Period determination of variable sources from Gaia DR3 using the nonuniform discrete Fourier transform

Pranger, A

Period determination of variable sources from Gaia DR3 using the nonuniform discrete Fourier transform

THESIS

submitted in partial fulfillment of the
requirements for the degree of

BACHELOR OF SCIENCE
in
MATHEMATICS AND ASTRONOMY

Author : A. Pranger
Student ID : s2513323
Supervisor : dr. A.G.A. Brown, dr. ir. O.W. van Gaans

Leiden, The Netherlands, July 6, 2023

Period determination of variable sources from Gaia DR3 using the nonuniform discrete Fourier transform

A. Pranger

Leiden Observatory, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

July 6, 2023

Abstract

There are many different classes of variable astrophysical sources (which have a luminosity that varies over time), all having a certain physical phenomenon causing their variability. This results in different characteristic light curves, often containing periodicities within certain ranges of frequencies. The Gaia satellite telescope has gathered photometric data of variable sources, at semi-random nonuniform observation times, influenced by the Gaia scanning law. This research aims to use the nonuniform fast Fourier transform (NUFFT) to retrieve the main frequency of the brightness variations of the variable source from the photometric Gaia Data Release 3 data, where it is assumed that the underlying signal has one main frequency. The main goals are to investigate whether the frequency with maximal power in the NUFFT periodogram is the main frequency of the underlying signal and whether it is possible to distinguish between in this way correctly and incorrectly determined frequencies. To this end a simulation of photometric data is used, where the time series are taken from actual Gaia DR3 data and the signal is simulated as a sine wave with a known frequency and a signal to noise equal to 5. Taking the frequency with maximal power from the corresponding periodograms results in a correct retrieval in about 90% of the simulated cases. A positive correlation between the number of data

points or visibility periods and the fraction of correctly determined frequencies is found. The incorrectly determined frequencies are most likely caused by spurious periods or aliasing. Furthermore, a method to compute a false alarm probability (FAP) for the determined frequency was investigated, but turned out to give no useful results as almost all FAPs were equal to zero. Therefore, further research on other methods is necessary to find out how to correctly identify the main frequency.

Contents

Chapter 1

Introduction

Variable sources have a luminosity that varies over time. Examples of variable sources are pulsating stars, planetary transits, rotating stars and eclipsing binaries [1]. In order to determine whether a source is variable and, if so, what kind of variable source it is, the photometric data of such a source over time is investigated. For example, the decomposition of the light signal into its underlying frequencies can help us to learn more about which and how many frequencies are most important within the signal, which can in turn uncover more about the cause of the variability. This information can, therefore, contribute to the classification of the source. By classifying many variable sources, large samples of similar types of sources are obtained, which gives the opportunity to investigate these classes of sources even further [2].

With photometric data of more than 11 million sources in Data Release 3 (DR3), the Gaia satellite telescope data is very suitable to classify variable sources. As this telescope is a satellite moving around multiple axes, it is not possible to take photometric data of one source at uniform time intervals, but the time intervals are also not completely random. In order to find the main frequency of a large amount of variable sources, we need to compute the frequency spectra of their flux signals. Therefore, we need to have a method to compute the frequency spectrum (also called periodogram) for nonuniformly sampled data. This problem has been investigated for many years and there are multiple ways to perform this task. One such method is called the Lomb-Scargle periodogram [3], which is also used by the Gaia Data Processing and Analysis Consortium for their determination of the periods of variable sources [4]. Another method uses the nonuniform discrete Fourier transform.

The Fourier transform is an operator that turns a signal into the spec-

trum of frequencies of which the signal consists. Since its first introduction the Fourier transform has been investigated further and many variants of the computation of the transform have been introduced, some specifically designed for the type of data it is to be used on. For example, the so called discrete Fourier transform (DFT) is designed specifically to yield the underlying frequency spectrum of a discrete signal that is sampled uniformly, both in the time and frequency domain (meaning that the sample distance between two data points, both in time and in frequency, is a fixed value). Implementations of this theoretic discrete Fourier transform are generally referred to as fast Fourier transform (FFT). A variant of the DFT is the so called nonequispaced / unequally sampled / nonuniform discrete Fourier transform (NDFT / USDFT / NUDFT), which does not require the discrete signal to have a uniform sampling. In this report we will further call this transform the nonuniform discrete Fourier transform. Again, the implementation of this transform is called the nonuniform fast Fourier transform (NUFFT) and this can be computed in different ways [5, 6]. As the Gaia photometric data is not uniformly sampled, the NUFFT should theoretically be a correct method to perform the task of recovering the frequency spectra of variable sources within this data.

