Search for low-energy periodic neutrino sources with ANTARES

Master's Thesis in Physics

presented by

Maximilian Eff

Date: 20.01.2023

Erlangen Centre for Astroparticle Physics
Friedrich-Alexander-Universität Erlangen-Nürnberg

Supervisor: Prof. Dr. Uli Katz

Abstract

ANTARES was a neutrino telescope submerged in the Mediterranean Sea and has been operating continuously from 2008 to its end in 2022. It is a Cherenkov detector equipped with a 3 dimensional matrix of hundreds of photosensors. Its main goal is the detection of high-energy neutrinos from astrophysical sources.

In this thesis, an analysis is developed and presented, to search for periodic low-energy neutrino fluxes in the photosensor counting rates of ANTARES. This switch from reconstructed neutrino events to the pure photosensor counting rates allows to undercut the energy threshold for the reconstruction process. The basis of this analysis is the Fast Fourier Transformation, which transforms the counting rates from the time domain into the frequency domain, allowing to easily identify hidden periodic signals contained in the rates. The Fourier power is chosen as the test statistic for this analysis, and its distribution under the assumption of only background described. Challenges posed by the available ANTARES data set are described and suitable solutions presented. Established techniques from pulsar astronomy are revised and included into this analysis. At the end, a sensitivity study for the implemented analysis and a short data set is performed, and the analysis for this blinded data set is executed.

Contents

1 Introduction

Astronomy is a field of science studying extraterrestrial phenomena around the the universe. It developed from the observation of the stars at the night sky with the bare eye to modern multi-messenger astronomy. Light and electromagnetic radiation is not anymore our solely access to studying the universe, but is nowadays expanded by the survey of cosmic rays, neutrinos and gravitational waves. Each of these messengers acts as carrier of unique information about its originating celestial body and/or the intermittent medium. The jointly consideration of these messengers therefore allows to surpass previously unattainable limits in the studying of the most extreme objects of the universe like black holes, neutron stars, gamma-ray bursts and many more. A recent milestone in multi-messenger astronomy is the first joint detection of a binary neutron star merger, via gravitational waves and multiple intervals of the electromagnetic spectrum [3].

Neutrino astronomy focuses on the neutrino messenger from astrophysical objects. Due to its exclusively weakly interacting behaviour, neutrinos rarely interacts with matter, unlike photons or cosmic rays. Due to this, neutrinos travel in a straight path from their source to Earth and therefore point back to their origin. This grants access to many astrophysical phenomena unavailable to the other messengers and allows to directly study the astrophysical source. Many theoretical models of astrophysical objects predict the generation of astrophysical neutrinos. However, so far the proven extraterrestrial sources of neutrinos are limited to the Sun and the supernova 1987A, as well as IceCube's identification of the blazar TXS 0506+056 in 2017 as another likely source [26] or IceCube's detection of an extraterrestrial diffuse high-energy neutrino flux of unknown origin [17].

The detection of neutrinos is a quite challenging task, as the neutrino only participates in the weak interaction. It can therefore not be directly measured. Instead, the usual detection principle works as follows. A neutrino interacting with matter, creates secondary charged particles like muons or hadronic showers. If such charged particles traverse a dielectric medium, like water, the medium along the path will be polarised and subsequently emit photons when returning to its ground state. If the speed of the charged particle surpasses the speed of light in said medium, the emitted spherical waves will overlap and constructively interfere with each other. This leads to the characteristic cone-shaped light emission called Cherenkov light. If the speed of the charged particle is below the local speed of light, then no constructive interference appears, as the wave fronts move faster, than new ones are emitted. This Cherenkov light can be detected using light detectors like photomultiplier tubes (PMT).

Based on this detection concept, astronomical neutrinos are searched for using so-called neutrino telescopes. These telescopes usually work by monitoring the Cherenkov light

produced by the secondary charged particles of neutrino interactions. For this, large amounts of PMTs observe the volume of a transparent medium. A neutrino event can be reconstructed from coincident hits in multiple PMTs. This allows to infer the arrival direction of the incident neutrino, and its energy. However, for a successful reconstruction process, the neutrino needs to surpass some energy threshold, that depends on the detector layout. Neutrinos below this threshold will not cause sufficient light emissions for hits in multiple PMTs. Due to the very low interaction rate of the neutrino, the observed volumes are usually huge to obtain a useful neutrino detection rate. Important neutrino telescopes are the ANTARES [10] and KM3NeT [22] project underwater in the Mediterranean sea, but also IceCube [20] deep in the antarctic ice or the deep underwater the Baikal Gigaton Volume Detector in Lake Baikal located in Russia [13].

Possible other approaches to identify a neutrino signal, besides the expensive reconstruction process of individual neutrino events, is to deduce changes of the neutrino flux directly from the PMT counting rate. By omitting the reconstruction, no neutrinos below the energy threshold are rejected, allowing to also detect low energy neutrino sources. Potential detectable variations could be for example a strong increase of the flux over a short period of time causing a short spike in the counting rate, or a periodic flux inducing a periodic pattern in the PMT rates. As an origin for spike shaped flux variations supernova explosions can be considered, which are with SN1987A already a verified neutrino sources. Nevertheless this is still an intensively researches in neutrino astronomy [6]. Pulsars have been considered by several works in literature as candidates for neutrino emission along different energy ranges over the last decades, and these will be the subject of this master thesis.

Sources of periodic neutrino fluxes may be pulsars, as a periodic neutrino emission similar to their periodic electromagnetic emission is conceivable.

Pulsars were first discovered in 1967 by Jocelyn Bell Burnell [19] and belong to the most extreme astrophysical objects. They are fast spinning neutron stars, dense objects with a mass on the order of one solar mass and a radius of about 10 km as well as with very strong magnetic fields ranging from 10^{11} G to 10^{15} G. Strong particle acceleration occurs at these magnetic poles and causes the emission of strong electromagnetic radiation from these regions. Pulsars differ from ordinary neutron stars due to their rotational axis and magnetic field axis being not aligned (see Figure 1.1). This causes a rotating beam of radiation, which will be perceived by an observer within the traversed path as a pulsating emission. Typical pulsar periods range from milliseconds to hundreds of seconds.

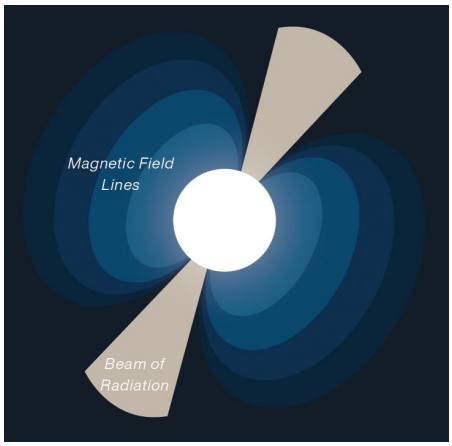


Figure 1.1:
Simple sketch of a pulsar showing the characteristic emission cones at the magnetic poles. Courtesy NASA/JPL-Caltech.

Various models attempt to explain these violent astrophysical objects, several also proposing the

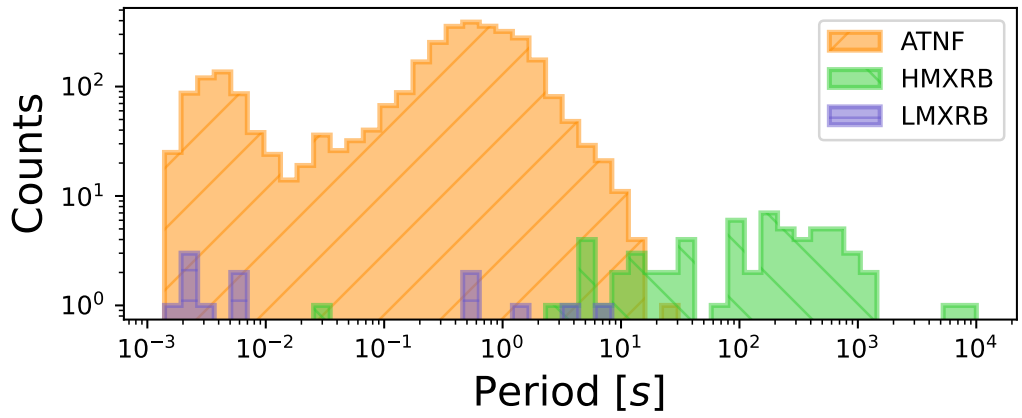


Figure 1.2:
The spin period distribution of known and catalogued pulsars (ATNF), high-mass X-ray binaries (HMXRB) and low-mass X-ray Binaries (LMXRB).

emission of neutrinos. They can be divided into
two categories. First the production of low energy neutrinos in the $\approx 100\,\text{keV}$ to $\approx 1\,\text{MeV}$ range due to pair-annihilation processes in the hot stellar plasma, e.g. [27], and secondly the production of high energy TeV neutrinos in pulsar wind nebular, a nebula surrounding the central pulsar, due to the production and subsequent decay of charged pions, e.g. [8].

So far however no neutrinos emitted from pulsars have been detected. Previous analyses investigated correlations between the arrival direction of neutrinos and known pulsars, e.g. [2], and the expected diffusive neutrino flux emitted from pulsar populations, e.g. [18]. From the absence of any significant neutrino detections, upper limits about the neutrino emissions from the examined pulsars were set.

An extensive list of known pulsars and their properties from multiple surveys is the Australia Telescope National Facility (ATNF) Pulsar Catalogue [25, `http://www.atnf.csiro.au/research/pulsar/psrcat`]. Figure 1.2 displays the spin period distribution of the pulsars of this catalogue.

In this thesis the development of an analysis in the search of a periodic neutrino signal in the ANTARES PMT counting rates is described. Established pulsar search tools from radio and gamma astronomy are investigated for their possible use in neutrino astronomy.

1.1 ANTARES Neutrino Telescope

The ANTARES neutrino telescope is located in the Mediterranean Sea, 40 km off the coast of Toulon, France, in a depth of 2.5 km. It detects neutrinos by observing the Cherenkov radiation produced in sea water from secondary charged leptons, originating from the weak interaction of the neutrino with the detector medium. The Cherenkov

light is detected with a 3 dimensional grid of photomultiplier tubes (PMT) in the deep sea. Each node in the grid is occupied by a so-called Optical Module Frame (OMF), containing three Optical Modules (OM), which act as containers for the PMTs, as well as the Local Control Module (LCM), the required on site electronics. Part of the LCM are the Analogue Ring Sampler (ARS) chips, which digitizes the analogue signal from the PMTs. To minimize the dead time by the digitization process, each OM is equipped with two ARS. 25 vertically aligned OMFs form a line and are connected to each other and the seabed, from where the cables go to the shore station for further data processing. In total there are 885 installed OMs [5]. Figure 1.3 displays the ANTARES detector layout.

The neutrino events are reconstructed from causally connected Cherenkov hits in multiple PMTs. For a successful reconstruction process, the incoming neutrino needs to surpass some energy threshold, which is for the ANTARES layout about 20 GeV [5]. In Figure 1.4 the neutrino detection principle is illustrated.

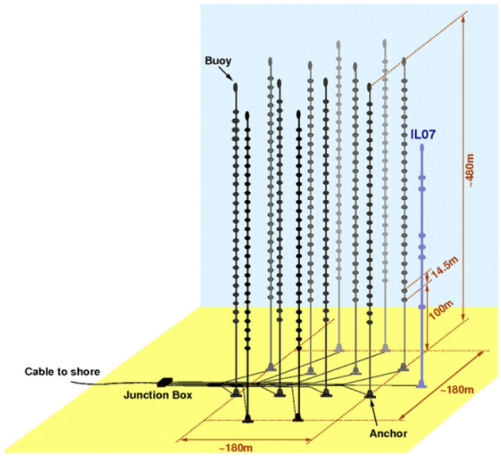


Figure 1.3:
Schematic view of the ANTARES detector. The lines are aranged in an octagonal configuration on the seabed.
Taken from [5].

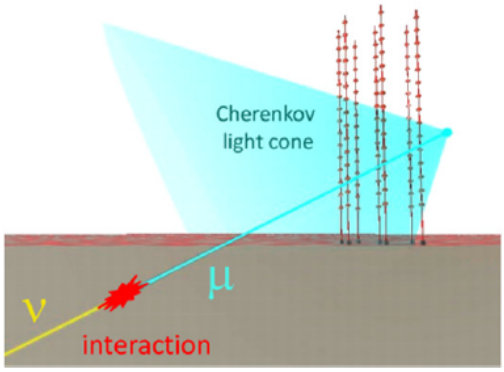


Figure 1.4:
Detection principle of ANTARES. A Neutrinos interacts with the medium around the detector and creates a muon which produces Cherenkov light, that can be detected by light sensors.
Taken from [5].

The goal of ANTARES as neutrino telescope is to detect astrophysical neutrinos. It is therefore optimized in its layout to detect upwards going neutrinos, using the earth as additional shielding to remove atmospheric muon contamination. The background of the PMT counting rate is dominated by the radioactive decay of ^{40}K and marine bioluminescence. ^{40}K is a radioactive potassium isotope contained within the salt of the sea water. With a half life of $1.2 \cdot 10^9$ years two of its decay channels emit Cherenkov

radiation contribution to the background rate:

$$^{40}\mathrm{K} \rightarrow {}^{40}\mathrm{Ca} + e + \bar{\nu}_e$$
$$^{40}\mathrm{K} \rightarrow {}^{40}\mathrm{Ar}^{*} + (\bar{\nu}_e) \rightarrow {}^{40}\mathrm{Ar} + \gamma$$

The contribution of ^{40}K to the background is constant over time. Exceedings of this constant baseline are therefore ascribed to bioluminescent activity from bacteria present in the deep sea [11].

2 Analysis Description

The central idea of this analysis is to apply the Fast Fourier Transformation onto the ANTARES PMT counting rates to search for possible periodic low-energy neutrino signals. In the following chapter, the developed analysis is described and advantages and challenges explained. The first section reviews the utilized data set and the Fourier power as suitable test statistic. The second section addresses issues which arise for the computation of the FFT and the obtained Fourier spectrum and presents suited solutions. In the third and final section further improvements to increase the sensitivity of this analysis are proposed.

2.1 Ingredients and expected Challenges of this Analysis

The fundamental aspects of this analysis can be summed up in the following three points:

- Properties of the used ANTARES data set
- Definition of a test statistic
- Behaviour of the test statistic for only background and background with signal

They are presented below to provide the basis for the analysis.

2.1.1 ANTARES PMT Rates

The search for a periodic neutrino source in this analysis is not performed on the reconstructed neutrino events, which are usually used in neutrino astronomy analyses, but instead directly on the pure PMT counting rates, of the optical modules.

These PMT rates are stored as i3-files and for this analysis first transformed into tabular csv-files and thereafter for faster access into HDF5-files. Each of these HDF5-files contains exactly one data taking run, and is therefore similarly indexed by the run number. Each rate value in such a HDF5-File is identified by four 'coordinates'. The first one is the time coordinate, which is expressed as number of frames since start of the data taking run. Together with the unix starting time of the run, and the sampling time, i.e. the time between two frames, $t_{\text{Sampling}} = 0.104858\,\text{s}$, the unix time of each data point can be inferred. The second coordinate is the LCM-ID which specifies the OMF and therefore the node within the detector grid. The third one is the ARS-ID. As there are two ARS per OM and three OM per OMF, it ranges from 0 to 5. The LCM-ID and the ARS-ID therefore serve as location coordinates of each data point. The last one, specifies the type of the rate. 'rateOff' denotes the offshore rate, which is sampled directly at the

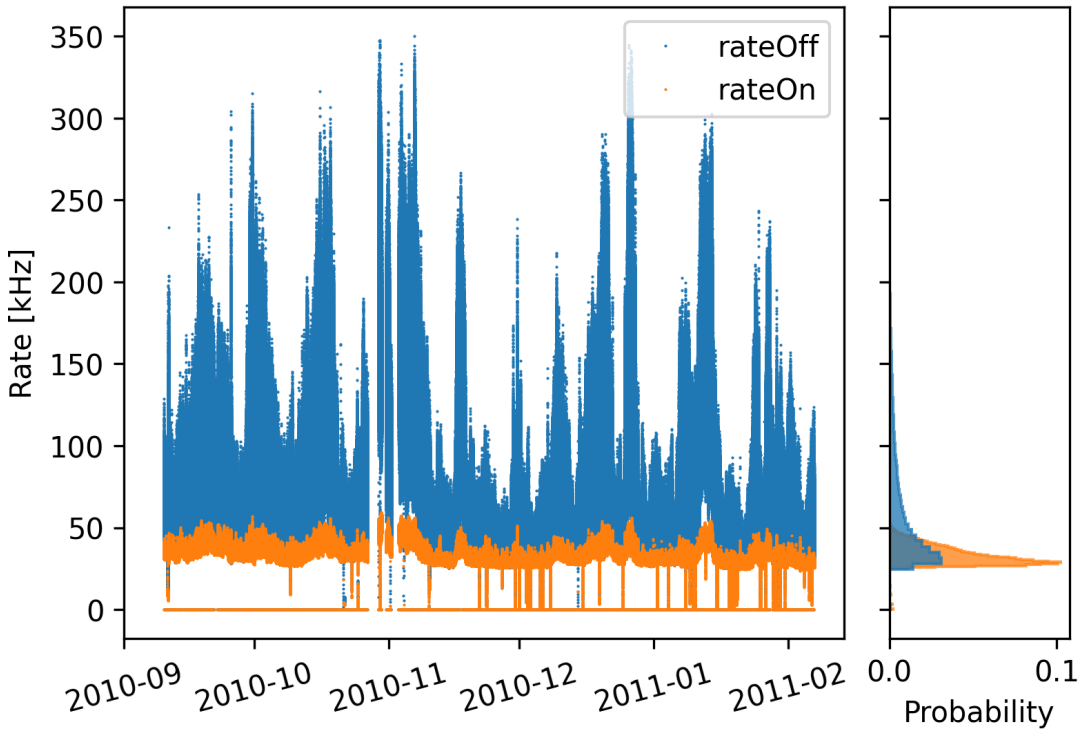


Figure 2.1:
The normalized rate values in the chosen period over the days since 2010-09-10. The right plot shows via a histogram the distribution of the rates. It can be seen, that the rateOff extend to much higher values, than the rateOn.

anode of the PMT as an 8-bit (count) rate value up to several MHz. 'rateOn' denotes the onshore rate and is calculated after the data acquisition system of ANTARES dealt for truncation and saturation effects. This reduces the possible value range, allowing for a better sampling resolution using 8-bits. For one OM, the two ARS will output therefore the same rateOff value, but different rateOn values. The rate of an OM for a fixed moment in time is given by the average of its two ARS rates.

Further information stored in the header of this file, is the unix starting time of the run, the number of samples of the run, the run number, the sampling time, as well as a list of all available LCM-IDs and the number of active ARS.

In this analysis the whole ANTARES telescope is effectively considered as one single photodetector. The rates over all LCMs and ARS are summed up for each moment in time (but separately for rateOff and rateOn), removing the two location coordinates from the data set. To consider for malfunctioning OMs or yet missing OMs due to the (in the past) still ongoing construction process, the average rate per OM is calculated by dividing by the number of active ARS.

So far, this rate data has not been used in any ANTARES analysis. The data set which is used in the later part, consists of 1531 ANTARES data taking runs from 2010-09-10 to 2011-02-06, it can be seen in Figure 2.1. This period was chosen, due to its relatively low baseline rate. The set contains $109359135 \approx 2^{26.7}$ sample points. For the full ANTARES lifetime of approximately 14 years, the number of sample points in the time series is about $2^{32} \approx 10^{9.6}$.

As this analysis makes direct usage of the PMT rates and no distinct neutrino events are considered, instead only deviations from the background rate caused by all collective neutrino responses for any given moment in time are investigated. An advantage of this approach is that also neutrinos below the detectors energy threshold for the reconstruction process contribute to the PMT rates and are therefore accessible with this analysis. This therefore yields a new way of detecting low energy neutrinos via their periodic response in the counting rates. One downside of putting the reconstruction process aside, is however the loss of any information about the arrival direction of the incoming neutrinos. Hence if a significant deviation from the background is identified, the origin of the corresponding neutrino signal can not be pinpointed based on the rates. Moreover, as neutrino telescopes can not be pointed in a direction of the sky, as photon based telescopes can, possible signals from all directions will be simultaneously present in the detector and overlap in the rates.

