth. To be more specific, the data in use were collected over a bandwidth of 800 MHz partitioned into 4096 non-overlapping frequency channels. Channel 4096 corresponds to 1100 MHz, whereas channel 1 corresponds to 1900 MHz. The data are in the form of high-resolution complex-valued channelized voltages at two polarizations, polarization 0 and polarization 1. To calculate the PSD from the complex-valued channelized voltages, real and imaginary parts at each discrete time and frequency location are squared and summed.

The aforementioned RFI detection methods are tested on observations containing both RFI and pulsar signals. The pulsar in question, PSR J1713+0747, is a millisecond pulsar with period $P = 4.5$ ms, a pulse width of 1 ms, and a DM of 15.97 $\rm cm^{-3}$ pc (Foster, Wolszczan & Camilo 1993). These data also contain known RFI from: (i) the Iridium satellite communication system over the frequency range 1620–1626 MHz (channels 1402–1433); (ii) FAA radar originating from the Bedford, NC station is seen at

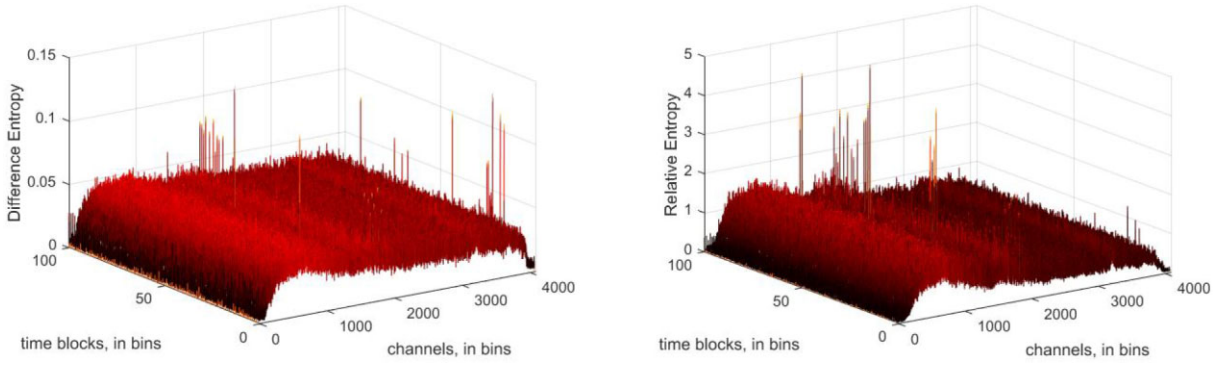

Figure 5. Left: Normalized spectral entropy as a normality test. Right: Spectral relative entropy as a test for normality.

Table 1. This table describes the characteristics of the data set containing astronomical pulses, RFI, and noise.

PSR	J1713+0747
Sampling interval (μs)	5.12
Length of data (s)	1.6646144
Number of time samples	325 120
Number of frequency channels	4096
Bandwidth (MHz)	800
Centre frequency (MHz)	1500
Number of bits	8 signed

1255 and 1305 MHz (channels 3302 and 3046); (iii) a collection of unknown sources that exists around 1500 MHz (channel 2048). Periodic RFI from GPS-L3 Communications is found at 1381 MHz (channel 2657).

The data were collected by our colleagues at GBO on eight different occasions and thus spaced in time and saved as eight distinct data files each of size 5 GB. The raw complex voltages saved in each file were sampled at a Nyquist frequency of 1600 MHz, then converted to channelized voltages by means of the short-time FT. The complex channelized data files are available to us as real and imaginary parts each of size 4096 frequency channels, 325 120 time samples, and two polarizations. After channelization, the sampling interval is reduced to 5.12 μs. There are about 1.66 s of test data in each file. This means that the number of time samples representing a complete period of the pulsar is about 892 and the maximum number of pulses that can be found in the test data is approximately 370. The right column of Table 1 summarizes the parameters of the data set.

Each file was shared with us in Matlab data file format and was saved in five non-overlapping chunks of size 65 024-by-4096 at a time resulting in a phase discontinuity of the pulsar pulse at the end of each 65 024-th time sample. Therefore, to avoid any misleading results, we partitioned each file into five chunks, each containing 65 024 time samples and 4096 frequency channels. To distinguish between files and chunks, we named each file as *mat number*, with numbers ranging between 0 and 7, and each chunk as *chunk number*, with numbers ranging between 0 and 4.

Each data chunk was further broken down into non-overlapping, consecutive segments containing all of the frequency channels and 512 time samples. Thus, 127 segments of 512 time samples and 4096 frequency channels were formed per each *chunk number* file.

The MAD algorithm, SW test for normality, SK, SRE, and SE are applied to each data segment. Since there is no way to definitively know what and where RFI signals are, there is no definitive ground truth. To analyse the effectiveness of the newly proposed methods in the detection and mitigation of RFI, we compare their performance on the pulsar signal to noise against the performance of the MAD and SK methods, both known in the literature, on the same metric.

To see a pulse in the data of J1713+0747, several processing steps must be applied. First, the channelized voltages must be converted to PSD by summing the squares of both the real and imaginary valued channelized voltages. Next, the data must be dedispersed using the DM value of the pulsar. Subsequently, integration of the dedispersed data over frequency components is completed. A depiction of a single chunk of *mat* 0 after the application of the signal processing steps is shown in Fig. 6. Several pulses of the pulsar are clearly seen in each panel between 0.1 and 0.15 s.

4 EXPERIMENTAL RESULTS

Given raw channelized voltage data as described in Section 3 and a list of prospective RFI detection and mitigation methods applied to the raw data, we now illustrate the performance of the RFI detection and mitigation methods. We adopt S/N as an objective measure of performance. The S/N of a single folded pulse is a traditional metric to measure the quality of astronomical signals when searching for pulsars (Lorimer & Kramer 2005).

After the inspection of the eight data files, we selected two files, *mat* 0 and *mat* 2, due to the unique types of RFI present in the data. The first file *mat* 0 contains several broad-band RFI signals, while *mat* 2 has a presence of strong RFI signals varying in frequency. Both types of RFI present challenges for modern RFI detection methods. Fig. 6 and Tables 2–6 display the results of our analysis of five different RFI detection methods defined in Sections 2.2, 2.3, and 2.4 in application to the five chunks of *mat* 0. Tables 7–11 demonstrate the results of our analysis in application to the five chunks of *mat* 2.