In this report we investigate whether the NUFFT periodogram can be used to recover the main frequency from variable source photometric data of the Gaia satellite. More specifically, we assume that the underlying signal of the variable source consists of one main frequency. Then we use the NUFFT to compute the corresponding periodogram and we analyse whether the method of taking the frequency with maximal power from this periodogram gives us the correct main frequency of the underlying signal and whether it is possible to distinguish between correctly and incorrectly determined frequencies.

To this end, we firstly discuss the background information needed to understand the research area. This consists of theory about the variable sources and information about the Gaia DR3 data used. Secondly, we consider the theory behind the nonuniform discrete Fourier transform, which tells us what the general limitations and points of attention of this method are. Thirdly, we elaborate on the method used to do the period search using the NUFFT, which does not only consist of the application of the NUFFT, but also of the determination of the most important frequency from the recovered periodogram. Continuing, we apply this method on simulated photometric Gaia data, of which we know the simulated frequency and can thus investigate whether we retrieve the correct frequency. Lastly, we discuss all results and draw the conclusions.

Chapter 2

Background information

In this chapter we discuss the most important background information necessary to understand the research conducted. We start by explaining what variable sources are, giving some examples of classes of variable sources caused by different physical phenomena and showing how their luminosity signals can differ. Secondly, we introduce the data that is used in this research and elaborate on how the Gaia scanning law influences the time series of the data, causing spurious periods. Lastly, we investigate theoretically how the nonuniform discreteness of the data influences the resulting periodogram.

2.1 Variable sources

There are many different kinds of variable sources on the sky, all giving rise to a luminosity signal that varies over time. They can be classified by the physical phenomena that cause their variability. These phenomena can be either intrinsic or extrinsic, which respectively means that the source itself has a varying luminosity or the variability is caused by other visual effects (with respect to us as observer).

An example of an intrinsic variable is a pulsating variable, which is a star that has a radius which alternatingly increases and decreases over time. The luminosity of these stars varies with the radius, which can be either periodic, semi-regular or irregular. Examples of light curves resulting from pulsating stars are shown in the first and second to sixth rows of Figure 2.1. Another example of intrinsic variables are eruptive variables, which are stars that for example undergo mass ejection. These flares of ejecting mass cause irregular variability of the luminosity.