2.1.2 Properties of the FFT and Definition of a Test Statistic

To identify periodic signals in a time series, an often used method is the Fast Fourier Transform (FFT) algorithm. It converts a signal from its (usually) time domain into the frequency domain. This allows to identify periodic signals hidden by noise in the time domain much more easily, as they tend to clearly stand out above background in the frequency domain. A simple example of this ability is demonstrated in Figure 2.2. The FFT is an commonly used tool in physics and engineering, well studied and understood in literature.

First the properties of the FFT and the underlying DFT (Discrete Fourier Transform) are described, followed by the definition of a suitable test statistic for this analysis.

2.1.2.1 Introduction to the Discrete and Fast Fourier Transform

For a uniformly spaced time series z_n of length N, i.e. $n \in [0, N-1]$, the kth element of the discrete Fourier transform is defined as

$$\hat{z}_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} z_n \cdot \mathrm{e}^{-\mathrm{i}2\pi kn/N} \tag{2.1}$$

with $k \in [0, N-1]$ and $\hat{z}_k$ often called Fourier coefficient. [15] By this, a sequence in time domain is transformed to a sequence in frequency domain of same length. This therefore yields a way to detect periodicities in a time series, by analysing the Fourier transformed.

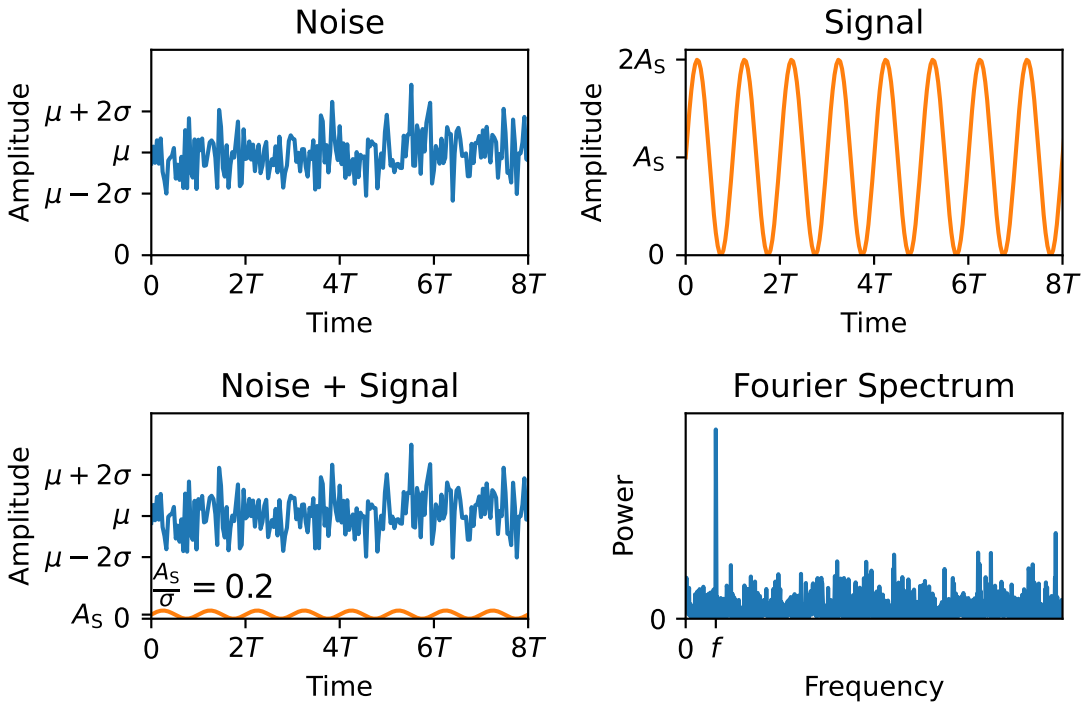


Figure 2.2:
Using the FFT, periodicities unrecognizable in time domain, can often be easily detected in frequency domain.
The top left shows a time series of pure Gaussian noise, with a mean μ and standard deviation σ. The top right shows a time series of a sinusoidal signal with frequency f, i.e. period $T = 1/f$, and amplitude A_S. The bottom left shows the superposition of the noise and signal. By bare eyes this is however indistinguishable to the above pure noise. The bottom right shows the Fourier power spectrum of the 'noise + signal' time series. The peak at frequency f clearly stands out above the background. It has to be noted, that the simulated time-series is 10 times longer than displayed here.

The time spacing dt between two successive elements of the time series determines the frequencies f_k associated with a wave number k by

$$f_k = \frac{k}{Ndt} = \frac{k}{T} \tag{2.2}$$

with T the total time duration of the time series. The frequency spacing is therefore given by $df = 1/T$ and is therefore also the frequency resolution of the DFT. The frequency $f_{\mathrm{Ny}} = N/(2T) = 1/(2dt)$ is the so called Nyquist frequency.

The sampling theorem states, that any signal can be exactly reconstructed from a series of uniformly spaced samples with spacing dt, if the signal contains no frequency components above the Nyquist frequency. If the DFT is applied to the samples of a signal with frequencies f above the Nyquist frequency, then this frequency will be projected into the available spectrum at $f_a = 2f_{\mathrm{Ny}} - f$ (for $f_{\mathrm{Ny}} < f < 2f_{\mathrm{Ny}}$). This is the so-called aliasing effect. [15] [21].

In general both, the input values z_n and the output values $\hat{z}_k$, can take complex values. In many applications however, the input time series is only real-valued. In this case, the Fourier coefficients are symmetric about the Nyquist frequency with $\hat{z}_{N-k} = \hat{z}_k^*$. It is therefore sufficient to calculate only half of the Fourier frequencies. For $k = 0$, i.e. 0 Hz, Equation 2.1 simply reduces to the sum of the time series values. As this contains no information about any periodicities, it is usually not displayed in the graphical representations of the Fourier coefficients. The plots of spectra therefore commonly show the interval $[df, f_{\mathrm{Ny}}]$, and contains due to this only half the number of points of the time series. [30]

The computational complexity of the DFT as defined in Equation 2.1 is given by $\mathcal{O}(N^2)$. The computation of any larger time series would therefore be practicable impossible. The Fast Fourier Transform is a implementation of the DFT, that reduces the complexity to $\mathcal{O}(N \log_2 N)$. This is done, by repeatedly splitting the time series into shorter sequences of even and odd indices and calculating the DFTs of each. The best performance can be achieved for time series with a length that is a power of 2 [15].

2.1.2.2 Definition of a Test Statistic

The complex valued Fourier coefficients $\hat{z}_k$ resulting from the DFT/FFT contain information about the strength of the corresponding signal in the form of the magnitude of $\hat{z}_k$. But also information about the phase of the underlying signal is available in the argument of the complex number.

As in many applications the phase has no important physical meaning and only takes an arbitrary value, due to the arbitrary start time of the data taking, it is often disregarded. Instead usually only the magnitude is used to assess the strength of a signal. Two common ways exist for this. First by directly using the absolute value, i.e. the magnitude, $|\hat{z}_k|$ which is called the Fourier amplitude. The second convention, is by using the absolute square $|\hat{z}_k|^2$, called the Fourier power. Both approaches are in principle equivalent, this analysis however makes use of the Fourier power, as in this case the background distribution takes the form of the well studied χ^2-distribution.

As a periodic signal in the time series has a strong response in the Fourier powers, distinguishable from background as can be seen in Figure 2.2, this quantity serves as a suitable test statistic for hypothesis testing and to decide whether a signal is present or not.

The p-value is an often used quantity in hypothesis testing. It corresponds to the probability of observing an event at least as extreme as the one observed, under the assumption of the so-called null hypothesis. If this probability is smaller than some predefined significance level, then the null hypothesis is rejected [31]. In this analysis, the null hypothesis corresponds to the absence of any periodic signal and the presence of only background. Hence, if the p-value of a Fourier power is smaller than the significance level α, then a periodic signal is detected.

2.1.3 Behaviour of the Fourier Power in the presence of noise and signal

To perform the hypothesis testing the distribution of the Fourier powers, under the assumption of the null hypothesis, i.e. only background, is required. The usual approach in particle physics to obtain an only-background scenario is to perform numerous advanced simulations of the detector, and deriving therefrom a distribution of the test statistic. Unfortunately, this is not possible for this analysis, as there are to many unknowns about the data taking conditions in ANTARES during each run.

However, the distribution of the Fourier powers for hypothesis testing can be calculated under fairly simple assumption. One of these, is the assumption of frequency independent white noise. In the reality of ANTARES this is not the case, as colored noise is present, causing an excess in the low frequency region. This will be taken care of later using the Red Noise Filter described in subsection 2.2.2, which transforms the obtained frequency dependent spectrum into a white spectrum.

In the following the derivation of the Fourier powers under the assumption of only background, but also signal plus background is described. In literature this can be found e.g. in [24], [30] and [32]. As they tend to be rather short and often omitting important steps, a full description is presented here.

Signal First assume a continuous sinusoidal signal of the form

$$S(t) = A_{\mathrm{S}} \cdot \sin(2\pi f t) \tag{2.3}$$

for any time t, with the signal amplitude A_{S} and the frequency f of the signal. However experimental data is usually not continuously stored, but instead discretized. The discrete signal can therefore be written in the form

$$S_n = A_{\mathrm{S}} \cdot \sin(2\pi f T n/N) \tag{2.4}$$

where the continuous time is replaced with $t = Tn/N$. Here, T denotes the total recorded time, N the total number of discrete points and $n \in [0, N-1]$ the index of a given point.

Noise For the noise of this model, some normal distributed noise with a probability density function (pdf) of the form[1]

$$f_{\mathrm{N}}(x) = \varphi(x|\mu_{\mathrm{N}}, \sigma_{\mathrm{N}}) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{N}}}} \cdot \exp\left(-\frac{(x-\mu_{\mathrm{N}})^2}{2\sigma_{\mathrm{N}}^2}\right) \tag{2.5}$$

is assumed, where x denotes the value of the noise. μ_{N} and σ_{N} are the mean and the standard deviation of the underlying normal distribution. Note that the standard deviation σ_{N} can be interpreted as the amplitude of the noise.

Noise + Signal Considering a time series of noise, distributed according to Equation 2.5, and injecting a signal by adding Equation 2.4, the pdf of the value with index n in the time series is given by

$$f_{\mathrm{SN}}(x, n) = \varphi(x|\mu_{\mathrm{N}} + S_n, \sigma_{\mathrm{N}}) \tag{2.6}$$

where the mean of the normal distribution is shifted by the momentary value of the signal.

Effect of the DFT The next step is to find the corresponding probability distributions of the complex valued Fourier coefficients.

For real valued input series, i.e. $z_n \in \mathbb{R}$, the DFT, as defined in Equation 2.1 can be separated into the real and imaginary part:

$$\hat{x}_k = \mathrm{Re}(\hat{z}_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} z_n \cdot \cos(-2\pi kn/N) \tag{2.7}$$

$$\hat{y}_k = \mathrm{Im}(\hat{z}_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} z_n \cdot \sin(-2\pi kn/N) \tag{2.8}$$

Change of Random Variable Consider a random variable X with a pdf given by $f_X(x)$. By multiplying X with some factor $a \in \mathbb{R}$, one can therefore create a new random variable $Y = a \cdot X$. The pdf of the new random variable is then given by

$$f_Y(y) = \frac{1}{a} \cdot f_X\left(\frac{y}{a}\right). \tag{2.9}$$

Therefore, as z_n is normal distributed according to Equation 2.6 the pdf of each summation term $x_{n,k} \coloneqq c_{n,k} \cdot z_n$ with $c_{n,k} \coloneqq \cos(-2\pi kn/N)/\sqrt{N}$ in Equation 2.7 and $y_{n,k} \coloneqq s_{n,k} \cdot z_n$ with $s_{n,k} \coloneqq \sin(-2\pi kn/N)/\sqrt{N}$ in Equation 2.8 can be written as

[1] To simplify the notation, we shall denote the pdf of a normal distribution with mean μ and standard deviation σ by $\varphi(x|\mu, \sigma)$.

$$f_{x_{n,k}}(x) = \frac{1}{c_{n,k}} \cdot f_{\mathrm{SN}}\left(\frac{1}{c_{n,k}} \cdot x\right) = \varphi\left(x | c_{n,k} \cdot (\mu_{\mathrm{N}} + S_n), c_{n,k} \cdot \sigma_{\mathrm{N}}\right) \tag{2.10}$$

$$f_{y_{n,k}}(y) = \frac{1}{s_{n,k}} \cdot f_{\mathrm{SN}}\left(\frac{1}{s_{n,k}} \cdot y\right) = \varphi\left(y | s_{n,k} \cdot (\mu_{\mathrm{N}} + S_n), s_{n,k} \cdot \sigma_{\mathrm{N}}\right). \tag{2.11}$$

This means, that each summation term $x_{n,k}$ is distributed like a normal distribution with shifted mean $\mu_{x_{n,k}} = c_{n,k} \cdot (\mu + S_n)$ and scaled standard deviation $\sigma_{x_{n,k}} = c_{n,k} \cdot \sigma_{\mathrm{N}}$, and respectively $y_{n,k}$ with mean $\mu_{y_{n,k}} = s_{n,k} \cdot (\mu + S_n)$ and standard deviation $\sigma^2_{y_{n,k}} = s_{n,k} \cdot \sigma_{\mathrm{N}}$.

Sum of Random Variables Both, real and imaginary part of a complex Fourier coefficient are therefore just the sum of N normal distributed random variables with varying mean and variance. The pdf of a random variable, that is constructed as a sum of N normal distributed random variables, is still a normal distribution. The mean of this new pdf is simply the sum of the means, and the variance similarly simply the sum of the variances [29].

A similar conclusion can be drawn from the central limit theorem (CLT), that states that the distribution of the sum of N random variables with finite mean and finite standard deviation converges towards a normal distribution for $N \to \infty$. These calculations therefor also hold for any non-normal distributed noise, that meets the requirements of the CLT, like for example uniformly distributed noise [29].

The means and variances for the pdfs of the real coefficient $\hat{x}_k$ and the imaginary coefficient $\hat{y}_k$ can therefore be easily obtained. For the full calculation see subsection 5.3.2.

Distribution of the Fourier Coefficients The real and imaginary part of the Fourier coefficients are again normally distributed. The pdf of the complex valued Fourier coefficient can therefore be interpreted as a 2-dimensional normal distribution in the complex plane. For frequency bins that contain only background, i.e. $fT - k \neq 0$ $\Rightarrow M_{\hat{y}_k} = 0$, the complex normal distribution is centered around the origin. If a sinusoidal signal is present, i.e. $fT - k = 0$, then the complex normal distribution is shifted downwards on the imaginary axis. Figure 2.3 illustrates these resulting distributions.

$$M_{\hat{x}_k} = \sum_{n=0}^{N-1} \mu_{x_{n,k}} = 0 \tag{2.12}$$

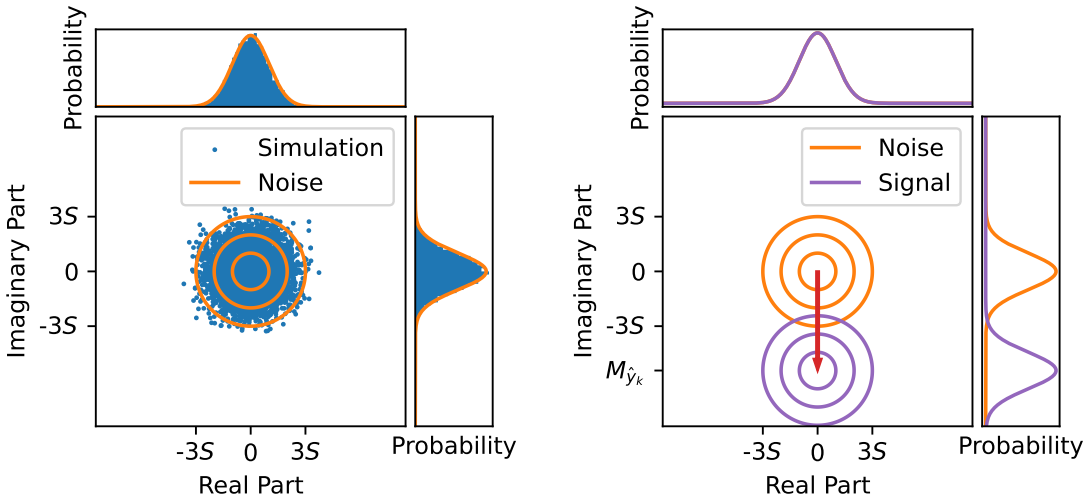
$$M_{\hat{y}_k} = \sum_{n=0}^{N-1} \mu_{y_{n,k}} = \begin{cases} -\frac{1}{2} \cdot A_\mathrm{S} \cdot \sqrt{N} & \text{if } [fT - k] = 0 \\ 0 & \text{if } [fT - k] \in [1, N-1] \end{cases} \tag{2.13}$$

$$S^2_{\hat{x}_k} = \sum_{n=0}^{N-1} \sigma^2_{x_{n,k}} = \frac{1}{2}\sigma^2_\mathrm{N} \tag{2.14}$$

$$S^2_{\hat{y}_k} = \sum_{n=0}^{N-1} \sigma^2_{y_{n,k}} = \frac{1}{2}\sigma^2_\mathrm{N} = S^2_{\hat{x}_k} = S^2 \tag{2.15}$$

$$\Rightarrow f_{\hat{x}_k}(x) = \varphi\left(x|0, S\right) \tag{2.16}$$

$$\Rightarrow f_{\hat{y}_k}(y) = \varphi\left(y|M_{\hat{y}_k}, S\right) \tag{2.17}$$



(a)
The FFT of some simulated white noise is calculated, and the resulting complex valued Fourier coefficients plotted on the complex plane. Due to the pure noise they are distributed according to a complex normal distribution centered around the origin.

(b)
If additional to the background a sinusoidal signal is present, the complex normal distribution of the one Fourier coefficient corresponding to the signals frequency is shifted downwards on the imaginary axis.

Figure 2.3:
The complex normal distributions of the complex valued Fourier coefficients. The contour circles indicate the regions 1, 2 and 3 standard deviations.

Distribution of the Fourier Spectrum The complex valued Fourier coefficients are usually displayed by calculating the absolute square and plotting this over the frequency to obtain a spectrum of Fourier powers. The absolute square is given by

$$|\hat{z}_k|^2 = \hat{x}_k^2 + \hat{y}_k^2. \tag{2.18}$$

Hence, $|\hat{z}_k|^2$ is just the sum of the square of two normal random variables.

In general, a random variable that is constructed as the sum of the square of ν normally distributed random variables with zero mean and unit variance is distributed according to a chi-squared distribution with ν degrees of freedom, also denoted as χ^2_ν [29]. Similarly, a random variable that is constructed as the sum of the square of ν normally distributed random variables with varying mean and unit variance is distributed according to a noncentral χ^2_ν distribution with ν degrees of freedom [28]. Therefore in this case a (noncentral) χ^2-distribution with 2 degrees of freedom is obtained. As the variance of $\hat{x}_k$ and $\hat{y}_k$ is in general $\neq 1$, they first need to be normalized, by dividing them by S. This introduces the additional scale parameter s in the (noncentral) χ^2-distribution, that takes the value $s = S^2$. For further information about the central and noncentral χ^2-distribution see section 5.1 and section 5.2.

For those frequency bins where $fT - k \neq 0$, i.e. no signal is contained, the normal distribution of both the real and imaginary part each have a mean of 0. Therefore the pdf of this pure noise is given by the 'central' χ^2_2-distribution with the pdf:

$$f_{\chi^2_2}(x|S^2) = \frac{1}{2S^2} \mathrm{e}^{-x/(2S^2)}. \tag{2.19}$$

For the frequency bin $fT - k = 0$ that contains a sinusoidal signal, the normal distribution of the imaginary part has a non-zero mean. Therefore the pdf of $|\hat{z}_k|^2$ is described by the more general noncentral χ^2_2-distribution. Its pdf is given by

$$f_{\mathrm{NC}\chi^2_2}(x|\lambda, S^2) = \frac{1}{2S^2} \mathrm{e}^{-(x/S^2+\lambda)/2} \cdot I_0\left(\sqrt{\lambda x/S^2}\right), \tag{2.20}$$

with λ the so-called noncentrality parameter defined as $\lambda = (M_{\hat{y}_k}/S)^2$ and the modified Bessel function I_0. Figure 2.4 displays the pdfs of the two kinds of χ^2-distributions.