4.1 Performance analysis of *mat* 0

As mentioned earlier, the data file *mat* 0 was selected for analysis due to its unique content. The file contains several broad-band RFI signals. One of them is shown in the form of 'RFI masks' in Figs 7–10. Since complex channelized voltage data are represented by real and imaginary parts, a mask is generated per each part, then a single combined mask is generated as a product of the two masks. Different RFI detection methods are applied to *chunk* 0 of *mat* 0 yielding several combined masks, one per each method. The chunk is of size 65 024 time samples and 4096 frequency channels. It is partitioned into 127 non-overlapping segments, each composed of 512 time samples and 4096 channels. The RFI detection methods are applied to 512 time samples in every channel and every segment. If a

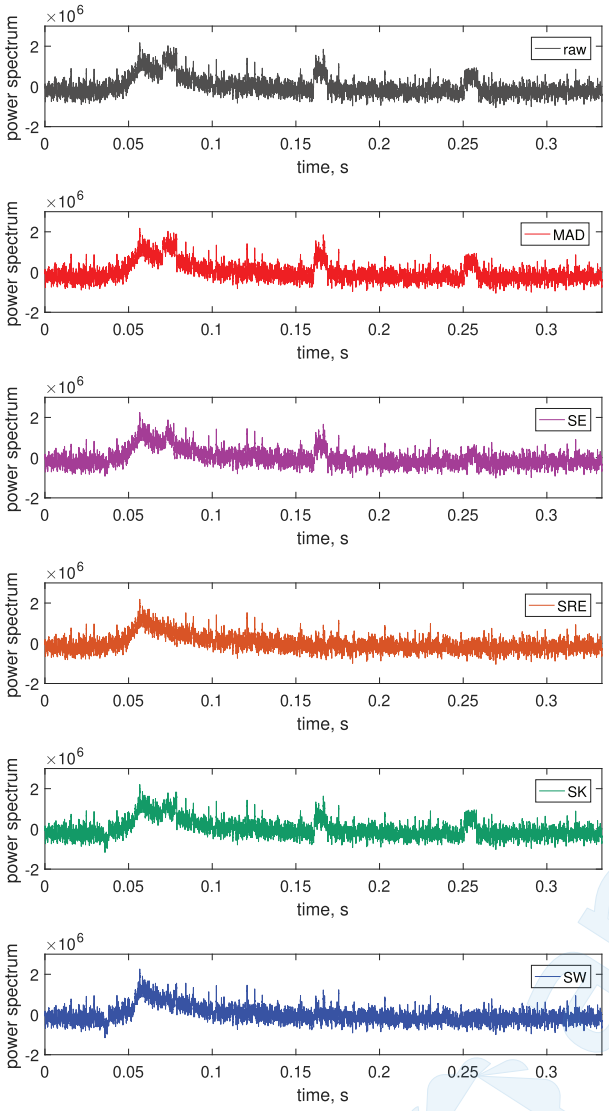

Figure 6. Power spectrum displayed as a time series. The time series contains 57 pulsar pulses embedded in noise and the remaining RFI.

Table 2. The S/N values on *chunk 0* of *mat 0* for different thresholds and different RFI removal methods of median absolute deviation (MAD), spectral entropy (SE), symmetrical spectral relative entropy (SRE_s), asymmetrical spectral relative entropy (SRE_a), spectral Kurtosis (SK), and Shapiro–Wilks (SW). The maximum value of each method is marked in bold. The S/N of the raw data is 15.81.

	chunk 0 of mat 0						
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	13.65	15.21	15.96	15.88	15.05	15.52	0.01
3.5	15.03	15.95	**16.33**	16.11	15.49	15.50	0.005
4	15.78	**16.35**	16.14	**16.43**	15.60	15.64	0.0025
4.5	15.82	16.30	16.29	16.38	15.83	16.18	0.001
5	15.93	16.29	16.30	16.34	15.90	16.35	0.0005
5.5	**15.94**	16.23	16.21	16.20	16.17	16.33	0.00025
6	15.89	16.20	16.23	16.20	15.97	**16.36**	0.0001
6.5	15.87	16.15	16.28	16.17	**16.18**	16.12	0.00005
7	15.85	16.07	16.23	16.20	16.18	16.13	0.00001

Table 3. The S/N values on *chunk 1* of *mat 0* for different thresholds and different RFI removal methods. The S/N of the raw data is 17.42.

	chunk 1 of mat 0						
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	15.68	17.50	17.29	17.78	17.42	16.72	0.01
3.5	16.99	16.88	**17.48**	**18.12**	17.33	17.31	0.005
4	17.40	17.06	17.35	17.55	17.26	17.32	0.0025
4.5	**17.45**	17.22	17.34	17.04	17.10	17.05	0.001
5	17.41	17.28	17.33	17.15	**17.47**	17.26	0.0005
5.5	17.42	17.47	17.23	17.33	17.42	17.45	0.00025
6	17.42	**17.62**	17.31	17.33	17.40	**17.54**	0.0001
6.5	17.43	17.49	17.25	17.31	17.36	17.49	0.00005
7	17.42	17.43	17.22	17.28	17.44	17.39	0.00001

Table 4. The S/N values on *chunk 2* of *mat 0* for different thresholds and different RFI removal methods. The S/N of the raw data is 16.92.

	chunk 2 of mat 0						
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	13.53	16.12	15.78	16.22	**17.17**	15.66	0.01
3.5	15.33	16.73	15.81	16.30	17.13	16.18	0.005
4	16.17	16.92	16.55	16.62	17.00	15.59	0.0025
4.5	16.48	16.90	16.98	16.66	16.87	16.38	0.001
5	16.58	16.73	17.04	17.22	17.03	16.91	0.0005
5.5	16.66	16.92	**17.25**	**17.33**	16.91	16.96	0.00025
6	16.73	17.24	17.23	17.23	16.86	**17.10**	0.0001
6.5	16.82	17.30	17.16	17.17	16.96	17.07	0.00005
7	**16.87**	**17.32**	17.20	17.23	16.70	16.70	0.00001