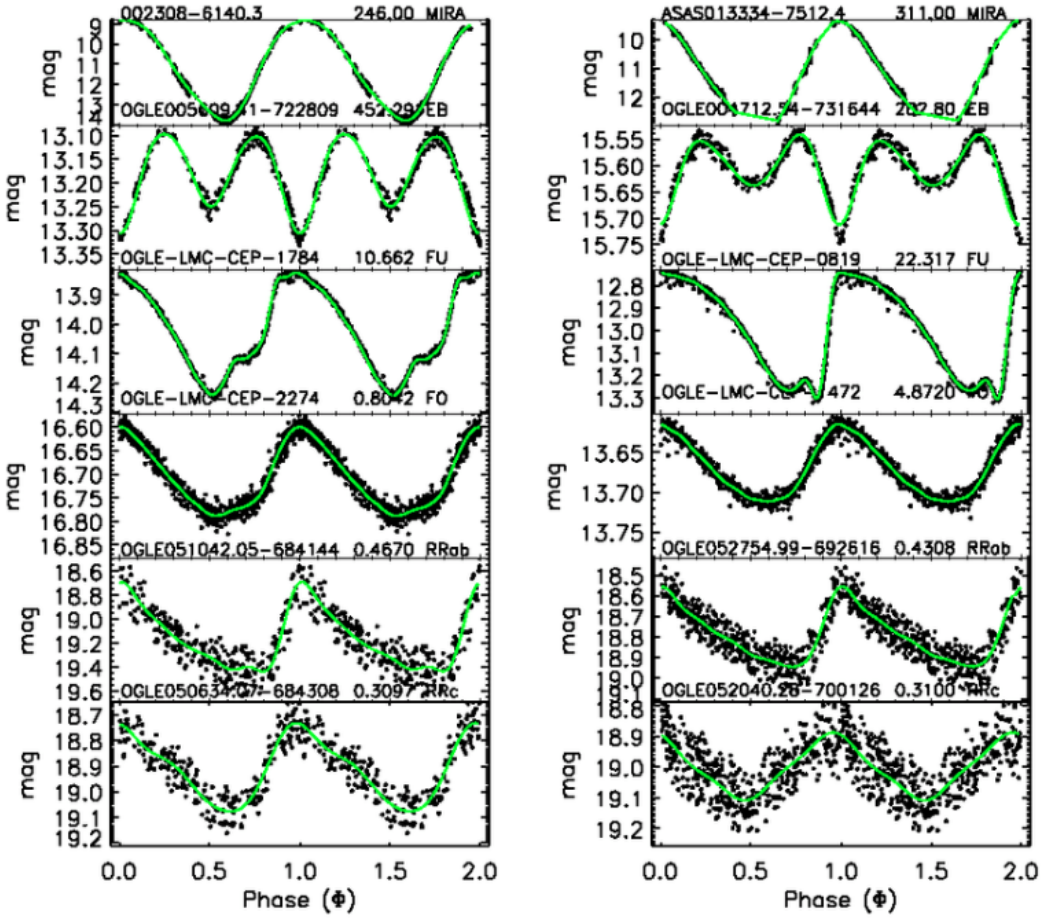

Figure 2.1: *Folded light curves of different types of variable sources. Two phases are plot for each source. The caption at the top of each panel shows the variable name, period in days and type of variables respectively. From the top down we have Mira variables (MIRA), Eclipsing binaries (EB), Cepheid variables (Fundamental mode (FU) and First Overtone (FO)) and RR Lyrae variables (RRab, RRc). Sources in the same row are of the same class. The green lines show optimal fits of the light curves. From [7].*

Examples of extrinsic variables are planetary transits, eclipsing binaries and rotating stars. A planetary transit causes the light of the host star to be blocked for a certain amount of time, which results in a dip in the luminosity signal of the star. As the planet's orbit around the star is periodic, the varying luminosity signal of the star will show the same periodicity. Eclipsing binaries cause variability similar to that of planetary transits, as the blocking of light caused by one of the stars in front of the other results in a decrease of the total luminosity of the binary system. However, for

eclipsing binaries both eclipsing objects emit light, but they do not have to have the same luminosity nor the same size. Because of this, there are effectively two different eclipsing moments: star 1 in front of star 2 and the other way around. Each of these two eclipsing moments results in its own specific dip in the total luminosity of the system, which alternate each other. Examples of light curves resulting from eclipsing binaries are shown in the second row of Figure 2.1. Rotation of stars can also cause variability in the luminosity, for example because star spots move in and out of sight. These variations are again periodic, because of the periodic rotation of the star.

All these different classes of variable sources can be categorised according to their physical origin. Such a categorisation is shown in Figure 2.2, where sources other than stars are also included. More detailed explanations of the physical processes behind many variable source classes can be found in [1].