Conclusion The χ^2-distribution describes the Fourier response of pure noise, and can therefore be used to model the background of the frequency spectrum. This allows for hypothesis testing, as the probability of a power value being larger or equal to some value x is given by the complementary cumulative distribution function (ccdf):

$$\bar{F}_{\chi^2_2}(x|S^2) = \mathrm{e}^{-x/(2S^2)} \tag{2.21}$$

The noncentral χ^2-distribution describes the frequency response of a sinusoidal signal plus noise. The expected frequency response of a signal is therefore given by the expected value of the noncentral χ^2-distribution:

$$\mathrm{E}_{\mathrm{NC}\chi^2_2} = (2 + \lambda)\, S^2 \tag{2.22}$$

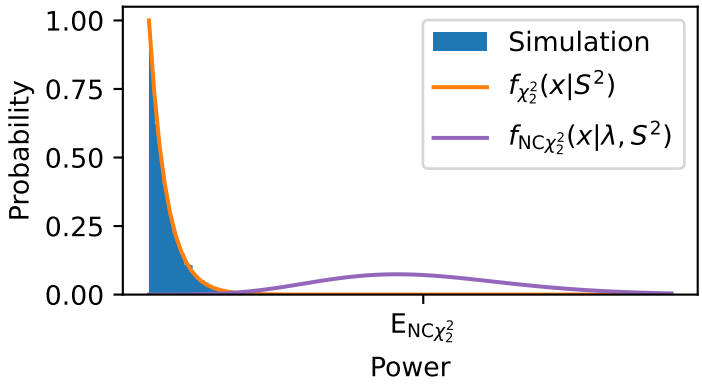


Figure 2.4:
Distribution of the Fourier powers of some simulated white noise and the probability density functions of the 'central' and noncentral χ^2-distribution. It can be seen, that the expected Fourier response of a (sufficiently strong) signal clearly deviates from only noise distribution.

This expected value of the Fourier response therefore allows to analytically calculate the expected outcome of further analysis steps and estimate their influence on the detection sensitivity.

Annotations

- To calculate the p-value of an expected Fourier response, it is in general sufficient to calculate
$$p = \bar{F}_{\chi^2_\nu}(\nu + \lambda) \tag{2.23}$$
as possible location and scale parameters present in $\mathrm{E}_{\mathrm{NC}\chi^2_\nu}$ cancel the location and scale parameters of the ccdf $\bar{F}_{\chi^2_\nu}$. ν denotes the degrees of freedom of the underlying χ^2-distribution and λ the noncentrality parameter of the signal.

- As the experimental data is usually some quantity derived counting numbers of events, negative values bear no physical meaning. Therefore one would need to add a constant offset of A_S to the signals in Equation 2.3 and Equation 2.4, such that the signal is always ≥ 0. This constant offset however is canceled out in the calculation of $M_{\hat{x}_k}$ and $M_{\hat{y}_k}$. See subsection 5.3.2. Similarly also for the description of the normal distributed noise, the parameters μ and σ need to be chosen such that no negative values appear.

- For an injected cosine wave, $M_{\hat{x}_k}$ will instead take the value $\frac{1}{2}A_\mathrm{S}\sqrt{N}$. If both a sine and a cosine wave of the same frequency are present, $M_{\hat{x}_k}$ and $M_{\hat{y}_k}$ will take the corresponding values, and both contribute to the noncentrality parameter λ. As any real valued signal can be written as a sum of sine and cosine waves, its Fourier series, the derivation of the noncentral χ^2 distribution holds for any periodic signal.

- If $[fT - k] \in \mathbb{R}/\mathbb{Z}$, i.e. is not an integer, then $M_{\hat{y}_k} > 0 \; \forall k$. This is the so-called scalloping effect, that appears if a signal frequency f doesn't exactly match with a discrete bin frequency f_k sampled by the DFT. If this happens, the power of the signal will be spread over the whole spectrum, as $P_k = P_0 \operatorname{sinc}^2 [\pi (k - fT)]$,[2] where P_0 is the corresponding Fourier power of the signal in the integer case, and P_k the contributing power to the kth bin. While the sinc function still keeps most of the power in the bin closest to the frequency, some power is lost to the neighbouring bins, and therefore some loss of sensitivity appears. In the worst caste, were the signal frequency lies exactly between two bin frequencies, nearly 60% signal power is lost, and on average $\approx 23\%$ are lost [24] [30].

- Another common way to display the Fourier coefficients is to calculate the absolute value instead of the absolute square. This will then result in the so-called (noncentral) chi distribution, that is closely related to the (noncentral) chi-squared distribution and shares many properties. As however the chi-squared distribution is usually better known, it is used here instead.

2.2 Additional Issues and Solutions

The description of the analysis so far made use of some assumptions, which are however not true in reality. This ranges from the actual spectrum containing non-white noise, as well as a discontinuous time series due to breaks in the data taking process. A further issue appears, if the analysis presented in this thesis is re-performed on the full ANTARES data set due to the sheer size of the data volume. Further descriptions and handy solutions to these issues will be given in the following section.

2.2.1 Averaging Spectra

The analysis performed at the end of this thesis consists of 2^{27} data points. This is short enough to compute the FFT within a reasonable time and the available memory. However the data set for the full lifetime of ANTARES will contain approximately 2^{32} data points. This exceeds the available computing resources for a simple application of the FFT algorithm.

The proposed approach to to perform the analysis with the full data set, is to split the full time series into N_{Avg} smaller equally sized time series of suitable size. One can then calculate the power spectra of all sub data sets and average them into one final spectrum. Setting the number of points for the small FFTs to N_0, the total number of points N is determined by $N = N_{\text{Avg}} \cdot N_0$.

The complexity of the FFT is given by $\mathcal{O}(N \log_2 N)$. For the averaging FFTs, the complexity can therefore be written as $\mathcal{O}([N_0 \log_2 N_0] \cdot N_{\text{Avg}}) = \mathcal{O}(N \log_2 N_0)$. Hence for a fixed N_0, the complexity only grows linearly with N, i.e. $\mathcal{O}(N)$, instead of $N \log_2 N$.

[2] The sinc function is here defined as $\operatorname{sinc}(x) = \sin(x)/x$.

On top of the increased computational effort, changes in the data taking conditions are another reason to split up the available data set. Certain properties like for example the background due to bio-luminescent activity or the PMT efficiencies are required to stay approximately constant. As however some of these change over time, it might be preferable to split the full data set into segments of approximately constant detector and environmental properties.

The averaging procedure changes the previously derived background statistic of a χ^2-distribution with 2 degrees of freedom, as each bin is not anymore the the sum of two normal distributed random variables but instead $2N_{\mathrm{Avg}}$, i.e.

$$|\hat{z}_k|^2 = \frac{1}{N_{\mathrm{Avg}}} \sum_{n=0}^{N_{\mathrm{Avg}}-1} \hat{x}_{k,i}^2 + \hat{y}_{k,i}^2. \tag{2.24}$$

This simply results in a χ^2-distribution with $\nu = 2N_{\mathrm{Avg}}$ degrees of freedom, i.e. a $\chi^2_{2N_{\mathrm{Avg}}}$-distribution. Also due to the averaging, the scaling parameter, i.e. the variance of the underlying normal distribution, changes and instead takes the value $S^2 = \frac{1}{2}\sigma_{\mathrm{N}}^2/N_{\mathrm{Avg}}$. The noncentrality parameter is then given by the sum over all means $M_{\hat{x}_{k,i}}$ and $M_{\hat{y}_{k,i}}$

$$\lambda = \sum_i^{N_{\mathrm{Avg}}} \frac{M_{\hat{x}_{k,i}}^2 + M_{\hat{y}_{k,i}}^2}{S^2} = \sum_i^{N_{\mathrm{Avg}}} \frac{\left(\frac{1}{2}A_{\mathrm{S}}\sqrt{N_0}\right)^2}{S^2} = \frac{1}{2}\left(\frac{A_{\mathrm{S}}}{\sigma_{\mathrm{N}}}\right)^2 N. \tag{2.25}$$

For the case of equal sized sub time series, this sum collapses again, to the same value, as the λ for the non averaging.

The complementary cumulative distribution function $\bar{F}$ and other quantities of interest for the case with averaging can be simply derived from the general forms of the χ^2-distribution in section 5.1 and noncentral χ^2-distribution in section 5.2.

This approach of the analysis by averaging, however introduces as a downside a loss of sensitivity due to the changed background statistics. A comparison of the p-value of the expecte Fourier response, as well as simulated p-values can be seen in Figure 2.5. There it can be observed, that the averaging process causes a drastic loss of sensitivity, but by increasing the total data set in size, the sensitivity can still be improved. Furthermore it can be noted, that the theoretical values describe the simulated values very well. It can therefore be concluded, that the expected values derived from the (noncentral) χ^2-distribution are a suitable way to analyse the behaviour of the FFT on signal and noise.

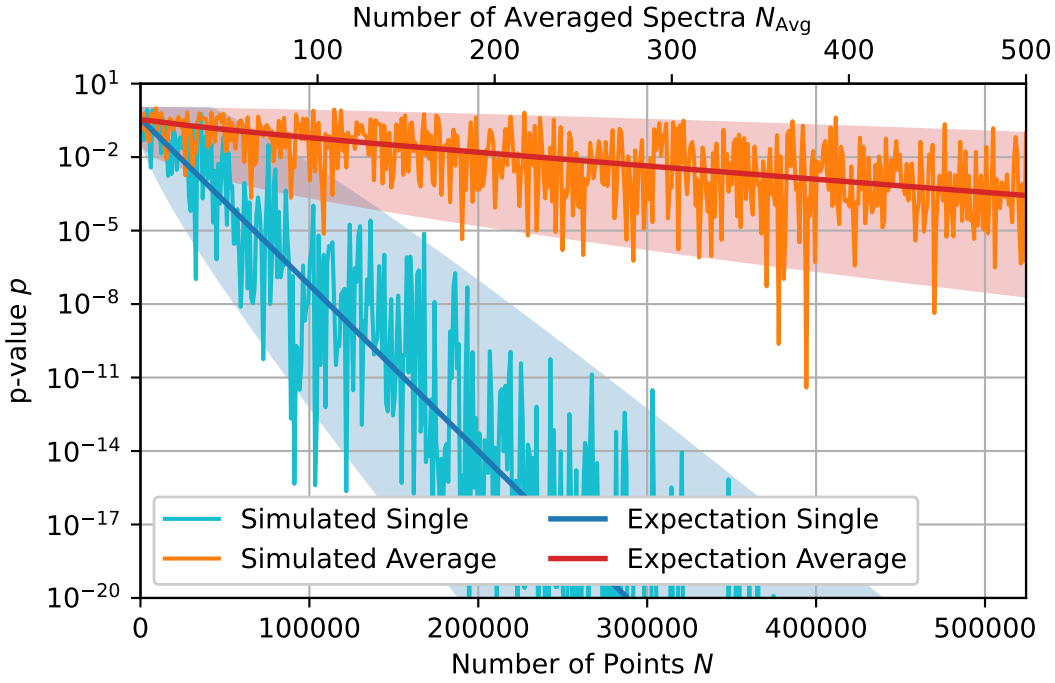


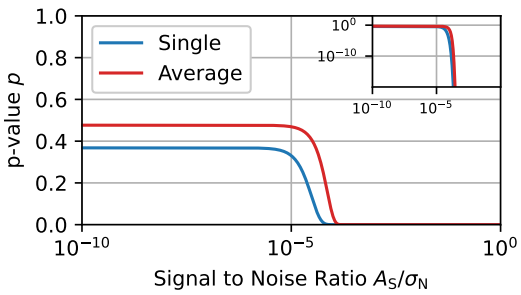
Figure 2.5:
Comparison of a single FFT to the averaged FFTs. A time series of N points is generated with Gaussian noise and a sinusoidal signal, with a signal to noise ratio of $2.5 \cdot 10^{-2}$. This time series is then split up into N_{Avg} smaller time series of size $N_0 = 2^{10}$, the power spectrum of each calculated and then these spectra averaged. Next the p-value of the power in the bin containing the signal frequency is calculated. The power spectrum and p-value are also obtained, for a FFT of the whole single time series. This process is repeated for increasing N and the resulting simulated p-values plotted. Furthermore, the expected values of the p-values are displayed, with the shaded areas being the region of ± 2 standard deviations around the expected value of the Fourier power. Increasing the total number of points N, increases according to this procedure for the averaged FFTs only the number of averaged spectra N_{Avg}, not the number of point of the underlying short time series.

The loss of sensitivity due to this averaging process can be estimated by calculating the the p-value of the expected Fourier response for the single case p_{S} and the averaged case p_{Avg}. The p-value is obtained using the complementary cumulative distribution function (ccdf) $\bar{F}_{\chi^2_\nu}$. The ratio $p_{\mathrm{S}}/p_{\mathrm{Avg}}$ serves then as a measure for the loss of sensitivity, as a value close to 1 indicates no significant loss, while a value close to 0 indicated a strong loss. This ratio can be written as

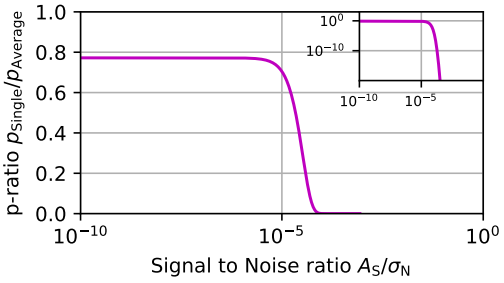
$$\frac{p_{\mathrm{S}}}{p_{\mathrm{Avg}}} = \frac{\bar{F}_{\chi^2_2}\left(2 + \frac{1}{2}\cdot\left(\frac{A_{\mathrm{S}}}{\sigma_{\mathrm{N}}}\right)^2 \cdot N\right)}{\bar{F}_{\chi^2_{2N_{\mathrm{Avg}}}}\left(2N_{\mathrm{Avg}} + \frac{1}{2}\cdot\left(\frac{A_{\mathrm{S}}}{\sigma_{\mathrm{N}}}\right)^2 \cdot N\right)} \tag{2.26}$$

and depends only on the signal to noise ration $A_{\mathrm{S}}/\sigma_{\mathrm{N}}$, the total number of points N and the number of averaged spectra N_{Avg}. (This equation can not be reasonably further simplified, as N_{Avg} appears within $\bar{F}_{\chi^2_{2N_{\mathrm{Avg}}}}$ in the upper limit of a sum, as well as in the terms of this sum, see section 5.1 Equation 5.5.)

Figure 2.6a shows the p-values corresponding to the expected Fourier response for the full ANTARES analysis as a function of the signal to noise ratio (SNR), with sub data sets of the size $N_0 = 2^{27}$. It can be seen, that up to a SNR of $\approx 10^{-5}$ the single FFT won't be able to differentiate a signal from background. For the averaged FFT, this detectability is slightly delayed. Hence there will be a SNR interval that is lost due to the averaging. In Figure 2.6b the ratio of the p-values (Equation 2.26) for the full ANTARES analysis is displayed. There it can be seen, that a significant loss of sensitivity occurs for SNR larger than $\approx 10^{-5}$. This coincides with the beginning of the detectability for the single FFT in Figure 2.6a. This indicates, that the averaging FFT will always have a significant loss of sensitivity, compared to the single FFT. This however tends to be in regions, where the p-value is already so low, that this loss of sensitivity wouldn't affect the claim of a discovery. The loss of sensitivity for lower SNRs is effectively irrelevant because both FFT approaches are unable to differentiate between signal and noise. It can be therefore concluded from Figure 2.6, that it is of great importance, to choose the size N_0 of the to-average FFTs as large as possible, to minimize the inevitable loss of sensitivity introduced by averaging.



(a)
p-value of the expected Fourier response in dependence of the signal to noise ratio for the single and averaged FFT. The plot in the right top, is the same plot in log-log-scale.



(b)
Ratio of the p-values of the expected Fourier response in dependence of the signal to noise ratio. The plot in the right top, is the same plot in log-log-scale.

Figure 2.6:
p-values and loss of sensitivity over the signal to noise ratio for the full ANTARES analysis $N \approx 2^{32}$, with $N_0 = 2^{27}$ the size of the to-average FFTs and $N_{\text{Avg}} = 22$ the number of averaged spectra.

2.2.2 Red Noise Filter

2.2.2.1 Description of the Red Noise Filter

The Red Noise Filter is one of the most important steps of this analysis, as it allows to transform the ANTARES background into a white spectrum, which in turn allows to estimate the sensitivity without the need to perform complicated simulations.

The assumptions for the noise in the previous derivation of the χ^2-distribution are not fulfilled in real experimental data. Actual experimental spectra are usually not only filled with white noise (no frequency dependence), but instead with some colored noise showing a frequency dependence. An example for ANTARES PMT rates can be seen in Figure 2.7. The frequency dependent behaviour can be clearly seen. In the plot with linear x-axis, a steep decline below $\approx 1\,\text{Hz}$ is visible. The double logarithmic plot reveals, that a change in the slope occurs at $\approx 0.1\,\text{Hz}$. This dominance of lower frequencies is sometimes called 'red noise'. Therefore as the spectrum is filled with non-white noise, the background can not be anymore simply described by a χ^2-distribution.

The standard approach followed in particle physics to search for signal within a background dominated dataset is to compare the observed data with the expectations in the case that an equivalent dataset was only composed by background. More specifically, a test statistic is derived from observable quantities, and its value is compared with the corresponding distribution for the background only case. Normally experimental data are very difficult to represent by means of a simple mathematical model, and such a distribution has to be derived by performing simulations. This is however rather difficult for this analysis as the detector is very complex and many necessary parameters of the

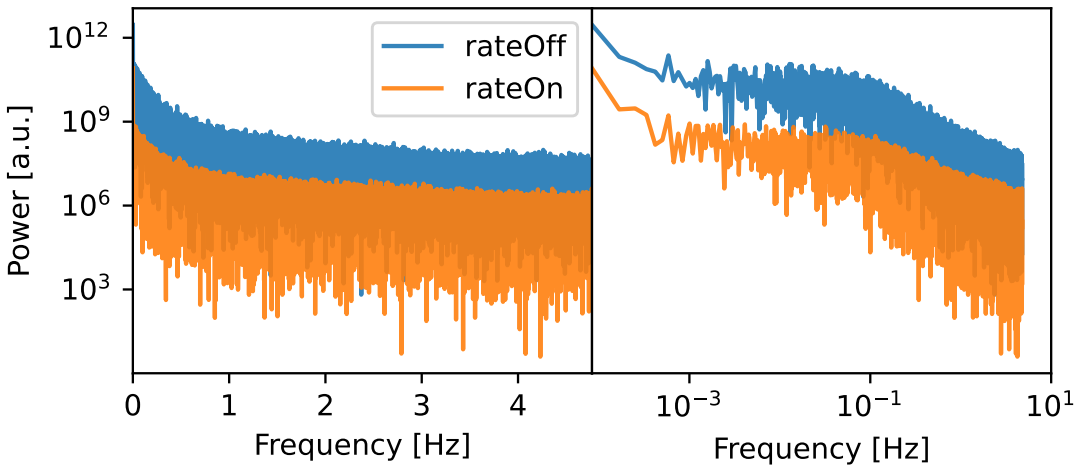


Figure 2.7:
Fourier power spectrum in log-lin scale and log-log scale of the ANTARES run 051870. Changing frequency dependencies can be seen over the spectrum.

data taking runs are not available.

The standard procedure in pulsar searches is instead a Red Noise Filter, also called Whitening Algorithm. It allows to transform the frequency dependent spectrum into a frequency independent white spectrum [24]. Applying this to the ANTARES data set would allow to perform the sensitivity study in an analytical way and overcome the obstacle of performing simulations.

The Red Noise Filter is applied after the FFT is performed. It works by splitting the frequency spectrum into smaller segments. For each segment, the empirical mean μ and empirical standard deviation σ are calculated, and then the segment normalized according to

$$X'_k = \frac{X_k - \mu}{\sigma} \tag{2.27}$$

with X_k and X'_k the value of the kth frequency bin before and after normalization. The normalized segments therefore have zero mean and unit standard deviation. If this is done properly for all segments, the complete spectrum loses (in good approximation) any frequency dependence and can be considered as a white spectrum.