Table 5. The S/N values on *chunk 3* of *mat 0* for different thresholds and different RFI removal methods. The S/N of the raw data is 16.03.

	chunk 3 of mat 0						
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	13.26	16.18	15.96	16.39	16.00	15.17	0.01
3.5	14.60	16.26	15.76	15.62	16.11	15.47	0.005
4	15.29	16.21	16.03	15.97	16.06	15.79	0.0025
4.5	15.56	16.36	15.56	16.63	16.52	15.88	0.001
5	15.82	16.41	16.70	16.85	**16.58**	16.25	0.0005
5.5	15.99	**16.43**	**16.92**	16.85	16.55	16.27	0.00025
6	**16.04**	16.01	16.78	**16.87**	16.56	16.36	0.0001
6.5	16.04	15.81	16.60	16.69	16.56	**16.38**	0.00005
7	16.03	15.68	16.75	16.67	16.43	16.33	0.00001

Table 6. The S/N values on *chunk 4* of *mat 0* for different thresholds and various RFI removal methods. The S/N of the raw data is 16.93.

	chunk 4 of mat 0						
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	13.66	16.45	15.42	15.41	16.25	16.12	0.01
3.5	15.51	**16.96**	16.03	16.50	16.80	16.70	0.005
4	16.32	16.61	16.72	16.49	**17.25**	**16.81**	0.0025
4.5	16.65	16.33	16.51	16.57	17.10	16.52	0.001
5	16.78	16.43	16.63	16.58	16.84	16.43	0.0005
5.5	16.94	16.42	16.61	16.57	16.64	16.57	0.00025
6	16.94	16.58	16.76	16.65	16.62	16.43	0.0001
6.5	**16.95**	16.22	**16.81**	16.70	16.37	16.48	0.00005
7	16.95	16.26	16.74	**16.70**	16.55	16.75	0.00001

Table 7. The S/N values on *chunk* 0 of *mat* 2 for different thresholds and different RFI removal methods. The S/N of the raw data is 14.06.

chunk 0 of mat 2							
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	12.64	**16.56**	15.72	15.56	16.54	14.31	0.01
3.5	14.10	16.41	15.86	15.89	**17.12**	15.39	0.005
4	14.43	16.25	16.06	**16.06**	16.48	**16.57**	0.0025
4.5	**14.53**	16.11	15.95	15.99	16.48	16.32	0.001
5	14.52	16.36	15.98	15.84	16.52	16.03	0.0005
5.5	14.42	16.38	16.11	16.01	16.34	16.40	0.00025
6	14.31	16.46	16.07	16.05	15.78	16.10	0.0001
6.5	14.18	16.40	15.99	16.06	15.86	16.08	0.00005
7	14.12	15.58	**16.12**	16.05	15.78	16.17	0.00001

Table 8. The S/N values on *chunk* 1 of *mat* 2 for different thresholds and different RFI removal methods. The S/N of the raw data is 17.50.

chunk 1 of mat 2							
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	15.68	16.36	**18.43**	**18.27**	17.00	**17.87**	0.01
3.5	17.17	17.47	17.93	17.60	17.21	16.97	0.005
4	17.70	17.92	17.68	17.90	**17.44**	16.51	0.0025
4.5	**17.89**	17.89	17.95	17.89	17.27	17.13	0.001
5	17.83	17.94	17.75	17.81	17.37	17.41	0.0005
5.5	17.77	18.01	18.01	17.94	17.07	17.44	0.00025
6	17.75	**18.18**	17.97	17.83	16.99	17.51	0.0001
6.5	17.75	18.00	17.88	17.96	17.41	17.61	0.00005
7	17.73	18.05	17.88	17.90	17.06	17.70	0.00001

Table 9. The S/N values on *chunk* 2 of *mat* 2 for different thresholds and different RFI removal methods. The S/N of the raw data is 13.93.

chunk 2 of mat 2							
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	15.21	13.46	**14.70**	14.23	17.92	18.17	0.01
3.5	16.61	13.63	14.25	**14.70**	17.71	17.97	0.005
4	**16.96**	13.83	14.09	13.60	17.51	**18.20**	0.0025
4.5	16.88	13.95	14.08	13.95	17.73	17.89	0.001
5	16.84	14.12	14.09	13.99	17.43	17.67	0.0005
5.5	16.68	14.34	13.98	14.05	17.72	17.71	0.00025
6	16.48	**14.36**	13.87	13.87	17.72	17.71	0.0001
6.5	16.31	14.21	13.81	13.88	17.66	17.71	0.00005
7	16.16	14.06	13.86	13.87	**18.18**	17.57	0.00001

Table 10. The S/N values on *chunk 3* of *mat 2* for different thresholds and different RFI removal methods. The S/N of the raw data is 15.53.

chunk 3 of mat 2							
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	14.36	18.43	16.45	16.61	18.42	18.25	0.01
3.5	15.79	18.49	16.64	16.94	17.92	18.33	0.005
4	16.46	**18.58**	17.24	17.09	18.44	18.34	0.001
4.5	**16.48**	18.39	17.45	17.42	18.81	**18.60**	0.0025
5	16.48	18.27	17.66	17.82	18.07	18.27	0.0005
5.5	16.16	18.14	17.84	17.97	18.24	18.28	0.00025
6	16.09	18.12	17.97	18.07	18.90	18.15	0.0001
6.5	16.07	18.14	**18.04**	18.04	18.88	18.15	0.00005
7	16.01	18.11	18.03	**18.13**	**18.91**	18.18	0.00001

Table 11. The S/N values on *chunk* 4 of *mat* 2 for different thresholds and various RFI removal methods. The S/N of the raw data is 15.54.

chunk 4 of mat 2							
Th	MAD	SE	SRE_s	SRE_a	SK	SW	α-level
3	13.64	15.55	14.81	14.52	15.76	14.13	0.01
3.5	15.07	15.61	15.38	14.93	16.22	15.64	0.005
4	15.49	15.83	15.38	15.73	16.37	16.05	0.0025
4.5	15.65	15.89	15.70	15.81	16.49	**16.25**	0.001
5	**15.69**	15.91	15.77	15.85	16.61	15.99	0.0005
5.5	15.62	15.87	15.92	15.85	**16.63**	16.04	0.00025
6	15.62	16.03	15.91	**16.06**	16.53	16.00	0.0001
6.5	15.62	16.00	**15.95**	15.98	16.50	16.14	0.00005
7	15.59	**16.10**	15.87	15.94	16.46	16.06	0.00001