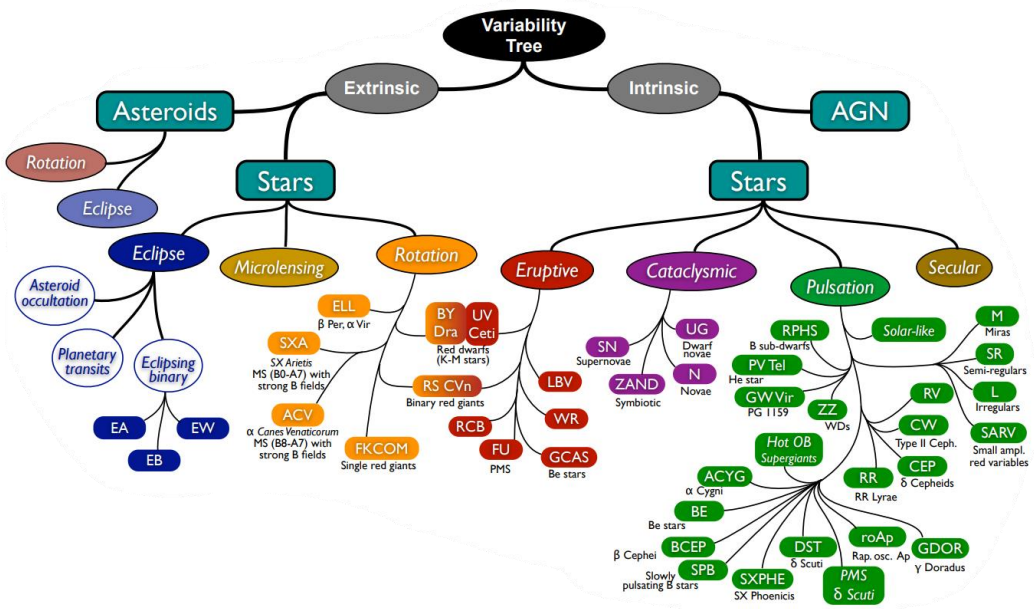

Figure 2.2: *Variability tree showing a categorisation of variable source classes. From [2].*

As one can see from Figure 2.1, the luminosity signals for different classes of variable sources can be quite different. Most classes, however, have very specific light curves, with a clearly recognisable periodicity and /or form of the signal, as we can also see from the similarities between the light curves situated on the same row in Figure 2.1, which come from

sources of the same class. The luminosity signals can therefore be used to classify observed variable sources.

Many of the classes do have periodic variability. Important describing parameters of such luminosity variability are the main frequency and amplitude of the signal. As these parameters are influenced greatly by the physical origin of the variability, as can also be seen in Figure 2.1 where similar sources have periods of the same order of magnitude, many classes have specific frequency and amplitude ranges for their luminosity signals. Thus, retrieving these parameters from the data helps with the classification of the source.

2.2 Gaia DR3 photometric data

In this research, photometric data of variable sources from Gaia Data Release 3 [8] is considered. The ultimate goal is to determine the periods of such variable sources by using the nonuniform Fourier transform on the data and investigating the resulting periodogram. The way the data is gathered by the Gaia satellite affects the nonuniformity of the time samplings, which can influence the resulting periodogram. Therefore, we will discuss the data in this section.

Gaia DR3 includes photometric data of about 11 million variable sources, in each of the photometric bands G, G_{RP} and G_{BP}. The photometric data of one source in a certain band consists of the Julian Date at which the flux of the source is measured in units of days and the corresponding measured flux F in electrons per second in the given photometric band. In this research, we only work with the data in the G-band. An example of the considered photometric data for one source is given in Figure 2.3.

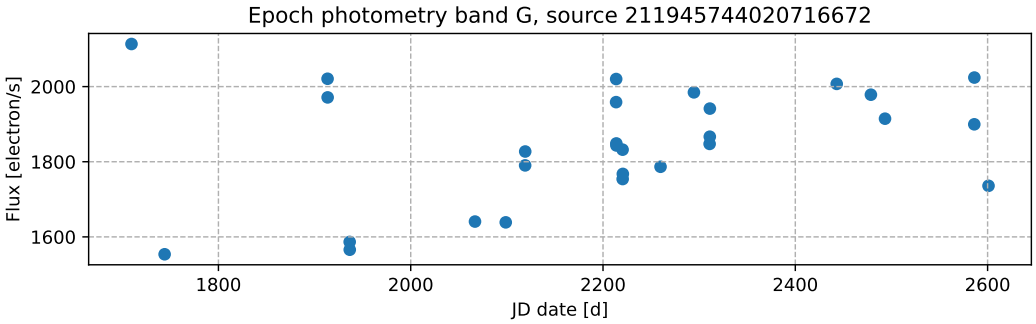

Figure 2.3: *The photometric data in the G-band from Gaia DR3 of one source (see title).*

One can see that the time between two successive measurements differs a lot. By eye, the measurement times seem to be quite random. However, they are not actually completely random.