These segments can either be all equal sized, or instead of variable size. The segments in low frequency regions usually need to rather small to obtain a white spectrum, due to the rather strong frequency dependence. Towards higher frequencies, larger segments in the already rather flat regions are preferable to avoid statistical fluctuations.

By construction, the whitened noise is not anymore described by a standard χ^2-distribution, but instead by an accordingly shifted and scaled version. The pdf can be

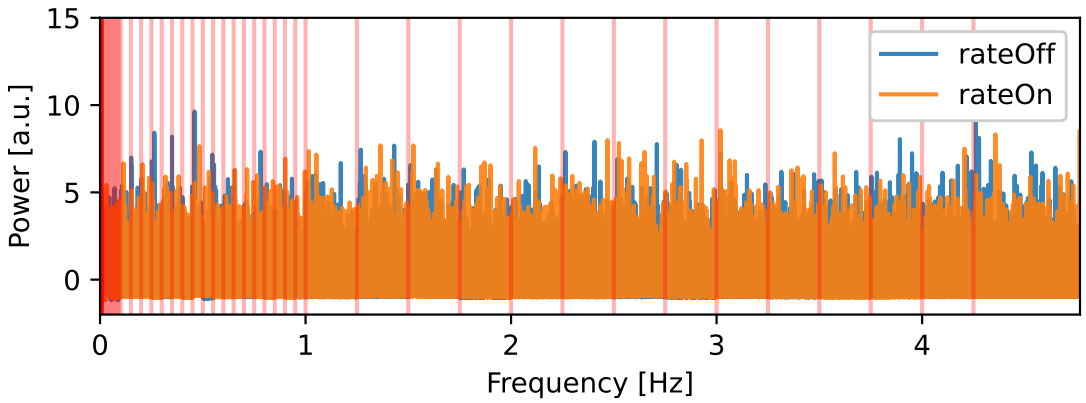


Figure 2.8:
Fourier power spectrum after the application of the Red Noise Filter of the ANTARES run 051870. The red lines indicate the borders of the segments. The frequency independence of the whole spectrum can be clearly seen.

given as

$$f_{RNF}(x, \nu, l_{\mathrm{RNF}}, s_{\mathrm{RNF}}) = f_{\chi^2_\nu}(x, l_{\mathrm{RNF}}, s_{\mathrm{RNF}}) \tag{2.28}$$

$$\text{with} \quad l_{\mathrm{RNF}} = -\sqrt{\frac{\nu}{2}} \quad \text{and} \quad s_{\mathrm{RNF}} = \frac{1}{\sqrt{2\nu}}. \tag{2.29}$$

$f_{\chi^2_\nu}$ denotes the pdf of the χ^2-distribution with ν degrees of freedom (see section 5.1). l_{RNF} and s_{RNF} are the location and the scale parameter of this new Red Noise Filter distribution and depend only on ν of the underlying χ^2-distribution. They can be directly derived from the mean and variance of the χ^2-distribution (see section 5.1), by setting these to to the normalized values (0 and 1) and solving for l and s. For the case without averaging, $\nu = 2$ and they take the values $l_{\mathrm{RNF}} = -1$ and $s_{\mathrm{RNF}} = 1/2$.

Figure 2.8 shows the successful applied Red Noise Filter on the previous spectrum displayed in Figure 2.7. The spectrum is now white and shows no frequency dependence.

2.2.2.2 Verification of a successful normalization

To check that the Red Noise Filter works as intended, it is applied individually to all available runs, with the run number ending by 0. The resulting distributions are then fitted in the location and scale parameter to a χ^2_2-distribution using the Method of Moments. The resulting fit parameters are in good accordance with the theoretical expected values, as can be seen in Figure 2.9.

Further in Figure 2.10 the combined histogram of all normalized segments of one run is displayed, together with an additional fit on this combined distribution, as well as the theoretical curve. Both curves describe the distribution very well. Some outliers are visible in the unsupported region $< l_{\mathrm{RNF}}$. These can occur if a segment has to many or large outliers disturbing the normalization process.

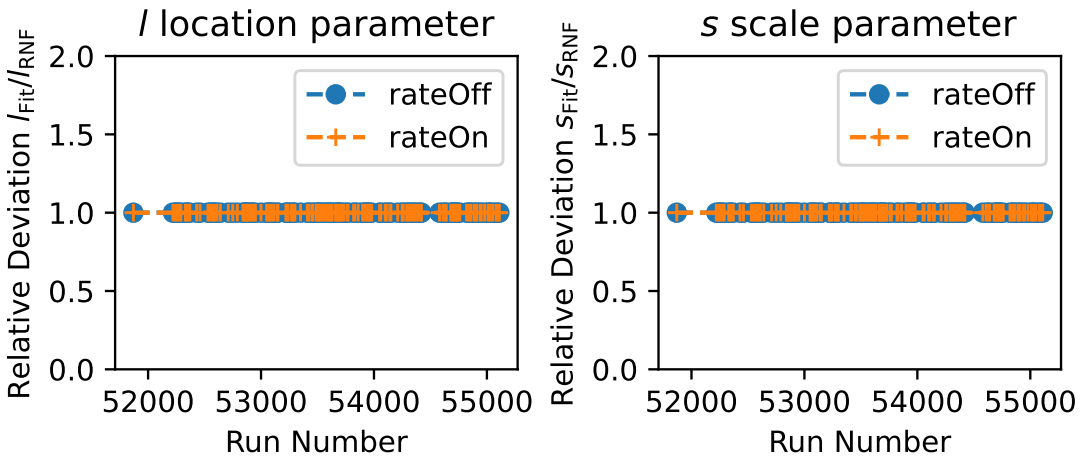


Figure 2.9:
The relative deviation of the fit parameters of the χ^2_2-distribution from the theoretical expectation over the run numbers. As this error is always close to one, it can be concluded, that the resulting spectra are in fact distributed like Equation 2.29.

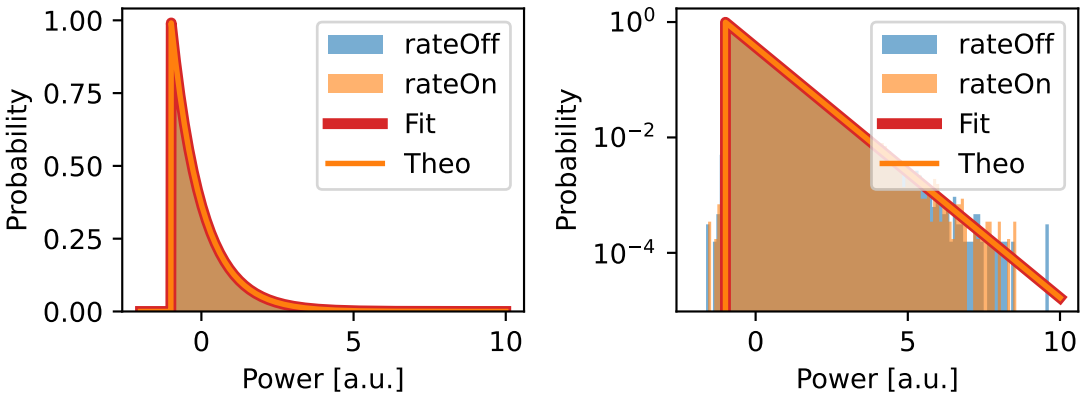


Figure 2.10:
Histogram of the Fourier powers after the application of the Red Noise Filter of the ANTARES run 051870. The left plot is in linear scale and the right plot in logarithmic. It can be clearly seen, that the theory of the shifted χ^2-distribution describes the whitened distribution very good, as well, as the fitted χ^2-distribution. Some powers below the theoretical limit l_{RNF} can be seen in the right plot. The arise if a segment has to large outliers.

As for this test each run was processed individually, the frequency spacing *df* becomes rather large and therefore the segments in the very low frequency regions only very sparsely populated. For the actual analysis all prepared runs are combined into one large time series. Therefore *df* is much smaller in the later performed analysis, improving the statistics of the segments.

Performing a goodness of fit test to verify the successful normalization turns out to be rather difficult. Attempts were done using the Kolmogorov–Smirnov test and the chi-squared test, but both failed to verify the χ^2-distribution of the whitened spectrum. Further a least square fit was executed with the resulting reduced chi-square being $\ll 1$. As a good fit yields a reduced chi-square close to 1, this also failed to quantify the goodness of the theoretical derived fit. The reason why these test failed, might be explained by the fact, that the underlying distribution of the whitened spectrum is not an actual χ^2-distribution, but only resembles it. Especially outliers in the unsupported regions $< l_{\mathrm{RNF}}$ seem to obstruct this testing. Despite these failures, this analysis idealized the normalized background to be χ^2-distributed, as it is assumed to still be a useful approximation to reality.

2.2.2.3 Decrease of sensitivity

Similar to the previously discusses outliers, a very strong signal has the potential to disturb the normalization process in a segment if the empirical mean μ and standard deviation σ differ to much from their ideal values.

To investigate the performance of the Red Noise Filter and a potential loss of sensitivity the noncentral χ^2-distribution can be used to model a signal and the 'central' χ^2-distribution to model the background. This allows to calculate the expected value of μ and σ in Equation 2.27 with a signal is present in the segment.

To do so, assume a segment containing N_{Seg} points, of which one, the signal, is distributed according to a noncentral χ^2_ν-distribution with noncentrality parameter λ and the rest according to a χ^2_ν-distribution. The location and scale parameters l and s take arbitrary values, describing the segment before its normalization.

The expected values are calculated to be (see section 5.4 for the full calculations):

$$\mathrm{E}[\mu] = \nu s + l + \frac{1}{N_{\mathrm{Seg}}}\lambda s \tag{2.30}$$

$$\mathrm{E}[\sigma^2] = 2\nu s^2 + \frac{1}{N_{\mathrm{Seg}}}4\lambda s^2 + \frac{N_{\mathrm{Seg}}-1}{N_{\mathrm{Seg}}^2}\lambda^2 s^2 \tag{2.31}$$

Putting these values into Equation 2.27 and setting X_i to the expected value of the noncentral χ^2-distribution, i.e. $X_i = \nu s + l + \lambda s$, one can calculate the expected real response of the Red Noise Filter for the assumed signal peak.

$$\frac{\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}[\mu]}{\sqrt{\mathrm{E}[\sigma^2]}} = \frac{(1 - 1/N_{\mathrm{Seg}})\lambda}{\sqrt{2\nu + 4\lambda/N_{\mathrm{Seg}} + 2\lambda^2(N_{\mathrm{Seg}}-1)/N_{\mathrm{Seg}}^2}} \tag{2.32}$$

A perfect normalization can be modeled by setting $\lambda = 0$ in $\mathrm{E}[\mu]$ and $\mathrm{E}[\sigma^2]$, describing the case, where the signal does not influence the normalization process. This simply replaces $\mathrm{E}[\mu]$ and $\mathrm{E}[\sigma^2]$ by the mean $\mathrm{E}_{\chi^2_\nu} = \nu s + l$ and variance $\mathrm{Var}_{\chi^2_\nu} = 2\nu s^2$ of the χ^2_ν-distribution. Putting these values in Equation 2.27 and again setting X_i to the expected value of the noncentral χ^2-distribution, the ideal expected response can be calculated.

$$\frac{\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}_{\chi^2_\nu}}{\sqrt{\mathrm{Var}_{\chi^2_\nu}}} = \frac{\lambda}{\sqrt{2\nu}} \tag{2.33}$$

This allows to calculate the p-value of the expected real and ideal Red Noise Filter response, using the ccdf of the underlying χ^2-distribution.

$$p_{\mathrm{RNF}} = \bar{F}_{\chi^2_\nu}\left(\frac{\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}[\mu]}{\sqrt{\mathrm{E}[\sigma^2]}} \;\middle|\; l_{\mathrm{RNF}}, s_{\mathrm{RNF}}\right) \tag{2.34}$$

$$p_{\mathrm{ideal}} = \bar{F}_{\chi^2_\nu}\left(\frac{\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}_{\chi^2_\nu}}{\sqrt{\mathrm{Var}_{\chi^2_\nu}}} \;\middle|\; l_{\mathrm{RNF}}, s_{\mathrm{RNF}}\right) \tag{2.35}$$

Figure 2.11 shows the resulting p-values over the signal to noise ration $A_\mathrm{S}/\sigma_\mathrm{N}$ (note $\lambda = (A_\mathrm{S}/\sigma_\mathrm{N})^2 \cdot N/2$ with N the length of the underlying time series), while Figure 2.12 displays the ration between the ideal and real case. It can be seen, that for strong signals, the real p-value will eventually diverge from the ideal p-value and converge towards some constant value as $\lim_{(A_\mathrm{S}/\sigma_\mathrm{N})\to\infty} p_{\mathrm{RNF}} = \bar{F}_{\chi^2_\nu}\left(\sqrt{(N_{\mathrm{Seg}} - 1)/2}\right)$. However for sufficiently large segments, this effect can be delayed towards acceptable small p-values. It can be concluded that, segments as large as possible are preferable to reduce the loss of sensitivity of potential signal.

Another implementation of the Red Noise Filter is possible, by replacing the mean in the normalization process with the median, as well as a 'standard deviation' that is calculated with respect to the median instead of the mean. The advantage of this is, that the empirical median is more stable with respect to outliers than the empirical mean. However for large outliers the 'standard deviation' will still take a to large value, and therefore disturb the normalization. One possible way of reducing the influence of strong signal outliers, for both mean and median implementation, could be to omit the maximum value of each segment in the calculation of the empirical normalization parameters. Keep in mind, that due to the scalloping effect, a signal can cause more than one outlier in a segment, (see annotations in subsection 2.1.3).

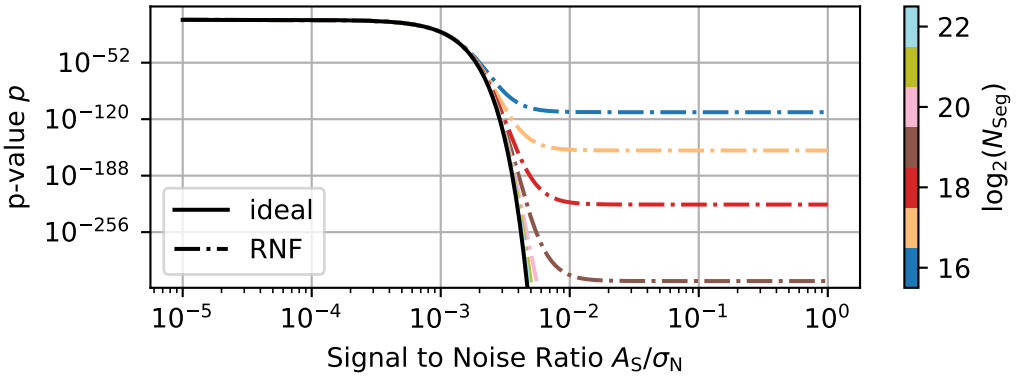


Figure 2.11:
p-value of the normalized signal peak over the signal to noise ratio, for varying segment sizes N_{Seg}. The length of the corresponding time series is set to 2^{27}. For sufficiently strong signals, the normalization process eventually fails. Increasing the segment size postpones this breakdown of the Red Noise Filter. The segment sizes used in the later analysis are $N_{\text{Seg}} = 2^{19}$ and $N_{\text{Seg}} = 2^{20}$.

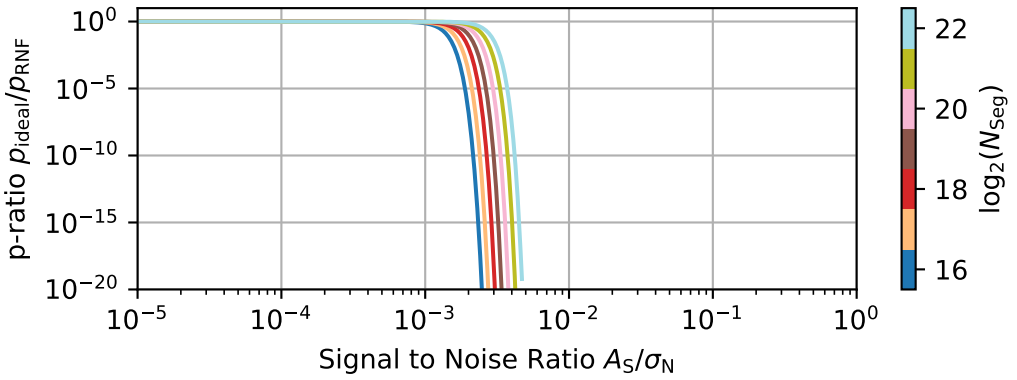


Figure 2.12:
Ratio of the ideal p-value to the real p-value over the signal to noise ratio, for varying segment sizes N_{Seg}. The length of the corresponding time series is set to 2^{27}. A ratio close to 1, indicates a successful normalization, while values much smaller 1 indicate a loss of sensitivity. Strong signals will cause a loss of sensitivity as the real p-value deviates from the ideal p-value. Increasing the segment size reduces this loss of sensitivity.

2.2.3 Data-Padding and Resampling

The definition of the Discrete Fourier Transform assumes a uniformly spaced input series. Experimental data is however not necessarily available in this uniformly spaced structure. This section will first discuss the issues of the time series structure of ANTARES and propose suitable solutions. In the following these solutions are further discussed and their implementation in this analysis presented.

The ANTARES data taking is performed in so-called runs of several hours, but with irregular recording time. Within a run, the data points are evenly spaced with $dt = t_{\text{Sampling}} = 0.104858\,\text{s}$. This allows to properly calculate the Fourier spectrum of each individual run. As an input series as long as possible is desirable to increase the sensitivity, this is not the implemented approach. Averaging the spectra as described in subsection 2.2.1 (taking the variable data set length into account) would be possible, however causes the average some undesired loss of sensitivity.

The performed approach is instead to connect multiple successive runs into one suitably long time series. However, the data taking runs of ANTARES do not connect seamlessly to each other. This (or also missing runs) causes gaps in the created time series, resulting in a non uniformly spaced dataset. By padding the data, i.e. introducing replacement values in these gaps, an evenly spaced time series can be constructed [24].

Regarding this, the implementation of the data taking of ANTARES presents one additional difficulty. Since the time difference between two consecutive runs is not necessarily an integer multiple of dt, data padding alone is therefore unable to create a perfect uniformly spaced input series series. It needs to be additionally resampled to obtain an evenly spaced time series.

Data-Padding Padding usually describes in FFT analyses the technique of appending a large number of artificial data point at the end of the datas set, to smoothen the spectrum. The two common procedures for this are the so called zero-padding and mean-padding. The first one appends data points of value zero, while the second one instead appends data points with value equivalent to the mean of the raw data set. Mean padding is in general preferable as zero-padding introduces low-frequency noise in the spectrum [30]. Figure 2.13 shows a simple example of this effect. Data-padding can also be used to fill gaps in the input data to obtain a evenly spaced time series [24].

Increasing the size of the data set by padding, does not increase the frequency resolution, despite the fact, that the padded spectrum has a narrower frequency spacing. Padding in the time domain merely interpolates the spectrum in the Fourier domain [16]. Moreover, as padding does not increase the power of the signal, the padded values need to be excluded from the sensitivity studies. Only the number of actual recorded data points needs to be considered [30].