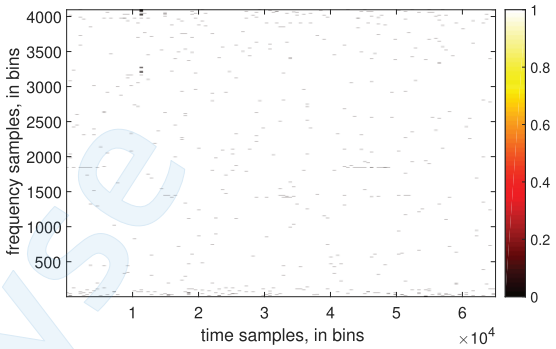

Figure 7. Mask generated by SE at the value of threshold set to 4σ. Small black intervals mark detected RFI, where the test rejected the Gaussian hypothesis. White intervals mark the part of the data where the test did not reject the Gaussian hypothesis.

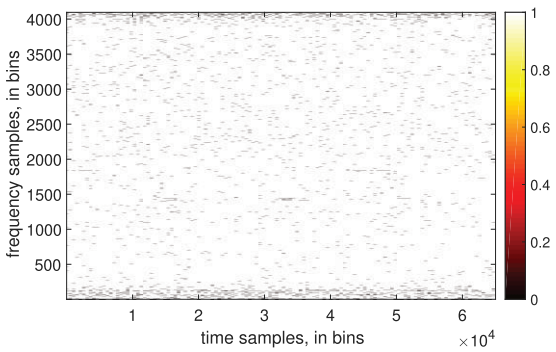

Figure 8. Mask generated by asymmetrical SRE at the threshold of 4σ. For further details, see Fig. 7.

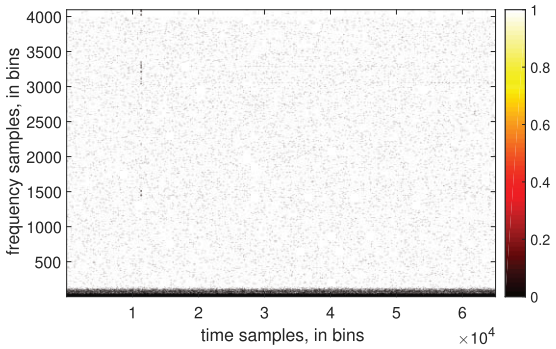

Figure 9. Mask generated by SK at the threshold of 3σ. For further details, see Fig. 7.

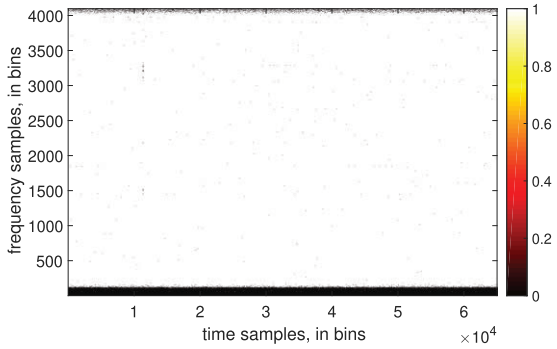

Figure 10. Mask generated by SW using an α-level set to 10^{-4}. For further details, see Fig. 7.

particular test detects the presence of RFI, the 512 time samples are replaced with zeros, otherwise, they are replaced with ones. Black lines and bars mark detected RFI while white space represents the portion of the data free of RFI as determined by each RFI detection method. The illustrations are provided for zero polarization of the data. The RFI masks are shown for SE at the threshold of 4σ, for SRE at the threshold of 4σ, for SK at values of the threshold of 3σ, and for SW at the α-level of 10^{-4}. Note that although SK and SW were applied for the detection of narrow-band RFI (along each frequency channel), they also captured broad-band RFI, unlike MAD, SE, and SRE methods. We do not show the RFI mask for MAD since, regardless of the threshold, it does not display any essential RFI signals.

After RFI masks are generated, several signal processing steps are applied to the data to arrive at the plots in Fig. 6. The steps are: (1) applying the combined RFI masks to real and imaginary parts of complex-valued channelized voltages (multiplying them one-by-one); (2) forming the power spectrum (spectrogram); (3) dedispersing the data; (4) integrating dedispersed data in frequency. The outcome is a power spectral time series.

Six integrated power spectral series are displayed in Fig. 6. The top panel shows the raw power spectrum. The second panel shows the power spectrum after the MAD method at 5.5σ was applied. The third from the top panel presents the power spectrum after applying SE at 4σ. The fourth panel shows the power spectrum after the application of SRE at 5σ. The fifth panel displays the power spectrum after applying SK at 6.5σ and the panel at the bottom shows the power spectrum after applying the SW test with α set to 10^{-4}. Note that the thresholds were selected to maximize the performance of each detection method as will be explained below.

To quantify the performance of the proposed RFI detection methods, we compute the S/N values of a folded pulse for different methods. The results are summarized in Table 2. To arrive at each S/N, a single folded pulse is generated from the power spectrum shown in Fig. 6 using RIPTIDE (Morello et al. 2020), a Python implementation of the fast folding algorithm (Staelin 1969). Table 2 displays the found S/N values as a function of varying thresholds and α-levels. Thresholds for the methods of MAD, SE, SRE, and SK are varied between 3σ and 7σ. The values of α-level used by SW are varied between 0.01 and 10^{-5}.