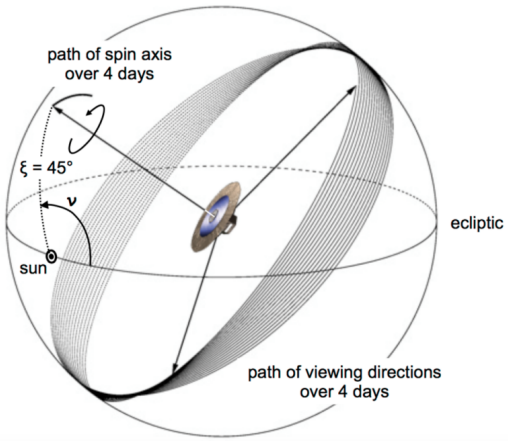

Figure 2.4: *Illustration of the scanning law of Gaia caused by the movement of the satellite telescope. It shows the path of the spin axis (**z**), and the corresponding path of the preceding viewing direction, during four days. From [9].*

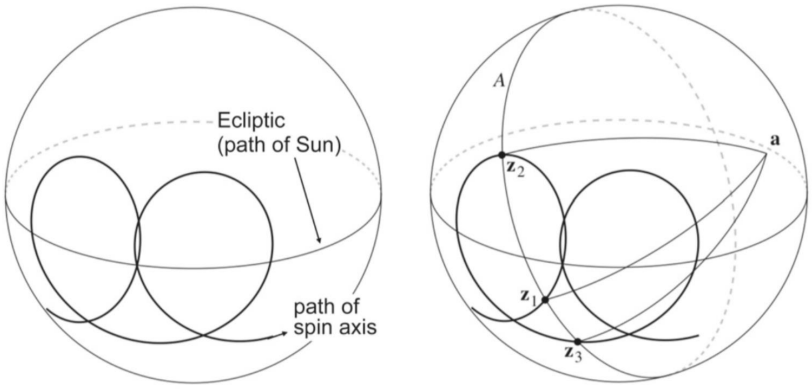

Figure 2.5: *Overview of the Gaia scanning law. Left: During the nominal scanning law, the spin axis **z** makes overlapping loops around the Sun at a separation of 45° and rate of 5.8 cycles per year. Right: One source at point **a** may be scanned whenever **z** is 90° from **a**, that is, on the great circle A at $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, etc. From [10].*

This semi-randomness can be explained by the Gaia scanning law, which describes the movement of the field-of-view of the satellite telescope on the sky. Gaia scans the sky using uniform revolving scanning, which maximises the uniformity of the sky coverage [9]. A few important aspects of

the scanning law include the rotation of the spacecraft around its spin axis z with a period of 6 hours, the fixed angle of 45° between the sun and the spin axis and the rotation of the spin axis around the sun with a period of about 63 days.

In Figures 2.4 and 2.5 an overview of the nominal scanning law is given. They show that the movement of the field-of-view indeed has a smooth and periodic nature, which causes a source to be observed at times that are definitely not uniformly spaced, but also not random. Furthermore, Figure 2.4 shows that the great circles describing the viewing directions overlap, causing sources close to these points to possibly be scanned multiple times over a short period. These clustered observations can be so close together, that they do not give extra information about the light curve of the source, as the flux hardly changes between the observations. This is why we later use the Gaia parameter called 'visibility_periods_used', which describes the number of visibility periods of a time series, as a measure of the effective number of observations. A visibility period is here defined as a cluster of observations separated from other clusters by a gap of at least 4 days.

The scanning law also induces that some parts of the sky are scanned more often and with other frequencies than other parts. The number of field-of-view observations as a function of ecliptic coordinates during the Gaia DR3 time range is shown in Figure 2.6. We use this information later in Section 4.2.1 to select variable sources with varying numbers of fov observations, in order to get a more diverse set of time series.

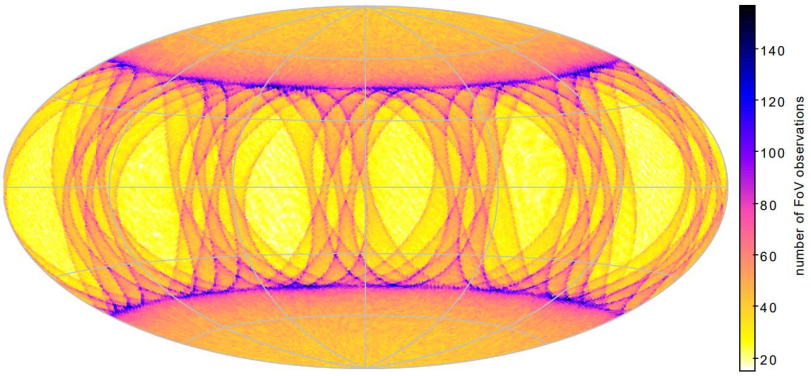

Figure 2.6: *Ecliptic coordinate plot with longitude zero at the centre and increasing to the left, showing the simulated number of field-of-view observations during the nominal scanning law phase of the Gaia DR3 time range. From [10].*