Resampling The resampling procedure used in this analysis works as follows. First the start time of the first run and the end time of the last run are obtained and the number of uniform bins calculated by dividing their difference with the sampling time dt. A new uniform time series with N_{Bins} and time spacing dt is created. A bin in this empty time

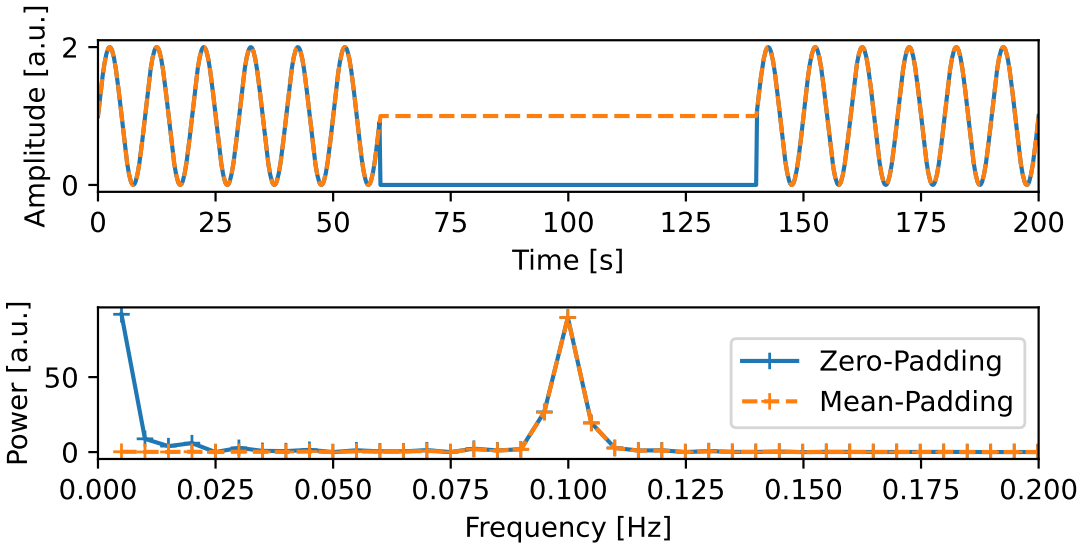


Figure 2.13:
A comparison of the effects zero-padding and mean-padding. It can be clearly seen, in the spectrum, that low frequency powers appear for zero padding.

series is then filled, with the value of the bin in the runs, that lies closest to it in time. In the next step, all remaining zero values are replaced by the mean of the input data. Therefore not only the padded values are set to the mean, but also the faulty zero-values present in the data, that can be seen in Figure 2.1. For the analysis performed in the later part of this thesis, the resampled time series was then mean-padded a second time to 2^{27} points.

2.3 Improving Sensitivity for more complex Scenarios

The described ingredients and tools so far are sufficient to perform an FFT analysis of the ANTARES PMT rates. Nevertheless the application of further techniques is desirable to increase the sensitivity for more complex scenarios. Two such procedures are presented in the following section, which deal with realistic signal shaped and additional correction of the time series.

2.3.1 Non sinusoidal Signals and Harmonic Summation

So far in the description of this analysis only pure sinusoidal signals are considered. This is however a rather unlikely model for many periodic sources. Instead they usually emit some pulse train, i.e. a periodic reappearance of some pulse. In the following the properties of non-sinusoidal signals are discussed, and a commonly used pulse model presented. Afterwards the Harmonic Summation is investigated, a tool to increase the

sensitivity of pulsed signals.

2.3.1.1 Non sinusoidal Signals

To describe non-sinusoidal periodic signals, many mathematical tools can be utilized. The following description of these is based on [15]. The continuous function of a pulse train with frequency f can be written as a sum of the underlying pulse shifted in time:

$$f_{\text{PulseTrain}}(t) = \sum_{n=-\infty}^{\infty} f_{\text{Pulse}}\left(t - \frac{n}{f}\right) \tag{2.36}$$

$$= \text{ш}_{1/f}(t) * f_{\text{Pulse}}(t) \tag{2.37}$$

It is often a handy way to write pulse trains by using the convolution between two functions $(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,\mathrm{d}\tau$ and the Dirac comb $\text{ш}_T(t) := \sum_{n=-\infty}^{\infty} \delta(t - nT)$ with period T.

Using this notation is rather simple to calculate the Fourier transform of such a pulse train. Let $\mathcal{F}(f)(y) = \int_{\mathbb{R}} f(x) \cdot \mathrm{e}^{-\mathrm{i}2\pi y \cdot x}\,\mathrm{d}x$ denote the continuous Fourier transformation, and the convolution theorem stating that $\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$.

$$\mathcal{F}(f_{\text{PulseTrain}}) = \mathcal{F}\left(\text{ш}_{1/f} * f_{\text{Pulse}}\right) \tag{2.38}$$

$$= \mathcal{F}\left(\text{ш}_{1/f}\right) \cdot \mathcal{F}(f_{\text{Pulse}}) \tag{2.39}$$

$$= \text{ш}_f \cdot \mathcal{F}(f_{\text{Pulse}}) \tag{2.40}$$

Note that the Fourier transform of the Dirac comb is also a Dirac comb, $\mathcal{F}(\text{ш}_T) = \text{ш}_{1/T}$. Therefore the Fourier transform of a periodic non-sinusoidal signal is a series of δ-peaks with spacing f, enveloped by the Fourier transform of the underlying pulse. In a spectrum, the frequency f is often called fundamental frequency, and its integer multiple the harmonics of higher order.

An often used pulse shape for estimation purposes, is the rectangular function $\text{rect}(t)$ [3], as it is the simplest pulse model, being either on or off. Its Fourier Transform is the sinc function $\text{sinc}(t)$. For a pulse train of rectangular pulses with frequency f, the power spectrum will contain harmonic peaks at integer multiples of f, following a sinc^2.

The term duty cycle denotes for pulse trains, the faction of a period in which a pulse is present. For a binary pulse train based on rect this is a straight forward definition. For other pulse shapes, such as a continuous Gaussian pulse, the width of the pulse needs to be defined separately. A common approach is the full width at half maximum (FWHM) [30].

Another important property of the Fourier transform is time scaling, sometimes also called similarity theorem. Scaling a function f in time according to $f\left(\frac{t}{a}\right)$, its Fourier

[3] $\text{rect}(t) = \begin{cases} 0, & \text{if } |t| > \frac{1}{2} \\ 1, & \text{if } |t| < \frac{1}{2} \\ \frac{1}{2}, & \text{if } |t| = \frac{1}{2} \end{cases}$ Varying definitions exist for the case $|t| = \frac{1}{2}$.

transform is $|a| \cdot \mathcal{F}(f)(a \cdot y)$. From this follows, that compressing a pulse in time, i.e. $0 < a < 1$, the envelop of the harmonics is stretched, and vice versa [15].

Due to this, pulse trains with narrow pulses compared to their pulse period, i.e. a low duty cycle, have a larger number of harmonics with relevant amplitude. The Harmonic Summing is a procedure that incorporate these harmonics of higher order into the candidate identification process of periodic signals [30].

2.3.1.2 Pulse Model: modified von Mises distribution

A frequently used pulse models in pulsar analyses is based on a modified von Mises distribution (MVMD): [30]

$$f_{\mathrm{MVMD}}(t, \kappa) = a \frac{\mathrm{e}^{\kappa \cos(2\pi f t)} - \mathrm{e}^{-\kappa}}{I_0(\kappa) - \mathrm{e}^{-\kappa}} \tag{2.41}$$

with the continuous time t, the frequency f and the shape parameter κ. I_0 denotes the Bessel function of zeroth order. κ determines the width of the pulses. For $\kappa \to 0$, it converges towards a sinusoid, and for $\kappa \to \infty$ the pulses converges towards a Gaussian pulse, where $1/\kappa$ corresponds to σ^2, the variance of the underlying normal distribution. The maximum value is

$$\max_{\mathrm{MVMD}} = \frac{2a \sinh(\kappa)}{I_0(\kappa) - \mathrm{e}^{-\kappa}} \tag{2.42}$$

and the full width at half-maximum (FWHM) as a fraction of a pulse is [4]

$$\mathrm{FWHM}_{\mathrm{MVMD}} = \pi^{-1} \arccos\Big(\kappa^{-1} \ln\big(\sinh(\kappa) + \mathrm{e}^{-\kappa}\big)\Big). \tag{2.43}$$

As this FWHM is already normalized with respect to the pulse period, Equation 2.43 describes the duty cycle of this pulse train.

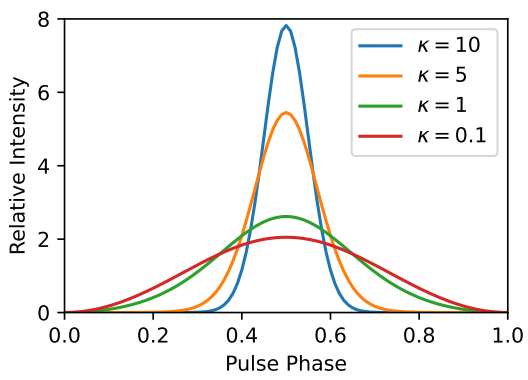
The advantage of the pulse profile in Equation 2.41 is, that it can be easily given in another form

$$f_{\mathrm{MVMD}}(t, \kappa) = a + \frac{2a \sum_{h=1}^{\infty} I_h(\kappa) \cos(2\pi h f t)}{I_0(\kappa) - \mathrm{e}^{-\kappa}} \tag{2.44}$$

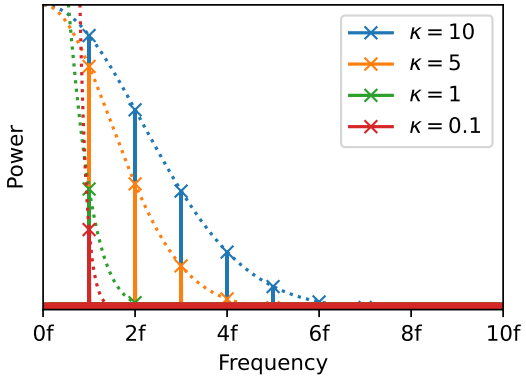
using the relation $\mathrm{e}^{x \cos(\theta)} = I_0(x) + 2 \sum_{h=1}^{\infty} I_h(x) \cos(h\theta)$, with I_h, the modified Bessel function of order h. Hence f_{MVMD} can be represented as a sum cosine waves, its Fourier series.

This makes it easy to calculate its Fourier response. In subsection 2.1.3 the noncentral χ^2-distribution is derived under the assumption of a single sinusoid. Its frequency only determines the index of the output bin in the spectra, but not the power distribution itself. This therefore allows now to consider each cosinusoid independently. Each response is therefore again a noncentral χ^2-distribution, only the signal amplitude A_{S} of each cosinusoid depends on the index h as $A_{\mathrm{S}}(h, \kappa) = \frac{2aI_h(\kappa)}{I_0(\kappa) - \mathrm{e}^{-\kappa}}$.

[4] Note, that the values given in [30] for Equation 2.42 and Equation 2.43 are incorrect.



(a)
Pulse profiles based on the MVMD. Increasing κ decreases the relative width (duty cycle) of the pulse.



(b)
Fourier power spectrum of a MVMD pulse train. For narrow peaks the number of significant harmonics of higher order increases.

Figure 2.14:
Properties of a pulse train based on the modified von Mises distribution as described in Equation 2.41 for different shape parameters κ.

2.3.1.3 Harmonic Summation

As stated earlier, the number of relevant harmonic peaks increases for narrow pules. It is therefore of great interest, to take them into consideration for the identification of a potential periodic signal.

The way this is usually done in pulsar search analysis is the so-called Harmonic Summation [24]. It is applied on the white spectrum after the Red Noise Filter and is given by the equation

$$X_{\mathrm{HS}}(k) = \sum_{h=1}^{H} x(hk) \tag{2.45}$$

where H is the total number of harmonics summed, x a value of the discrete power spectrum and k the frequency bin index. This sum is usually calculated multiple times for varying H as the number of relevant harmonics depends on the duty cycle of the signal [4]. Note that the output spectrum of the Harmonic Sum is truncated, to the interval $(0, \lfloor f_{\mathrm{Ny}}/H \rfloor]$, due to the fact that for higher frequencies not all corresponding harmonics lie within the available spectrum below the Nyquist frequency f_{Ny}.

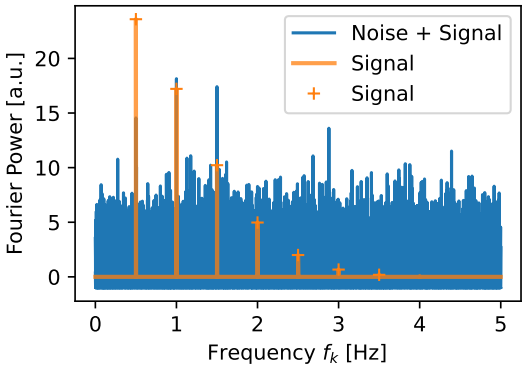
This summation process also changes the distribution of the test statistic under the background-only hypthesis. It is not anymore described by a χ^2-distribution with ν degrees of freedom, but instead a χ^2-distribution with $\nu \cdot H$ degrees of freedom. This is simply due to the additive property of the χ^2-distribution, that a sum of χ^2-distributed random numbers is still χ^2-distributed, where the new degree of freedom is the sum of the contributing degrees of freedom. Note that also the location parameter of the background distribution changes. Its new value is the sum of all location parameters of

the contributing distributions, hence $l_{\mathrm{HS}} = H \cdot l_{\mathrm{RNF}}$, and for the standard case of $\nu = 2$ therefore $l_{\mathrm{HS}} = -H$. The scale parameter keeps unchanged under the Harmonic Sum, i.e. $s_{\mathrm{HS}} = s_{\mathrm{RNF}}$. After the application of the Red Noise Filter and the Harmonic Summing, the pdf of the background can be written as

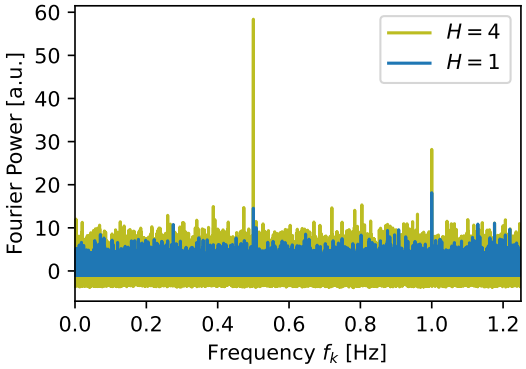
$$f_{\mathrm{HS}}(x, \nu H, l_{\mathrm{HS}}, s_{\mathrm{HS}}) = f_{\chi^2_{\nu H}}(x, l_{\mathrm{HS}}, s_{\mathrm{HS}}) \tag{2.46}$$

$$\text{with} \quad l_{\mathrm{HS}} = -H\sqrt{\frac{\nu}{2}} \quad \text{and} \quad s_{\mathrm{HS}} = \frac{1}{\sqrt{2\nu}}. \tag{2.47}$$

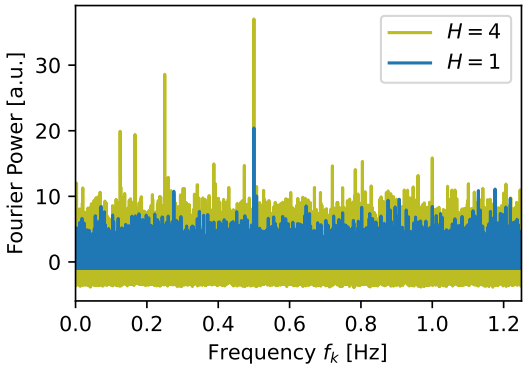
Note, that if the signal frequency f doesn't exactly match a bin frequency f_k, i.e. $f \cdot h \neq f_k \cdot h$, a drift in the harmonics can occur. Peaks of higher order, then lie next to their supposed position and therefore do not contribute to the calculation of the harmonic sum according to Equation 2.45. More elaborated algorithms exist, that try to capture this problem [4].



(a)
The Red Noise Filtered power spectrum of a MVMD-PulseTrain with white noise (blue) and the Fourier response of the pure MVMD-PulseTrain (not RNF-normalized). The signal peak at the fundamental and also the first harmonic can be seen, despite the presence of noise. The remaining harmonics, despite still of significant size, vanish in the noise, and their power therefore lost.



(b)
Spectrum of the Harmonic Sum for $H = 4$ and $H = 1$ (corresponds to not applying Harmonic Sum). It can be clearly seen, that due to the summing, the power at the signal frequency significantly increased, however also the background also increased.



(c)
If the frequency of a signal lies within a frequency bin with even index k, then false peaks can appear in the Harmonic Sum spectrum, as this frequency bin is then contained in the Harmonic Sum of lower frequencies.

Figure 2.15:
Some white Gaussian noise with variance 1 was created and a MVMD-PulseTrain injected with frequency $f \approx 0.5\,\text{Hz}$, $\kappa = 10$, $a = 4.5 \cdot 10^{-3}$. For Figure 2.15a and Figure 2.15b, the frequency was chosen to have an odd frequency index k, for Figure 2.15c the frequency was slightly changed to have an even frequency index k.

Loss and Gain of Sensitivity To investigate the potential impact on the sensitivity of the Harmonic Sum, the noncentral χ^2-distribution can be used to model the signal response. This allows to calculate the expected value of the Harmonic Sum X_{HS}.

First assume a time series containing N points, filled with normally distributed noise with standard deviation σ_{N} and a MVMD signal with amplitude a and shape parameter κ. The MVMD signal can be easily written as a series of cosinusoidal signal with amplitude $A_{\mathrm{S}}(h, \kappa)$. The expected Fourier response (after the application of an ideal Red Noise Filter) of the h'th harmonic order peak is given by the mean of the noncentral χ^2-distribution as

$$\mathrm{E}_{\mathrm{NC}\chi^2_\nu}(h) = \nu \cdot s_{\mathrm{RNF}} + \lambda(h) \cdot s_{\mathrm{RNF}} + l_{\mathrm{RNF}} \tag{2.48}$$

with

$$\lambda(h) = \left(\frac{A_{\mathrm{S}}(h,\kappa)}{\sigma_{\mathrm{N}}}\right)^2 \cdot \frac{N}{2} \tag{2.49}$$

$$= \left(\frac{a}{\sigma_{\mathrm{N}}}\right)^2 \cdot \left(\frac{I_h(\kappa)}{I_0(\kappa) - \mathrm{e}^{-\kappa}}\right)^2 N. \tag{2.50}$$

By using the linearity of the expected value, the expectation of the Harmonic Sum X_{HS} up to order H is given by

$$\mathrm{E}\left[X_{\mathrm{HS}}\right](H) = \sum_{h=1}^{H} \mathrm{E}_{\mathrm{NC}\chi^2_\nu}(h) = \nu H s_{\mathrm{RNF}} + \sum_{h=1}^{H} \lambda(h) \cdot s_{\mathrm{RNF}} + H l_{\mathrm{RNF}}. \tag{2.51}$$

This allows th calculate the p-value of the expected Harmonic Sum, by using the ccdf of the underlying χ^2-distribution.

$$p_{\mathrm{HS}}(H) = \bar{F}_{\chi^2_{\nu H}}\left(\mathrm{E}\left[X_{\mathrm{HS}}\right](H) \mid l_{\mathrm{HS}}, s_{\mathrm{HS}}\right) \tag{2.52}$$

To quantify the loss and gain of sensitivity, the ratio of p-values $p_{\mathrm{HS}}(1)/p_{\mathrm{HS}}(H)$ can be used. $p_{\mathrm{HS}}(1)$ denotes the the p-value if no Harmonic Sum is applied, i.e. $H = 1$ in Equation 2.52, and $p_{\mathrm{HS}}(H)$ the p-value for the Harmonic Sum up to order H. Figure 2.16 shows this ratio in dependence of the duty cycle of the MVMD (see Equation 2.43) for varying highest orders H. Ratios smaller than 1 indicate a loss of sensitivity, while ratios larger than 1 indicate a gain of sensitivity. It can be seen, that towards shorter duty cycles, as more and more harmonics of relevant size arise in the spectrum, the Harmonic Sum for increasing highest order H increases the sensitivity. From this is can be concluded, that the harmonic sum for multiply H should be calculated and individually investigated, to obtain the maximum sensitivity for each duty cycle interval.