Table 2 provides insight into the best performance delivered by each RFI detection method, given the data partitioning as described in Section 3. MAD, one of the two baseline methods selected for performance comparison, achieves the maximum S/N value of 15.94 at the threshold of 5.5σ. This is slightly above the untreated (raw data) S/N of 15.81. When the threshold is set to 3σ, as recommended

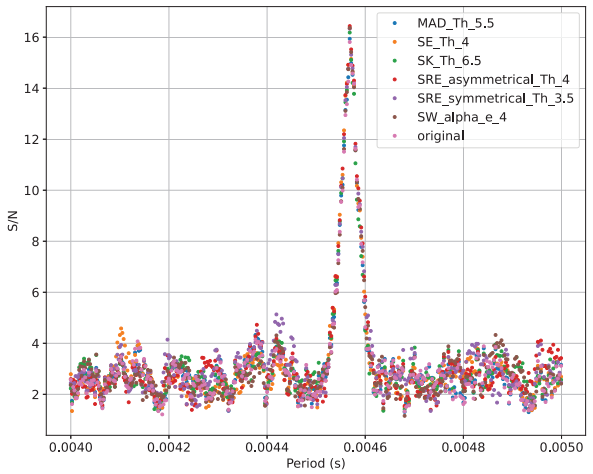

Figure 11. Periodogram results from RIPTIDE for the proposed and baseline RFI detection methods obtained from the data in *chunk* 0 of *mat* 0.

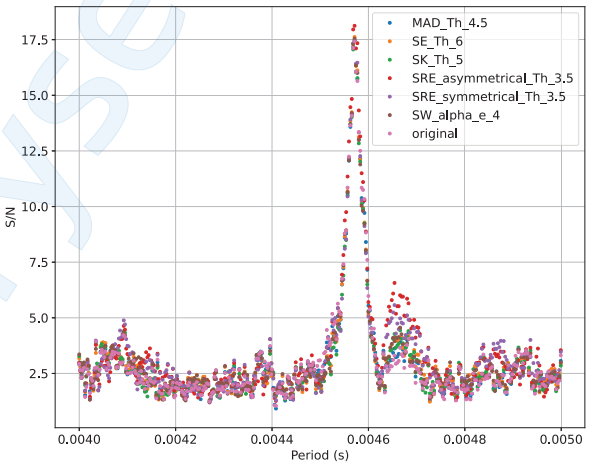

Figure 12. Periodogram results from RIPTIDE for the proposed and baseline RFI detection methods obtained from the data in *chunk* 1 of *mat* 0.

in Ramey et al. (2019), the S/N of MAD is below the S/N of raw (untreated) data. Each of the remaining four methods (SE, symmetrical SRE, asymmetrical SRE, SK, and SW), demonstrates a more significant performance improvement compared to MAD. As an example, SE achieves the best performance of 16.35 at 4σ, symmetrical SRE achieves S/N of 16.33 at the threshold 3.5σ, asymmetrical SRE achieves S/N of 16.43 at the threshold 4σ, SK reaches S/N of 16.18 at the threshold 6.5σ, and SW demonstrates the S/N value of 16.36 at α-level set to 10^{-4}. To conclude, our proposed methods are all better than the baseline methods of MAD and SK in the case of *mat* 0*chunk* 0. It should be also noted that asymmetrical SRE is performing better than symmetrical SRE. The plots of a single folded pulse for the choice of the best S/N value for the six RFI detection methods as well as for the original case are provided in Fig. 11.

To complete the analysis of *mat* 0, we process the data in *chunks* 1 through 4. The S/N value of raw data in *chunk* 1 of *mat* 0 is higher than the S/N of any other chunk. It is equal to 17.42 for *chunk* 1. Looking at the values provided in Table 3, there is no S/N value that is significantly higher than 17.42, pointing to the fact that the removal of RFI signals in this case is not that useful. None the less, our proposed methods of SE, symmetrical SRE, asymmetrical SRE,

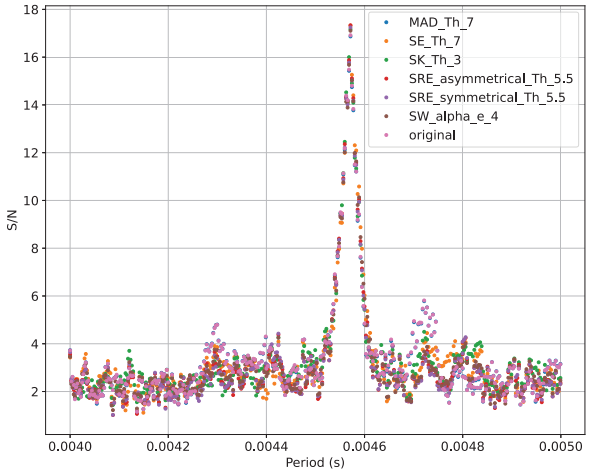

Figure 13. Periodogram results from RIPTIDE for the proposed and baseline RFI detection methods obtained from the data in *chunk* 2 of *mat* 0.

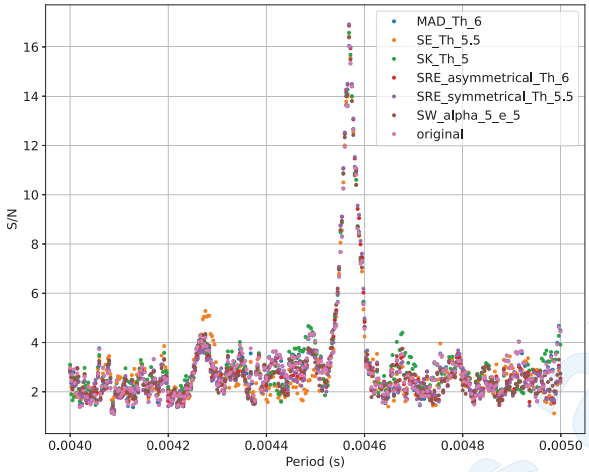

Figure 14. Periodogram results from RIPTIDE for the proposed and baseline RFI detection methods obtained from the data in *chunk* 3 of *mat* 0.

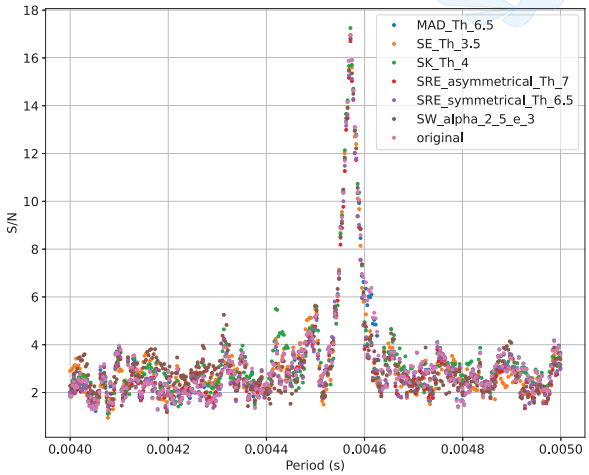

Figure 15. Periodogram results from RIPTIDE for the proposed and baseline RFI detection methods obtained from the data in *chunk* 4 of *mat* 0.

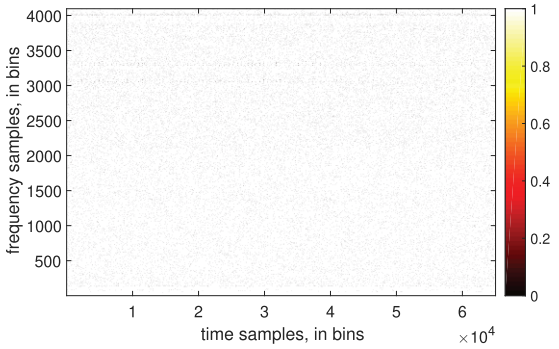

Figure 16. The mask generated by MAD at the threshold of 3σ when the MAD method is applied to *chunk* 2 of *mat* 2. For further details, see Fig. 7.

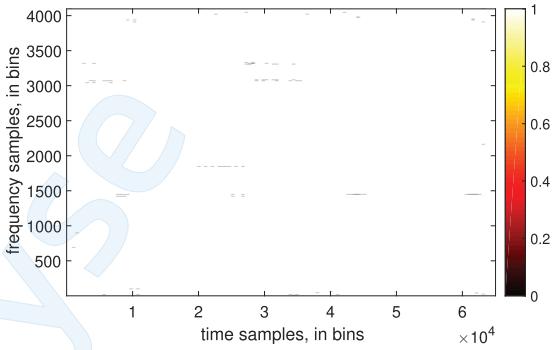

Figure 17. The mask generated by SE at the threshold of 6σ when the SE method is applied to *chunk* 2 of *mat* 2. For further details, see Fig. 7.

and SW all surpass the baseline methods of MAD and SK. The highest of the best S/N equal to 17.62 is achieved by SE with the threshold value set to 6σ, while the lowest of the best S/N equal to 17.45 is achieved by MAD with the threshold value set to 4.5σ. The respective single pulse plot for *chunk* 1 of *mat* 0 is provided in Fig. 12.

The improvement in S/N is much more noticeable when the S/N value of raw data is relatively low. For example, for *chunk* 2 (see Table 4) the S/N value of raw data is 13.93. The application of SK at a threshold of 6.5σ and SW at an α-level of 10^{-4} result in S/N values of 17.66 and 17.71, respectively, indicating that at a low value of S/N of the raw data, it is beneficial to detect and remove high in value RFI signals. The analysis of the S/N values as a function of the method of removal of RFI signals and varying threshold value (α-level for SW) in Tables 5 and 6 demonstrate a similar trend. The respective periodograms for the best values of S/N are shown in Figs 12–15.

4.2 Performance analysis of *mat* 2

The data in *mat* 2 contain another challenging type of RFI, a strong signal varying in frequency and time. While MAD, SE, and SRE RFI detection methods miss to flag this type of RFI signal which is demonstrated in Figs 16, 17, 18, and 19, SK and SW methods demonstrate the ability to detect and flag this type of RFI in *chunk 2* of *mat 2* as shown in Figs. 20 and 21. The S/N values for all chunks of *mat* 2 are displayed in Tables 7 through 11. Note that for *chunk* 2 flagging the RFI signal varying in frequency and time resulted in considerably improved S/N values for SK and SW compared to the S/N value of raw data. The plots of a single folded pulse for the choice of the best S/N value for the six RFI detection methods as

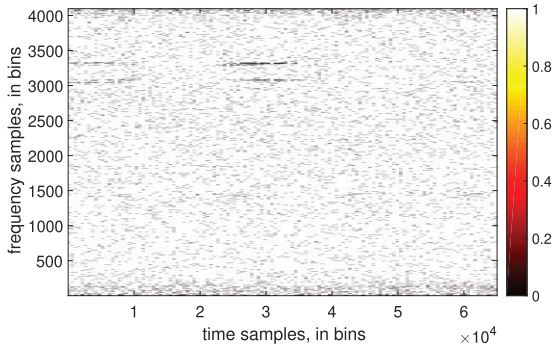

Figure 18. The mask generated by symmetrical SRE at the threshold of 3σ when the SRE method is applied to *chunk* 2 of *mat* 2. For further details, see Fig. 7.

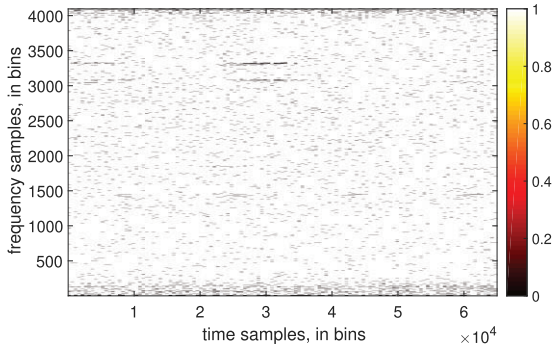

Figure 19. The mask generated by asymmetrical SRE at the threshold of 3.5σ when the asymmetrical SRE method is applied to *chunk* 2 of *mat* 2. For further details, see Fig. 7.

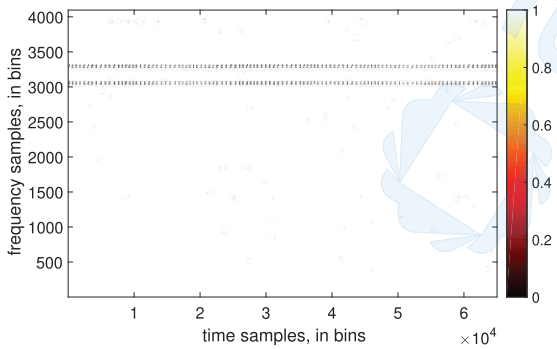

Figure 20. The mask generated by SK at the value of threshold set to 7σ when the SK method is applied to *chunk* 2 of *mat* 2. Note how well SK detects the RFI signals of varying frequency in the frequency range between 3000 and 3500. For further details, see Fig. 7.

well as for the case of raw data are provided in Figs 22 through 26 for *chunk* 0 through 4 of *mat* 2, respectively.