2.3.2 Barycentric Correction

For the FFT analysis of long time series, the rotation of the Earth and its motion around the Sun can not be neglected anymore. The relative motion between observer and target introduces a Doppler shift. This causes a continuous change of the signal frequency with

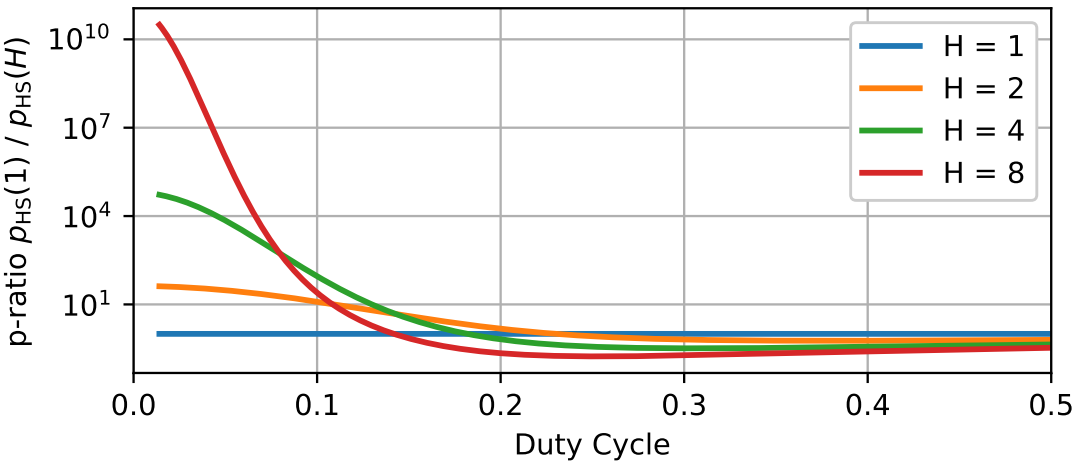


Figure 2.16:
Ratio of the p-value without summation to the p-value with Harmonic Sum over the duty cycle for varying highest harmonic order H. The length of the corresponding time series is set to $N = 10^{27}$ and the signal to noise ratio to $a/\sigma_N = 2 \cdot 10^{-4}$. The largest ratio of each interval indicates the best gain of sensitivity.

time. A changing frequency causes the peak in the Fourier spectrum to broaden over its neighbouring bins, causing a decrease of the sensitivity.

The established solution to this issue in pulsar analysis is the so called barycentric correction. It transforms the (topocentric) data of the telescope to the solar system barycenter, which is to a very good approximation an inertial reference frame. [24]

To perform this transformation the arrival time of each data point at the solar system barycenter needs to be calculated. As the relative motion is continually changing, the resulting time series is therefore not anymore uniformly spaced and needs to be resampled. The barycentric correction is therefore executed before the data-padding and resampling described in subsection 2.2.3.

This correction requires the observation time of each sample, and also the coordinates of the observing telescope itself, as well as the coordinates of the observation target.

The observer is approximated by the center of the ANTARES telescope with the coordinates $(42°48\text{N}, 6°10\text{E}, -2201\,\text{m})$. [5]

The coordinates of the observation target can be any coordinates on the sky, such as coordinates of already known possible sources, like pulsars.

In this analysis distinct neutrino events are not considered, and instead only deviations of the background rate, caused by all collective neutrino responses for a given moment in time, are investigated. Therefore no information about the arrival direction of the incident neutrinos is available in the PMT rates. Hence if a significant deviation from the background is identified, the origin of the corresponding neutrino signal can not be pinpointed based on the rates. Moreover, as neutrino telescopes can not be pointed in a

direction of the sky, as photon based telescopes can, possible signals from all directions will be simultaneously present in the detector and overlap in the rates.

The barycentric correction could therefore maybe used as a tool, to look at specific direction in the sky. This can be reasoned by the assumption, that only a signal from the corrected direction will be successfully restored and all possible signals from other directions sufficiently suppressed. However more investigation of this technique is required to estimate its influence and capabilities.

In this analysis, the calculation of the arrival time at the solar system barycenter is done using astropy's [12] 'Time' datatype and 'light_travel_time' method [1].

3 Analysis

In this chapter the previously described analysis will be performed on the data set introduced in subsection 2.1.1. The Harmonic Summing procedure and the barycentric correction depend on properties of the observation target. In the following first the selection criteria for the possible targets are described and the selected ones presented. Afterwards the sensitivity study is performed and a blinded analysis presented.

3.1 Selected Pulsars

Due to a general shortage of models describing the the neutrino emission pattern of pulsars, one can make the naive assumption that the neutrino emission properties, such as frequency and pulse width, are essentially identical to the catalogued photon emission properties.

Based on this assumption, the following selection criteria are applied to the ATNF catalogue:

- **Pulsar frequency within the spectrum:** The pulsars known frequency f_0 must be below the Nyquist frequency $f_{\text{Nyquist}} = 4.76834\,\text{Hz}$ of the ANTARES PMT rates. This is done to avoid aliasing effects of a potential signal peak to lower frequencies.
- **Exclude binary pulsars:** Binary pulsars are excluded from this analysis as their orbital motion causes an apparent change of their spin frequency over the course of the observation (similar to the motion of the earth and the barycentric correction). Their detection requires additional correction on the time series to take care of this continuous frequency change [30] [9].
- **Pulsars within a distance of** $5\,\text{kpc}$**:** According to the inverse-square law, for a fixed flux at the earth the required flux at the source grows with the distance squared. Therefore only pulsars close to earth, within $5\,\text{kpc}$ are take into consideration. Most of the selected are however within $2\,\text{kpc}$.
- **Suitably frequency and duty cycle combination for Harmonic Summation:** From the catalogued period and pulse width of a known pulsar, the duty cycle can be estimated. To test the Harmonic Summing procedure for narrow peaks up to the order H, the given pulsar frequency needs to be sufficiently low, such that the higher harmonics lie within the available spectrum.

The first two criteria are hard cuts of the analysis. They are based on the current capabilities of the used analysis tools. To lift these restriction the implementation of

further techniques is required which appropriately handle these cases. The later two criteria are soft cuts and have no intrinsic meaning. They are merely required to limit the wast number of pulsars to a suitably small number that can be processed in this test analysis.

Based on these criteria 5 pulsars are chosen to use in this exemplary analysis. They have a variety of different frequencies and duty cycles, allowing to perform the Harmonic Summation procedure for different parameters. Table 3.1 shows the final observation targets and their properties relevant for this analysis.

PSRJ	RAJD (deg)	DECJD (deg)	F0 (Hz)	DIST (kpc)	Duty Cycle
J1704-6016	256.06166667	-60.28166667	3.264528	1.589	0.2938075
J0820-4114	125.06441667	-41.24311111	1.833364	0.571	0.2585043
J0750-6846	117.64937500	-68.77605833	1.092638	0.338	0.2059624
J2325+6316	351.30549833	63.28121167	0.696229	4.855	0.09134521
J0828-3417	127.06927293	-34.28529102	0.540857	0.354	0.0611168

Table 3.1:
List of the selected Pulsars, that will be used in this analysis. Taken from the ATNF Pulsar catalog.
The first column gives the name of the pulsar based on J2000 coordinate. The abbreviation 'PSR' meaning 'pulsating source of radio emission'.
The second and third column give the Right ascension and Declination in J2000 coordinates. The next two columns contain the barycentric rotation frequency and the best estimates of the distance. The duty cycle in the last column is obtained as the quotient of the catalogued width of pulse at 50% of peak and the barycentric period of the pulsar.

3.2 Sensitivity Study

The previous chapter has described how to calculate the sensitivity of the FFT analysis for a data set composed of white noise and a signal of a given amplitude. This can now be used to estimate the discovery potential of this analysis applied to the ANTARES data. For this, the magnitude of the background in the used data set is required. Moreover the magnitude of a potential signal emitted by a pulsar needs to be evaluated.

The time series of the PMT rates and the distribution of the rates are shown in Figure 2.1. Table 3.2 lists important statistical properties of this sample distribution. The amplitude of the noise A_{N} is equivalent to the standard deviation σ_{N} of the distribution, and can be read off in Table 3.2.

The amplitude of the signal is determined by the response of the ANTARES detector to an incoming neutrino flux. In absence of models for neutrino emission in the energy range from $\approx 10\,\mathrm{MeV}$ to $\approx 10\,\mathrm{GeV}$ the response of ANTARES to a flux consisting of a mono-energetic neutrino beam is evaluated. The following section explains the calculation

	rateOff [kHz]	rateOn [kHz]
Mean	62.46	32.96
Median	47.31	31.76
Standard Deviation	38.89	5.62

Table 3.2:
Sample statistics of the PMT rates of the used data set.

of the detactability of ANTARES for a neutrino flux of a fixed energy.

The signal amplitude can be calculated by assuming a neutrino flux $F_\nu(E)$ that reaches the detector periodically and evaluating the detector response to this flux. This can be done by multiplying the incoming flux with the probability of an incoming neutrino to interact in the detector medium and by the probability that a neutrino interaction will produce a detection.This can be expressed mathematically by the following equation.

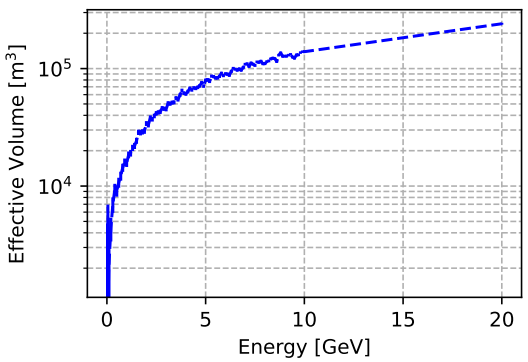
$$n_\nu(E) = F_\nu(E) \cdot \sigma(E) \cdot N_\mathrm{A} \cdot \rho \cdot V_\mathrm{eff}(E) \cdot \epsilon. \tag{3.1}$$

where F_ν is the incoming neutrino flux, σ the neutrino cross section, N_A is the Avogadro constant and ρ is the water density. The term V_eff is the so-called effective volume, and it is defined as the considered volume surrounding the detector multiplied by the probability that a neutrino interacting in this volume will produce hits on an ideal detector. The detector efficiency is taken into consideration as ϵ.

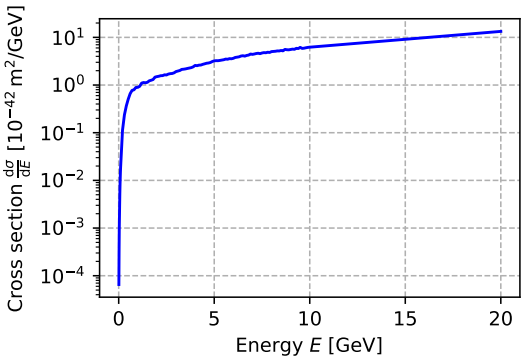
The calculation of the effective volume is based on simulations, where a certain number of neutrino interactions are generated in a spherical volume centered around an optical module. The final state particles are propagated through the water and the Cherenkov light emission is simulated. The generation volume is then multiplied by the ratio of events producing at least one hit on the PMT to the total number of generated events. Calculating the effective volume for ANTARES optical modules was due to software related issues not possible and remains to be done. For this analysis the ANTARES effective volume is estimated by using the KM3NeT effective volume. The main difference between the two optical modules is the photocathode area, with KM3NeT's being larger by a factor around 3.

Additional software issues didn't allow to produce the simulations for a hydrogen target and the effective volume of a KM3NeT optical module, shown in Figure 3.1a, corresponds to the electron neutrino charged current interactions with oxyen. Solving these issues is a task beyond the scope of this thesis and the subsequent calculations should be updated with the ANTARES effective volume once these issues are solved. Similarly to the effective volume, the interaction cross section σ is an energy dependent quantity. It is also obtained by extensive simulations. See Figure 3.1b for resulting values.

The PMT efficiency ϵ corresponds to the probability of an incident photon at the PMT to be detected. This is a property of the detector and changes over time. It is obtained by measuring the ^{40}K coincidence rate and is available for each OM in periods of ≈ 6 days. The combined detector efficiency ϵ used here, is obtained by averaging the efficiencies over all OMs. Figure 3.2 displays this detector efficiency over the selected



(a)
Effective Volume of the KM3NeT optical modules. The effective volume of ANTARES optical modules is estimated, by dividing this by the factor 3, the ratio of the photocathode area.



(b)
Charged current cross section of the neutrino with oxygen.

Figure 3.1:
The effective volume and interaction cross section as function of the neutrino energy.

period. A significant jump in the PMT efficiency is visible at the start of November 2010. This was possibly caused by a re-calibration of the detector. This jump discourages to perform a single FFT analysis on the entire available data set. This is because the required sensitivity estimations would be rather inaccurate by approximating the PMT efficiency by its total mean. Instead only the analysis for the second period with a PMT efficiency of about $\epsilon = 0.87$ is performed. This set starts at 2010-11-01 00:00:00 and ends at 2011-02-07 01:04:31.279 and contains $N = 72066798 \approx 2^{26.1}$ data points. The analysis of the averaged spectra of these two periods is desirable, however to include the varying PMT efficiency into the sensitivity estimations, more work needs to be done.

The signal amplitude A_S at which the sensitivity reaches 5σ and a discovery can be claimed needs to be estimated. The p-value is calculated as

$$p = \bar{F}_{\chi^2_2}\left(2 + \frac{1}{2}\left(\frac{A_\mathrm{S}}{\sigma_\mathrm{N}}\right)^2 \cdot N\right) \tag{3.2}$$

and is shown in Figure 3.3 for the ANTARES telescope using as noise amplitude σ_N the standard deviations given in Table 3.2. The Fourier spectrum will contain $N/2$ independent power values. This is a large enough number of points, such that even under the assumption of only background, a small number of power values will surpass the 5σ threshold just by chance. This is the so-called look-elsewhere effect. A simple approach to prevent erratic discoveries, is to multiply the p-value with the number of independent trials [14] or equivalently divide the significance levels by by the number of independent trials. The later one is used in this analysis.

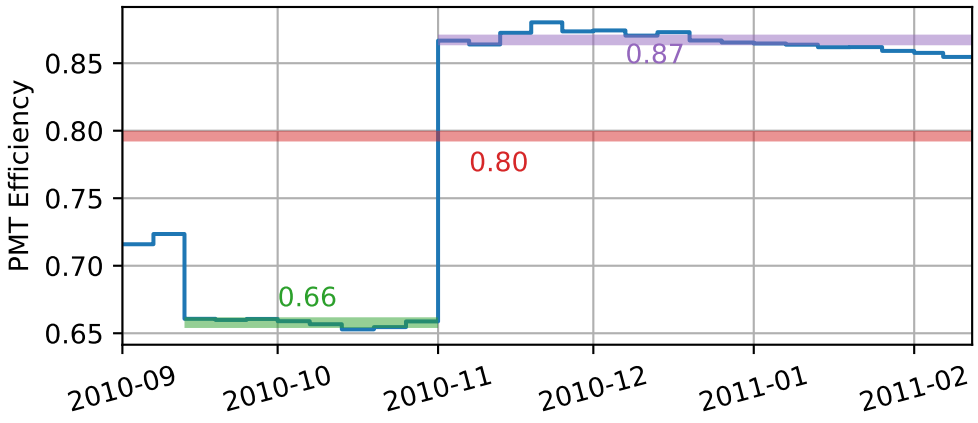


Figure 3.2:
ANTARES PMT efficiencies over time for the selected period. To sub periods can be identified. They are indicated by the colored horizontal line, at height of the mean of these sub periods. Additionally the total mean of this period is displayed. The colored number as the corresponding mean values.

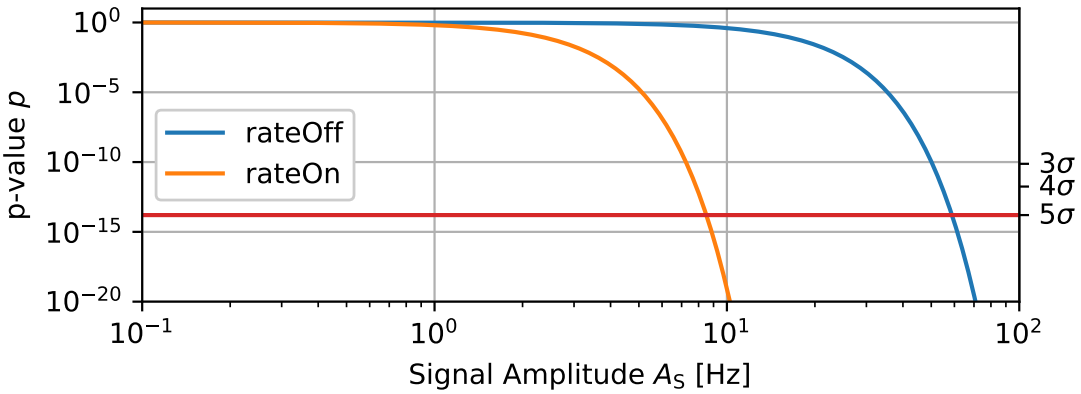


Figure 3.3:
The p-value for an incident periodic neutrino flux in ANTARES with signal amplitude A_S. The 5σ amplitude for rateOn is 8.53 Hz and for rateOff 59.1 Hz.

Figure 3.3 shows the discovery potential of ANTARES as a function of the signal amplitude for the data set equivalent to the one described in subsection 2.1.1. The signal amplitude can be related to physical parameters of the source of this signal. In particular, the signal amplitude depends on the incoming neutrino flux and on the energy of the interacting neutrinos. We can therefore show a more detailed figure to describe the ANTARES discovery potential as a function of these two parameters. This is shown in Figure 3.4 for the different sources described in Table 3.1. The color scale in these plots indicates the neutrino detection rate. Knowing the distance between the source and Earth, the flux can be extrapolated to the flux at the source by applying the inverse square law.

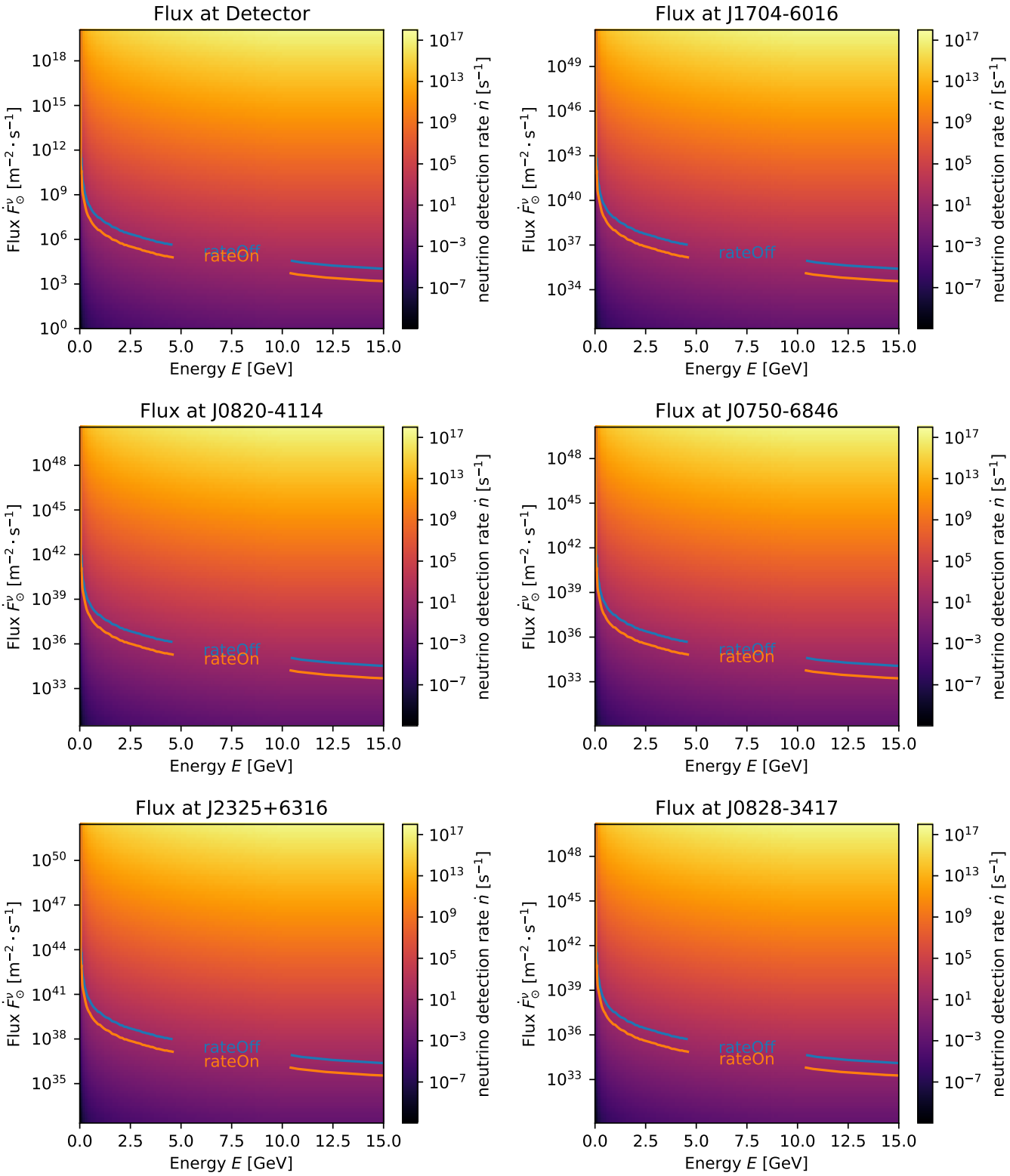


Figure 3.4:
Sensitivity plots for ANTARES of the used data set. A periodic flux located in a region above the contours should be detectable with a sensitivity of 5σ.