4.3 General observations

To summarize the performance of the tested methods for the detection and flagging RFI signals in astronomy data, the following general observations are made.

(i) In every analysed case, the application of SE, symmetrical SRE, asymmetrical SRE, SK, and SW resulted in an improved value

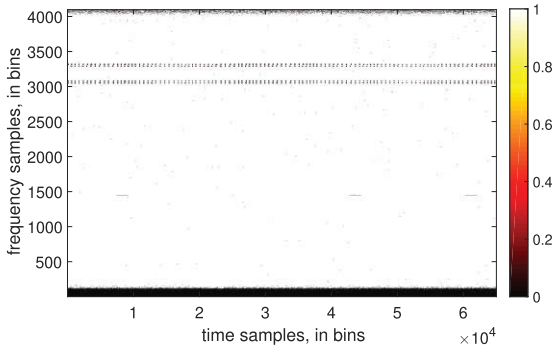

Figure 21. The mask generated by SW at the α level of 0.0001 when the SW method is applied to *chunk* 2 of *mat* 2. Similar to SK, SW detects RFI signals of varying frequency in frequency channels between 3000 and 3500. For further details, see Fig. 7.

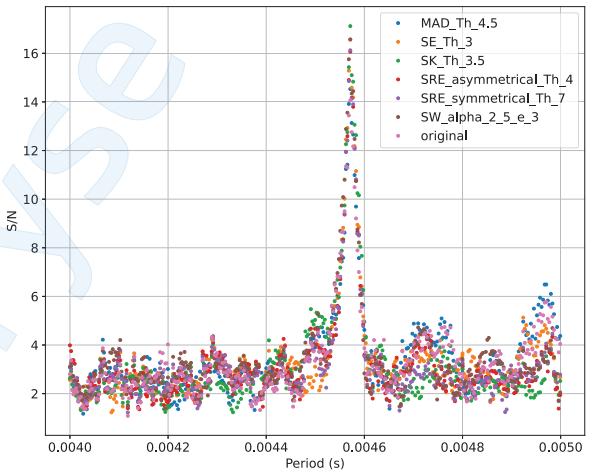

Figure 22. Periodogram results from RIPTIDE for and baseline RFI detection methods obtained from the data in *chunk* 0 of *mat* 2.

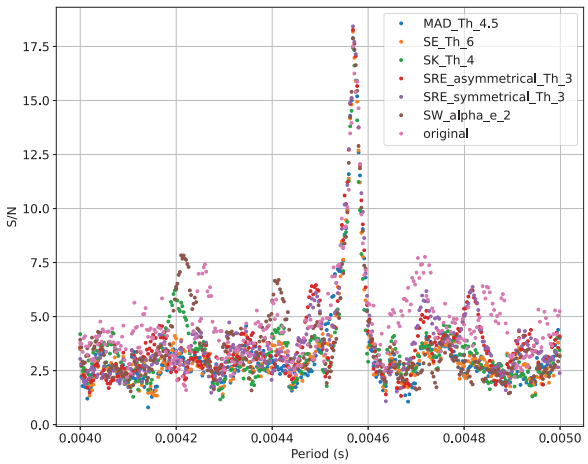

Figure 23. Periodogram results from RIPTIDE for and baseline RFI detection methods obtained from the data in *chunk* 1 of *mat* 2.

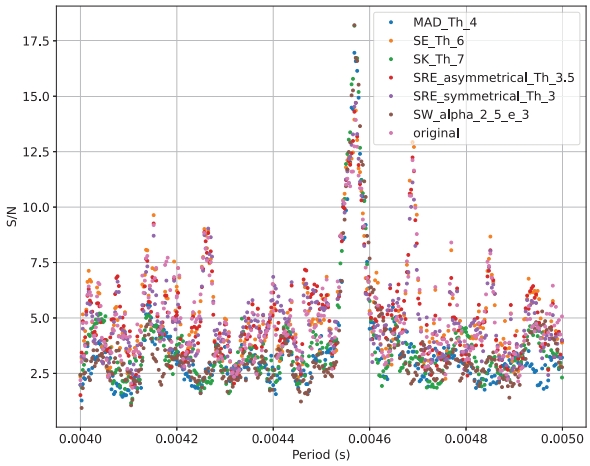

Figure 24. Periodogram results from RIPTIDE for and baseline RFI detection methods obtained from the data in *chunk* 2 of *mat* 2.

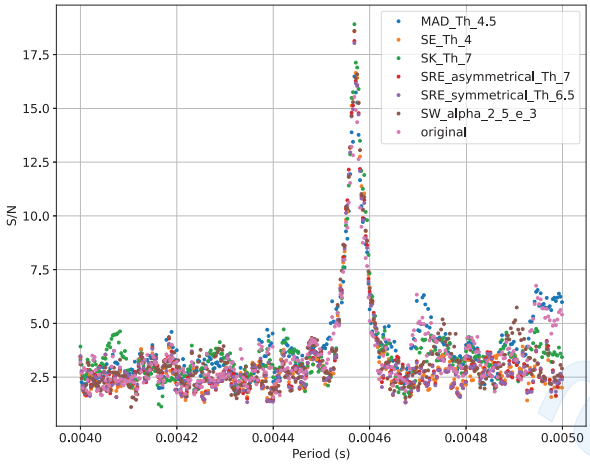

Figure 25. Periodogram results from RIPTIDE for and baseline RFI detection methods obtained from the data in *chunk* 3 of *mat* 2.

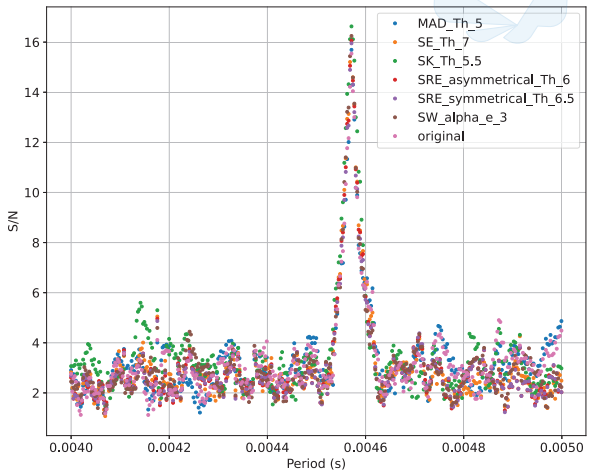

Figure 26. Periodogram results from RIPTIDE for and baseline RFI detection methods obtained from the data in *chunk* 4 of *mat* 2.

of S/N compared to the S/N of the raw data. Unlike the five methods above, the application of MAD on many occasions leads to a reduced value of S/N compared to the S/N of the raw data.

(ii) SE, symmetrical SRE, asymmetrical SRE, SK, and SW showcase their ability to detect broad-band RFI signals (e.g. *chunk* 0 in *mat* 0).

(iii) Varying in frequency and time RFI signals are best detected by SK and SW tests (see RFI in *chunk* 2 of *mat* 2) as well.

(iv) Raw channelized voltages yielding a high S/N of the folded pulse do not benefit from RFI detection and flagging methods.

(v) Asymmetrical SRE perfoms better than symmetrical SRE.

5 CONCLUSIONS

The range of statistical methods examined in this work is used as an indicator of how clean, RFI-free Gaussian distributed complex-valued frequency channel characteristics vary from the characteristics of RFI-contaminated channels. A demonstration of typical RFI environments was explored by applying MAD, spectral entropy (SE), spectral relative entropy (SRE), spectral Kurtosis (SK), and Shapiro–Wilks (SW) test for normality to complex-valued channelized voltage data collected with the GBT.

The S/N of a single folded pulse was selected as a means to compare the performance of the RFI detection methods. The application of MAD, SE, SRE, SK, and SW on the millisecond pulsar data of J1713+0747 illustrates that MAD does not always filter RFI effectively. Both MAD, SE, and SRE often keep the same RFI artefacts that are found in the original data. SK and SW successfully detect and remove both broad-band RFI signals and signals varying in frequency. All of the RFI detection tests except MAD increase the S/N of the pulsar data. In the future, further investigations of these methods on larger data sets are strongly encouraged.

DATA AVAILABILITY

The data used in this study are available upon request.

ACKNOWLEDGEMENTS

This research is partially supported by the National Science Foundation under Awards No. AST-2307581, the Natural Science Foundation of China (NSFC No. 61906149), and the Natural Science Foundation of Chongqing (cstc2021jcyj-msxmX1068). The authors would also like to thank their colleagues at the Green Bank Observatory and West Virginia University for providing the data set used throughout this research.

REFERENCES

Boyle J., Sclocco A., 2019, in RFI Workshop–Coexisting with Radio Frequency Interference (RFI). IEEE, New Jersey, p. 1

Buch K. D., Bhatporia S., Gupta Y., Nalawade S., Chowdhury A., Naik K., Aggarwal K., Ajithkumar B., 2016, J. Astron. Ins., 5, 1641018

Buch K. D., Naik K., Nalawade S., Bhatporia S., Gupta Y., Ajithkumar B., 2019, J. Astron. Instrum., 8, 1940013

Cover T. M., Thomas J. A., 2006, Elements of Information Theory. John Wiley and Sons, Inc, Hoboken, New Jersey

Dwyer R., 1983, in ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, New Jersey, p. 607

Ferrante A., Masiero C., Pavon M., 2011, in IEEE Transactions on Automatic Control, 57, 2561

Flanagan J. L., 1972, Speech Analysis, Synthesis and Perception. Springer-Verlag, New York

Ford J. M., Buch K. D., 2014, in 2014 IEEE Geoscience and Remote Sensing Symposium. IEEE, New Jersey, p. 231, http://ieeexplore.ieee.org/document/6946399/
Foster R. S., Wolszczan A., Camilo F., 1993, ApJ, 410, L91
Gary D. E., Liu Z., Nita G. M., 2010, PASP, 122, 560
Iglewicz B., Hoaglin D. C., 1993, How to Detect and Handle Outliers. ASQ Quality Press, Milwaukee, Wisconsin
Lorimer D., Kramer M., 2005, Handbook of Pulsar Astronomy. Cambridge Univ. Press, New York
Lorimer D. R., Bailes M., McLaughlin M. A., Narkevic D. J., Crawford F., 2007, Science, 318, 777
McLaughlin M. A. et al., 2006, Nature, 439, 817
Morello V., Barr E. D., Stappers B. W., Keane E. F., Lyne A. G., 2020, MNRAS, 497, 4654
Moulin P., Veeravalli V. V., 2019, Statistical Inference for Engineers and Data Scientists. Cambridge Univ. Press, New York
Nita G. M., Gary D. E., 2010, MNRAS, 406, L60
Nita G. M., Gary D. E., Liu Z., Hurford G. J., White S. M., 2007, PASP, 119, 805
Nita G. M., Hickish J., MacMahon D., Gary D. E., 2016, J. Astron. Ins., 5, 1641009
Nita G. M., Keimpema A., Paragi Z., 2019, J. Astron. Instrum., 8, 1940008
Ramey E., Joslyn N., Prestage R., Lam M., Hawkins L., Blattner T., Whitehead M., 2019, Seniors Honor paper/Undergraduate Thesis, Washington University in St. Louis
Saroff D., 2023, PhD thesis, Rochester Institute of Technology
Shannon C. E., 1948, Bell Syst. Tech. J., 27, 379
Shapiro S. S., Wilks M. B., 1965, Biometrika, 52, 591
Shen J.-L., Hung J.-W., Lee L., 1998, in ICSLP. IEEE, New Jersey
Staelin D. H. 1969, IEEE Proc., 57, 724
Taylor J., Denman N., Bandura K., Berger P., Masui K., Renard A., Tretyakov I., Vanderlinde K. 2019, J. Astron. Ins., 8, 1940004
Thornton D. et al., 2013, Science, 341, 53

This paper has been typeset from a TEX/LATEX file prepared by the author.