3.3 Blinded Analysis

After having calculated the discovery potential of the ANTARES telescope to the signal of interest, the next step in the analysis procedure is to apply the described analysis techniques to the ANTARES data set. Nevertheless, permission to look into the data should be granted by the ANTARES collaboration after a review of all the work presented above, and after solving the issues with the effective volume described in section 3.2. It is still valuable before analysing the data and looking at the results, to verify that the ANTARES data and the described procedure do not contain any undesired incompatibility problem. In order to do this, a blinding policy is followed where the real data are modified in a way that any possible signal like the one that the analysis is intended to find, is removed. In this case, the blinding is achieved by randomly shuffling the rate values in the time series. The result after applying the analysis to this data set should be a non detection and the power spectrum should be distributed as expected from background only.

The execution of the analysis starts with the selection of the data set. As described in the previous section, the period from 2010-11-01 00:00:00 to 2011-02-07 01:04:31.279 is selected. The first step is to perform the barycentric correction on the data points for the selected pulsar. The corrected time series is subsequently resampled and padded up to a length of 2^{27} points. The FFT is applied onto the rate values and the natural unnormalized spectrum obtained. The Red Noise Filter is applied in the next step. The spectrum is chosen to be partitioned into 16 segments of size 2^{19} in the low frequency region, and 56 segments of size 2^{20}. The resulting spectra are thereby sufficiently whitened. If possible for the candidate, the harmonic sum is performed for the highest harmonic orders $H = 2, 4, 8$. The resulting spectra are then plotted for further investigation. Additionally peaks surpassing a predefined significance level α are written into a text file for closer inspection.

Figure 3.5 shows the performed analysis on the blinded data set without barycentric correction. This corresponds to the uncorrected spectrum directly present in the ANTARES detector. Figure 3.6 shows exemplary the resulting spectra plot for one of the selected pulsar candidates. A red line indicates the known pulsar frequency. In both figures, the first line shows the natural Fourier spectrum directly calculated from the available rates. In an unblinded analysis, the colored noise in the low frequency regions would be visible and open to investigations. Figure 3.5 additionally shows in its second line the natural spectrum in a double logarithmic representation, to make possible power laws in this colored noise visible. The next line shows the normalized spectrum, obtained after the application of the Red Noise Filter. If in range, the right y-axis is marked with the σ significance levels. Additionally the tick $E = 1$ marks the area, above which the expected number of points, assuming only background, is 1. The experience with the unblinded test runs, with run number ending in 0, showed that the normalization process in the very first segment usually fails, as the power values tend to differ by several order of magnitude (as can be seen in Figure 2.7). Therefore in the following normalized plots only the spectrum starting with the second Red Noise Filter segment is shown. The

following lines show the spectra of the Harmonic Sum for increasing H. If the pulsar frequency is not anymore contained for a given H, then the resulting plot is omitted.

As the data is sufficiently blinded, no signal of significant strength is present. Moreover, the blinding process destroys the colored background and replaces it by an effectively white background.

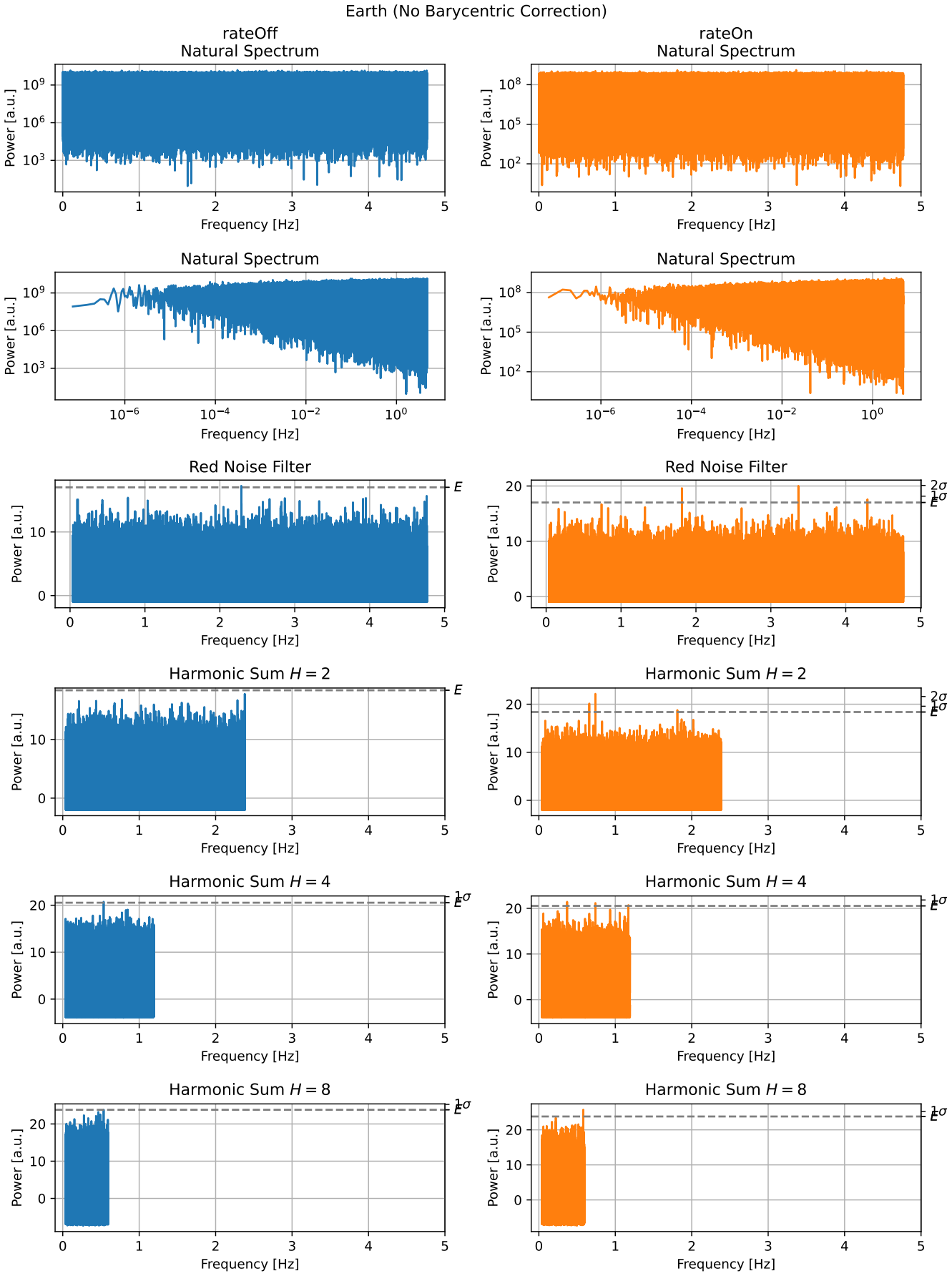


Figure 3.5:
Blinded analysis of the selected data set without applied barycentric correction.

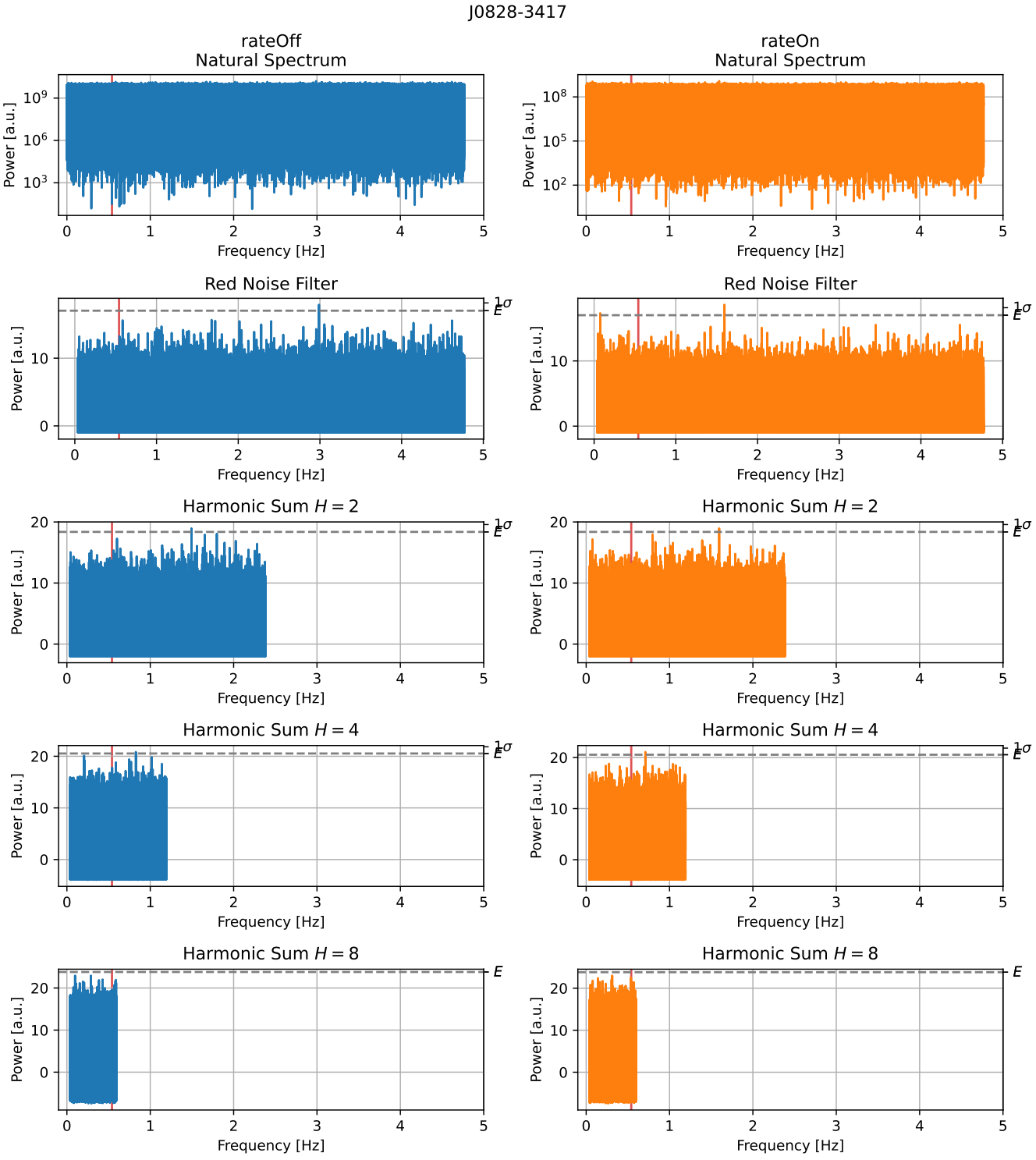


Figure 3.6:
Blinded analysis of the selected data set for the pulsar J0828-3417.

4 Summary and Outlook

In this thesis, an analysis was developed to search for periodic low-energy neutrinos sources with the ANTARES neutrino telescope. The investigation of periodicities directly in the PMT counting rates is performed by the application of the Fast Fourier Transformation.

Using an FFT, it is possible to perform an analytic calculation of the discovery potential of the ANTARES telescope to a periodic signal, which can be related to an incoming neutrino flux from a periodic source. This analysis presents a few challenges that have been covered in this work.

In particular, analysing a data set with arbitrary length, with discontinuities and with a non reproducibe background poses three challenges, which were tackled by methods like averaging, padding, resampling and by applying transformation algorithms as described in chapter 2.

Established techniques from pulsar astronomy were presented and their capabilities investigated. Properties of non-sinusoidal signals and a realistic pulse model for pulsars were presented. Using the Harmonic Summing increases the detection sensitivity of such signals. With the barycentric correction, the influence of the Earth's motion on a periodic signal can be counteracted, and possibly information about the direction of a signal acquired.

Possible observation targets for the assembled analysis were suggested and their detection sensitivity on the available test data set estimated. Finally, the implemented analysis was performed on the blinded test data set.

During the development of this work, some issues were encountered with the software required to calculate the ANTARES OM response to an incomming low energy neutrino flux. These issues remain to be solved and in the meantime an approximation based on the equivalent calculations for the KM3NeT DOM was used.

After the blinding policy of ANTARES for this proposed analysis is lifted, it can be performed on the suggested and additional candidates.

This analysis has the potential to be further improved and expanded, as numerous additional techniques were developed in the long history of pulsar surveys in radio astronomy.

Bibliography

[1] URL: https://docs.astropy.org/en/stable/api/astropy.time.Time.html#astropy.time.Time.light_travel_time.

[2] M. G. Aartsen et al. "IceCube Search for High-energy Neutrino Emission from TeV Pulsar Wind Nebulae". In: *The Astrophysical Journal* 898.2 (June 2020), p. 117. DOI: 10.3847/1538-4357/ab9fa0. URL: https://dx.doi.org/10.3847/1538-4357/ab9fa0.

[3] B. P. Abbott et al. "Multi-messenger Observations of a Binary Neutron Star Merger". In: *The Astrophysical Journal* 848.2 (Oct. 2017), p. L12. DOI: 10.3847/2041-8213/aa91c9. URL: https://doi.org/10.3847%2F2041-8213%2Faa91c9.

[4] K. Adámek, J. Roy, and W. Armour. "A novel greedy approach to harmonic summing using GPUs". In: *Astronomy and Computing* 40 (July 2022), p. 100621. DOI: 10.1016/j.ascom.2022.100621. URL: https://doi.org/10.1016%5C%2Fj.ascom.2022.100621.

[5] M. Ageron et al. "ANTARES: The first undersea neutrino telescope". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 656.1 (2011), pp. 11–38. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2011.06.103. URL: https://www.sciencedirect.com/science/article/pii/S0168900211013994.

[6] S. Aiello et al. "Implementation and first results of the KM3NeT real-time core-collapse supernova neutrino search". In: *The European Physical Journal C* 82.4 (Apr. 2022). DOI: 10.1140/epjc/s10052-022-10137-y. URL: https://doi.org/10.1140/epjc/s10052-022-10137-y.

[7] Hui Hui Dai Alan Jeffrey. *Handbook of mathematical formulas and integrals*. 4th ed. Academic Press, 2008. ISBN: 9780123742889. URL: https://libgen.fun/book/index.php?md5=737a2e57f816a1487c93ca9bfa598546.

[8] E. Amato, D. Guetta, and P. Blasi. "Signatures of high energy protons in pulsar winds". In: *Astronomy & Astrophysics* 402.3 (Apr. 2003), pp. 827–836. DOI: 10.1051/0004-6361:20030279. URL: https://doi.org/10.1051%5C%2F0004-6361%5C%3A20030279.

[9] Bridget C. Andersen and Scott M. Ransom. "A Fourier Domain "Jerk" Search for Binary Pulsars". In: *The Astrophysical Journal Letters* 863.1 (Aug. 2018), p. L13. DOI: 10.3847/2041-8213/aad59f. URL: https://dx.doi.org/10.3847/2041-8213/aad59f.

[10] *ANTARES - Astronomy with a Neutrino Telescope and Abyss environmental RESearch*. URL: https://antares.in2p3.fr/.

[11] ANTARES Collaboration and S. Escoffier. *The ANTARES detector: background sources and effects on detector performance*. 2007. DOI: 10.48550/ARXIV.0710.0527. URL: https://arxiv.org/abs/0710.0527.

[12] Astropy Collaboration et al. "The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package". In: *apj* 935.2, 167 (Aug. 2022), p. 167. DOI: 10.3847/1538-4357/ac7c74. arXiv: 2206.14220 [astro-ph.IM].

[13] *Baikal Gigaton Volume Detector*. URL: https://baikalgvd.jinr.ru/.

[14] Adrian E. Bayer and Uroš Seljak. "The look-elsewhere effect from a unified Bayesian and frequentist perspective". In: *Journal of Cosmology and Astroparticle Physics* 2020.10 (Oct. 2020), pp. 009–009. DOI: 10.1088/1475-7516/2020/10/009. URL: https://doi.org/10.1088%5C%2F1475-7516%5C%2F2020%5C%2F10%5C%2F009.

[15] Ronald N. Bracewell. *The Fourier Transform And Its Applications*. 3rd ed. McGraw-Hill series in electrical and computer engineering. Circuits and systems. McGraw Hill, 2000. ISBN: 0-07-303938-1.

[16] D. Donnelle and B. Rust. "The fast Fourier transform for experimentalists. Part I. Concepts". In: *Computing in Science & Engineering* 7.2 (2005), pp. 80–88. DOI: 10.1109/MCSE.2005.42.

[17] "Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector". In: *Science* 342.6161 (Nov. 2013). DOI: 10.1126/science.1242856. URL: https://doi.org/10.1126/science.1242856.

[18] Ke Fang et al. "IceCube constraints on fast-spinning pulsars as high-energy neutrino sources". In: *Journal of Cosmology and Astroparticle Physics* 2016.04 (Apr. 2016), p. 010. DOI: 10.1088/1475-7516/2016/04/010. URL: https://dx.doi.org/10.1088/1475-7516/2016/04/010.

[19] A.R. Hewish et al. "Observation of a Rapidly Pulsating Radio Source". In: *Nature* 217 (Feb. 1968). DOI: 10.1038/217709a0.

[20] *IceCube - Neutrino Observatory*. URL: https://icecube.wisc.edu/.

[21] Scott Ransom James Condon. *Essential Radio Astronomy*. Princeton Series in Modern Observational Astronomy. Princeton University Press, 2016. ISBN: 9780691137797.

[22] *KM3NeT - Opens a new window on our universe*. URL: https://www.km3net.org/.

[23] Don S. Lemons. *An introduction to stochastic processes in physics, containing On the theory of Brownian notion*. Johns Hopkins Paperback. The Johns Hopkins University Press, 2002. ISBN: 0801868661.

[24] D.R. Lorimer and M. Kramer. *Handbook of Pulsar Astronomy*. Cambridge University Press.

[25] R. N. Manchester et al. “The Australia Telescope National Facility Pulsar Catalogue”. In: *The Astronomical Journal* 129.4 (Apr. 2005), p. 1993. DOI: 10.1086/428488. URL: https://dx.doi.org/10.1086/428488.

[26] and Mark Aartsen et al. “Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A”. In: *Science* 361.6398 (July 2018). DOI: 10.1126/science.aat1378. URL: https://doi.org/10.1126%5C%2Fscience.aat1378.

[27] Alexander A Mushtukov et al. “Ultraluminous X-ray sources as neutrino pulsars”. In: *Monthly Notices of the Royal Astronomical Society* 476.3 (Feb. 2018), pp. 2867–2873. ISSN: 0035-8711. DOI: 10.1093/mnras/sty379. eprint: https://academic.oup.com/mnras/article-pdf/476/3/2867/24422267/sty379.pdf. URL: https://doi.org/10.1093/mnras/sty379.

[28] N. Balakrishnan Norman L. Johnson Samuel Kotz. *Continuous univariate distributions*. 2nd ed. Wiley Series in Probability and Statistics. Wiley-Interscience, 1995. ISBN: 9780471584940.

[29] N. Balakrishnan Norman L. Johnson Samuel Kotz. *Continuous Univariate Distributions, Vol. 1 (Wiley Series in Probability and Statistics)*. 2nd ed. 1994. ISBN: 9780471584957.

[30] Scott M. Ransom, Stephen S. Eikenberry, and John Middleditch. “Fourier Techniques for Very Long Astrophysical Time-Series Analysis”. In: *The Astronomical Journal* 124.3 (Sept. 2002), p. 1788. DOI: 10.1086/342285. URL: https://dx.doi.org/10.1086/342285.

[31] Rinaldo B. Schinazi. *Probability with statistical applications*. 2. ed. Boston [u.a.]: Birkhäuser, 2012. ISBN: 9780817682491.

[32] Meng Yu. “Harmonic Summing Improves Pulsar Detection Sensitivity: A Probability Analysis”. In: *The Astrophysical Journal* 868.1 (Nov. 2018), p. 8. DOI: 10.3847/1538-4357/aae51a. URL: https://dx.doi.org/10.3847/1538-4357/aae51a.

5 Appendix

5.1 chi-Squared Distribution

The chi-squared distribution, also χ^2-distribution, is a reoccurring probability distribution in this analysis and shall therefore here be described. [29]

Let $X_1, X_2, \ldots, X_\nu$ be ν independent and identically distributed random variables each following a normal distribution with zero mean and unit variance. Then $X^2 = \sum_{i=0}^{\nu} X_i^2$ is said to follow the χ^2-distribution with ν degrees of freedom (DOF), also denoted as χ^2_ν-distribution

The probability density function (pdf) is given by

$$f_{\chi^2_\nu}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \mathrm{e}^{-x/2} \tag{5.1}$$

for $x > 0$ with $\nu \in \mathbb{N}$ [1] and the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$.

This is the standardized form of the χ^2-distribution, as it appears in most text books. However, this analysis frequently requires a more general form, that incorporates the location parameter l and scale parameter s. The corresponding pdf can be written as

$$f_{\chi^2_\nu}(x|l,s) = \frac{1}{s} \cdot f_{\chi^2_\nu}(y) \qquad \text{with} \qquad y = \frac{x-l}{s}. \tag{5.2}$$

The location parameter considers a constant contribution in the above sum of random numbers. The scale parameter s corresponds to the variance of the underlying normal distribution. Hence for $l = 0$ and $s = 1$ the standardized form is restored.

The mean of this distribution is given by

$$\mathrm{E}_{\chi^2_\nu} = \nu s + l \tag{5.3}$$

and the variance by

$$\mathrm{Var}_{\chi^2_\nu} = 2\nu s^2. \tag{5.4}$$

The complementary cumulative distribution function (ccdf), also survival function, yields the probability of observing an event at least as extreme as the one observed. It can be written as

$$\bar{F}_{\chi^2_\nu}(x|l,s) = 1 - \frac{\gamma\left(\frac{\nu}{2}, \frac{y}{2}\right)}{\Gamma\left(\nu/2\right)} \underset{\text{even } \nu}{\overset{\text{for}}{=}} \mathrm{e}^{-y/2} \sum_{k=0}^{\nu/2-1} \frac{(y/2)^k}{k!} \tag{5.5}$$

[1] For the derivation of the distribution ν is usually considered to be an integer, hence the name 'degrees of freedom'. However, the distribution also holds for any positive real value of ν.

with the lower incomplete gamma function $\gamma(s,x) = \int_0^x t^{s-1}\mathrm{e}^{-t}\,\mathrm{d}t$. [2]

It follows from the definition of the χ^2-distribution, that the sum of χ^2-distributed random variables X_i with ν_i DOF, is still a χ^2-distributed random variable with $\sum_i \nu_i$ DOF. Furthermore it can be noted, that for large ν the χ^2-distribution converges towards a normal distribution.

5.2 Noncentral chi-Squared Distribution

The noncentral chi-squared distribution, also noncentral χ^2-distribution, is a generalization of the chi-squared distribution. [28]

Let $X_1, X_2, \ldots, X_\nu$ be ν independent distributed random variables, each following a normal distribution with mean μ_i and unit variance. Then $X^2 = \sum_{i=0}^{\nu} X_i^2$ is said to follow the noncentral χ^2-distribution with ν degrees of freedom and the noncentrality parameter λ, defined as

$$\lambda = \sum_{i=1}^{\nu} \mu_i^2. \tag{5.6}$$

The probability density function (pdf) is given by

$$f_{\mathrm{NC}\chi^2_\nu}(x|\lambda) = \frac{1}{2}\mathrm{e}^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{(\nu/2-1)/2} I_{(\nu/2-1)}\left(\sqrt{\lambda x}\right) \tag{5.7}$$

for $x > 0$ with $\lambda > 0$ and $\nu \in \mathbb{N}$.[3] With the modified Bessel function of the first kind $I_\alpha(x) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+\alpha+1)} \left(\frac{x}{2}\right)^{2k+\alpha}$.

This is the standardized form of the noncentral χ^2-distribution, as it appears in most text book. However, this analysis frequently requires a more general form, that incorporates the location parameter l and scale parameter s. The corresponding pdf can be written as

$$f_{\mathrm{NC}\chi^2_\nu}(x|\lambda,l,s) = \frac{1}{s}\cdot f_{\mathrm{NC}\chi^2_\nu}(y,\lambda) \qquad \text{with} \qquad y = \frac{x-l}{s} \tag{5.8}$$

$$\text{and} \qquad \lambda = \sum_{i=1}^{\nu} \frac{\mu_i^2}{s^2}. \tag{5.9}$$

The location parameter considers a constant contribution in the above sum of random numbers. The scale parameter s corresponds to the variance of the underlying normal distribution. Hence for $l = 0$ and $s = 1$ the standardized form is restored.

The mean of this distribution is given by

$$\mathrm{E}_{\mathrm{NC}\chi^2_\nu} = \nu s + l + \lambda s \tag{5.10}$$

and the variance by

$$\mathrm{Var}_{\mathrm{NC}\chi^2_\nu} = 2\nu s^2 + 4\lambda s^2 \tag{5.11}$$

[2] Here we also used the identity $\gamma(n+1,z) = n!\left(1 - e^{-z}\sum_{k=0}^{n}\frac{z^k}{k!}\right)$

[3] See footnote 1.

Furthermore, for large ν or large λ the noncentral χ^2-distribution converges towards a normal distribution.

5.3 Derivation of the Distribution of Fourier Powers

5.3.1 Sum or normal random variables

Let X_n be N independent random variables that are normally distributed, then their sum is also normally distributed [23], i.e.

$$\begin{aligned} X_n &\sim \mathcal{N}\left(\mu_n, \sigma_n^2\right) \\ Y &= \sum_{n=1}^{N} X_n \\ \Rightarrow Y &\sim \mathcal{N}\left(\sum_{n=1}^{N} \mu_n, \sum_{n=1}^{N} \sigma_n^2\right). \end{aligned}$$

5.3.2 Calculation of the Means and Variances of the real and imaginary parts of the Fourier coefficients

The real and and imaginary Fourier coefficients $\hat{x}_k$ and $\hat{y}_k$ in subsection 2.1.3 are the sum of N normal distributed random variables with varying mean $\mu_{x_{n,k}}$ and variance $\sigma^2_{x_{n,k}}$, and respectively $\mu_{y_{n,k}}$ and $\sigma^2_{y_{n,k}}$.

Using subsection 5.3.1 the means and variances of the normal distributions for the real and imaginary parts of the Fourier coefficients, given by Equation 2.10 and Equation 2.11 can be calculated as follows:

Means

$$\mu_{x_{n,k}} = \kappa \cdot \cos\left(-2\pi k \frac{n}{N}\right) \cdot (\mu_\mathrm{N} + S_n) \qquad S_n = A_\mathrm{S} \cdot \sin\left(2\pi f T \frac{n}{N}\right)$$

$$\mu_{y_{n,k}} = \kappa \cdot \sin\left(-2\pi k \frac{n}{N}\right) \cdot (\mu_\mathrm{N} + S_n)$$

The DFT defined as in Equation 2.1 used the so-called orthogonal normalization, meaning that the DFT and the inverse DFT are both normalized by the factor $1/\sqrt{N}$. Other used normalizations are the forward and backward normalizations, meaning, that the factor $1/N$ only appears in the DFT or respectively the inverse DFT. For the sake of generality, the normalization factor κ is used that can be either $\kappa = 1/\sqrt{N}$, $\kappa = 1/N$, $\kappa = 1$.

$$
\begin{aligned}
M_{\hat{x}_k} &= \sum_{n=0}^{N-1} \mu_{x_{n,k}} \\
&= \sum_{n=0}^{N-1} \kappa \cdot \cos\left(-2\pi k \frac{n}{N}\right) \cdot \left[\mu_{\mathrm{N}} + A_{\mathrm{S}} \cdot \sin\left(2\pi f T \frac{n}{N}\right)\right] \\
&= \kappa\mu_{\mathrm{N}} \sum_{n=0}^{N-1} \cos\left(-2\pi k \frac{n}{N}\right) + \kappa A_{\mathrm{S}} \sum_{n=0}^{N-1} \cos\left(-2\pi k \frac{n}{N}\right) \cdot \sin\left(2\pi f T \frac{n}{N}\right) \qquad (5.12) \\
&= \frac{\kappa A_{\mathrm{S}}}{2} \sum_{n=0}^{N-1} \sin\left(2\pi \frac{n}{N}\left[fT + k\right]\right) + \sin\left(2\pi \frac{n}{N}\left[fT - k\right]\right) \\
&= 0 \qquad \forall [fT - k] \in \mathbb{Z}
\end{aligned}
$$

$$
\begin{aligned}
M_{\hat{y}_k} &= \sum_{n=0}^{N-1} \mu_{y_{n,k}} \\
&= \sum_{n=0}^{N-1} \kappa \cdot \sin\left(-2\pi k \frac{n}{N}\right) \cdot \left[\mu_{\mathrm{N}} + A_{\mathrm{S}} \cdot \sin\left(2\pi f T \frac{n}{N}\right)\right] \\
&= \kappa A_{\mathrm{S}} \sum_{n=0}^{N-1} \sin\left(-2\pi k \frac{n}{N}\right) \cdot \sin\left(2\pi f T \frac{n}{N}\right) \qquad (5.13) \\
&= \frac{\kappa A_{\mathrm{S}}}{2} \sum_{n=0}^{N-1} \cos\left(2\pi \frac{n}{N}\left[fT + k\right]\right) - \cos\left(2\pi \frac{n}{N}\left[fT - k\right]\right) \\
&= \begin{cases} -\frac{1}{2} \cdot \kappa \cdot A_{\mathrm{S}} \cdot N & \text{if } [fT - k] = 0 \\ 0 & \text{if } [fT - k] \in \mathbb{Z}/\{0\} \end{cases}
\end{aligned}
$$

Here the identities

$$
\begin{aligned}
2\sin(u)\sin(v) &= \cos(u - v) - \cos(u + v) \\
2\sin(u)\cos(v) &= \sin(u - v) + \sin(u + v)
\end{aligned}
$$

$$
\sum_{n=0}^{N-1} \sin\left(2\pi \cdot kn/N\right) = 0 \ \forall k \in \mathbb{Z} \qquad \sum_{n=0}^{N-1} \cos\left(2\pi \cdot kn/N\right) = 0 \ \forall k \in \mathbb{Z}/\{0\} \qquad (5.14)
$$

are used together with the well known zero values $\sin(0) = 0$ and $\cos(0) = 1$. The summation identities can be derived from Lagrange's Trigonometric Identities [7]

$$
\sum_{n=0}^{N} \sin(n\theta) = \frac{\cos(\theta/2)}{2\sin(\theta/2)} - \frac{\cos\left(\left(N + \frac{1}{2}\right)\theta\right)}{2\sin(\theta/2)} \qquad (5.15)
$$

$$
\sum_{n=0}^{N} \cos(n\theta) = \frac{1}{2} + \frac{\sin\left(\left(N + \frac{1}{2}\right)\theta\right)}{2\sin(\theta/2)}. \qquad (5.16)
$$

Additionaly, using Lagrange's trigonometric identity $M_{\hat{y}_k}$ can be written in a more general form, that holds also for arbitrate non-integer $[fT-k] \in \mathbb{R}$, namely

$$M_{\hat{y}_k} = \frac{\kappa A_\mathrm{S}}{2}\left(1 + \frac{\sin\left(2\pi\frac{N-1}{N}[fT-k]\right)}{2\sin\left(2\pi\frac{1}{N}[fT-k]\right)} + \frac{\sin\left(2\pi\frac{N-1}{N}[fT+k]\right)}{2\sin\left(2\pi\frac{1}{N}[fT+k]\right)}\right). \tag{5.17}$$

For $[fT-k] \in \mathbb{Z}$ Equation 5.17 collapses as described above. However if $[fT-k] \in \mathbb{R}/\mathbb{Z}$, then $M_{\hat{y}_k} \neq 0\ \forall k$, i.e. if the signal frequency is not exactly sampled by the DFT, all frequency bins k experience a light shift on the imaginary axis, and not only the supposed frequency bin. This is the so-called scalloping effect.

Changing the signal from a sine to a cosine $M_{\hat{x}_k}$ will be shifted on the real axis towards $M_{\hat{x}_k} = +\frac{1}{2}\cdot\kappa\cdot A_\mathrm{S}\cdot N$, while $M_{\hat{y}_k}$ will remain at the origin. A combination of sine and cosine waves of same frequency, would therefore shift both means by the given values away from the origin. The noncentrality parameter of the noncentral χ^2-distribution would therefore be calculated as $\lambda = (M_{\hat{x}_k}^2 + M_{\hat{y}_k}^2)/S^2$.

Furthermore, in Equation 5.12 and Equation 5.13 it can be clearly seen, that any constant contribution to a signal would be canceled similarly as μ_N, the mean of the background noise.

Variances

$$\sigma_{x_{n,k}}^2 = \kappa^2\cdot\cos^2\left(2\pi k\frac{n}{N}\right)\cdot\sigma_\mathrm{N}^2 \qquad \sigma_{y_{n,k}}^2 = \kappa^2\cdot\sin^2\left(2\pi k\frac{n}{N}\right)\cdot\sigma_\mathrm{N}^2$$

$$S_x^2 = \sum_{n=0}^{N-1}\sigma_{x_{n,k}}^2 = \kappa^2\cdot\sigma_\mathrm{SN}^2\sum_{n=0}^{N-1}\cos^2\left(2\pi k\frac{n}{N}\right) = \kappa^2\cdot\sigma_\mathrm{N}^2\cdot\frac{N}{2} \tag{5.18}$$

$$S_y^2 = \sum_{n=0}^{N-1}\sigma_{y_{n,k}}^2 = \kappa^2\cdot\sigma_\mathrm{SN}^2\sum_{n=0}^{N-1}\sin^2\left(2\pi k\frac{n}{N}\right) = \kappa^2\cdot\sigma_\mathrm{N}^2\cdot\frac{N}{2} \tag{5.19}$$

Here the trigonometric identities

$$\sin^2(x) = \frac{1}{2}(1-\cos(2x))$$
$$\cos^2(x) = \frac{1}{2}(1+\cos(2x))$$

were used to transform the series of sine-squared into a series of sine as in Equation 5.14 and respectively for cosine.

The normalization factor κ of the DFT now gives some freedom about the exact form of the means and variances. In this analysis the symmetric orthogonal normalization ($\kappa = 1/\sqrt{N}$) is chosen, as then the χ^2 distribution for the background model is independent of the number of points N, as the variance S^2 also becomes independent of N. Note, that for this normalization the noncentrality parameter λ of the noncentral χ^2-distribution still depends on N. Note, that the chosen normalization factor is in practice however irrelevant, as due to the Red Noise Filter all χ^2-distributions will be normalized to zero mean and unit variance.

5.4 Expected Mean and Standard Deviation in the Red Noise Filter

Strong outliers have the potential to disturb the normalization process of the Red Noise Filter described in subsection 2.2.2. Using the properties of the noncentral χ^2-distribution to model a signal in a normalization segment, the response of the Red Noise Filter in dependence of the signal-to-noise ratio can be estimated. In the following the calculations to obtain the expected empirical mean and standard deviation are shown. To simplify the notation, the number of points in a segment is denoted by N instead of N_{Seg}.

Assumptions Assume $N-1$ points $\{X_1, \ldots, X_N\}/\{X_j\}$ distributed according to a χ^2_ν-distribution with arbitrary location parameter l and scale parameter s. Furthermore assume one point X_j distributed according to a noncentral χ^2_ν-distribution with the same arbitrary l and s, as well as with the noncentrality parameter λ.

Empirical Mean The expected value of the of the empirical mean μ_{RNF} of these N points can be calculated. For this, the linearity of the expected value and the known means of the 'central' and noncentral χ^2-distributions are used.

$$\begin{aligned} \mathrm{E}\left[\mu_{\mathrm{RNF}}\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[X_i\right] \\ &= \frac{1}{N}\sum_{i=1,i\neq j}^{N}\mathrm{E}\left[X_i\right] + \frac{1}{N}\mathrm{E}\left[X_j\right] \\ &= \frac{N-1}{N}\mathrm{E}_{\chi^2_\nu} + \frac{1}{N}\mathrm{E}_{\mathrm{NC}\chi^2_\nu} \\ &= \mathrm{E}_{\chi^2_\nu} + \frac{1}{N}\lambda s \end{aligned} \tag{5.20}$$

Empirical Standard Deviation The expected value of the empirical standard deviation σ_{RNF} of these N points is calculated in a similar manner, by first calculating the corresponding variance.

$$
\begin{aligned}
\mathrm{E}\left[\sigma^2_{\mathrm{RNF}}\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1,i\neq j}^{N}\mathrm{E}\left[\left(X_i - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] + \frac{1}{N}\mathrm{E}\left[\left(X_j - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1,i\neq j}^{N}\mathrm{E}\left[\left(X_i - \mathrm{E}_{\chi^2_\nu}\right)^2\right] + \frac{1}{N}\sum_{i=1,i\neq j}^{N}\mathrm{E}\left[\left(\mathrm{E}_{\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] \\
&\quad + \frac{1}{N}\mathrm{E}\left[\left(X_j - \mathrm{E}_{\mathrm{NC}\chi^2_\nu}\right)^2\right] + \frac{1}{N}\mathrm{E}\left[\left(\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] \\
&= \frac{N-1}{N}\mathrm{Var}_{\chi^2_\nu} + \frac{N-1}{N}\left(\mathrm{E}_{\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2 \\
&\quad + \frac{1}{N}\mathrm{Var}_{\mathrm{NC}\chi^2_\nu} + \frac{1}{N}\left(\mathrm{E}_{\mathrm{NC}\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2 \\
&= \mathrm{Var}_{\chi^2_\nu} + \frac{1}{N}4\lambda s^2 + \frac{N-1}{N}\left(-\frac{1}{N}\lambda s\right)^2 + \frac{1}{N}\left(\left(1-\frac{1}{N}\right)\lambda s\right)^2 \\
&= \mathrm{Var}_{\chi^2_\nu} + \frac{1}{N}4\lambda s^2 + \frac{N-1}{N^2}\lambda^2 s^2
\end{aligned}
\tag{5.21}
$$

Here additionally the following relation was used to decompose the variance,

$$
\begin{aligned}
\mathrm{E}\left[\left(X_i - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] &= \mathrm{E}\left[\left(\left(X_i - \mathrm{E}_{\chi^2_\nu}\right) + \left(\mathrm{E}_{\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)\right)^2\right] \\
&= \mathrm{E}\left[\left(X_i - \mathrm{E}_{\chi^2_\nu}\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}_{\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)^2\right] \\
&\quad + \underbrace{\mathrm{E}\left[\left(X_i - \mathrm{E}_{\chi^2_\nu}\right)\right]}_{=0}\left(\mathrm{E}_{\chi^2_\nu} - \mathrm{E}\left[\mu_{\mathrm{RNF}}\right]\right)
\end{aligned}
\tag{5.22}
$$

and similarly for X_j.

The expected value of the empirical standard deviation can then be simply obtained by taking the square root of $\mathrm{E}\left[\sigma^2_{\mathrm{RNF}}\right]$.

Acknowledgements

In this part I want to thank everyone who helped my during this thesis.

- PD Dr. Thomas Eberl, for the opportunity to work on this project, and the many experiences I have gained along this way.
- Rodrigo Gracia-Ruiz for the tireless help and support during this whole time.
- Felix Trunk for helping me with my questions and for proofreading this thesis.
- All members at ECAP who have accompanied me on my way.
- My friends and family for their unending support during this time.

Thank you very much!

Erklärung / Statement of Authorship

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

I hereby confirm that I have completed this thesis independently and only using the specified sources and tools.

Erlangen, 20. January 2023

Maximilian Eff