The periodicities that are present in the time sampling caused by the scanning law affect the resulting periodogram, as we will discuss in Chapter 3. In Figure 2.7 the distributions of the Baluev false alarm probabilities [11] for two photometric samples (consisting of respectively 73k and 38k time series) of Gaia DR3 are shown. The false alarm probability for one of the time series is provided by the distribution of the maximal periodogram value in the case there is no signal present in the data. The false alarm probability for a certain period for a certain time series thus indicates the probability to find that period having maximal power in the periodogram of photometric data without any signal being present in the data. In the figure the false alarm probabilities for each period for each of the time series of the photometric sample are scattered, where the probabilities equal to zero are omitted. The figure thus illustrates with the dense regions of dots at certain periods that indeed specific periods occur more often with maximal power, only because of the time sampling of the data (as there is no signal present when computing the false alarm probability). For example, the clear dense region at a period of 31.5 days is caused by the earlier mentioned rotation of the spin axis around the sun, which proceeds with a period of $2 \times 31.5 = 63$ days. The periods of the other dense lines correspond in similar ways to multiples of periods that are induced by the scanning law. These spurious periods are therefore important to take into account when computing the main period of a signal.

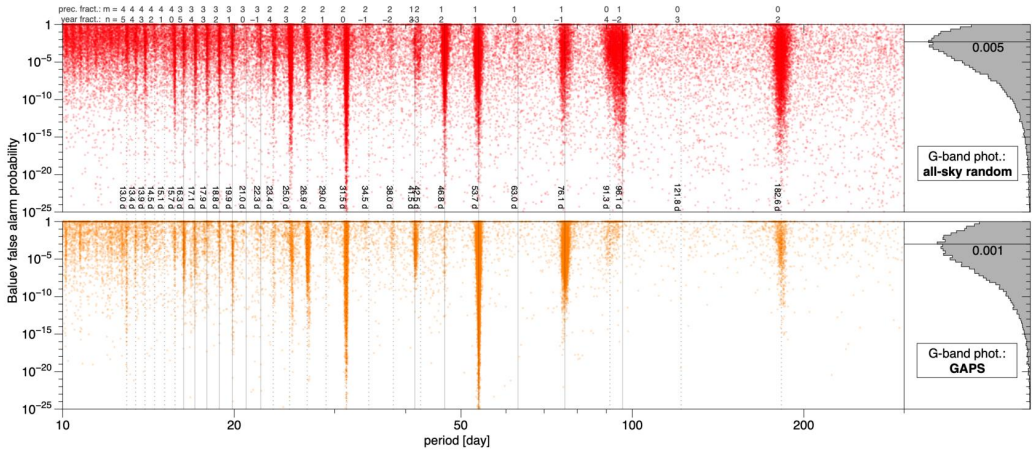

Figure 2.7: *Distributions of Baluev false alarm probabilities of two photometric samples, illustrating the highly significant nature of most of the spurious periods. From [10].*

The Gaia DR3 sources are given a variability flag by the Gaia Data Processing and Analysis Consortium [4], which indicates whether the source is variable or not. We only use the photometric *G*-band data of variable sources and we assume in our research that the data we use comes from a variable source.

Chapter 3

Theory of the periodogram

The data that we work with, the observations of the flux of a source at certain moments, are only a sample of the underlying signal, which is the continuous flux of the source. The goal of this research is to find the main frequency of this signal, assuming that the source has a varying flux which can be described as a signal with one main period, using the NUFFT periodogram. However, in order to decide which frequency is the main frequency that we are looking for in this periodogram, we need to understand what the actual meaning of such a periodogram is. Therefore, we dedicate this chapter to the analysis of the origin and mathematics of the periodogram. The content discussed in this chapter is inspired by [3, 12–14], where more information on the topics can also be found. As the continuous signal underlying the discrete observations is the original signal of which we want to determine the period, this chapter starts with the corresponding continuous Fourier transform.

3.1 Continuous Fourier transform

We can describe the continuous underlying signal as a continuous function $g : \mathbb{R} \to \mathbb{C}$ that assigns a flux to every time $t \in \mathbb{R}$. In order to be able to define the continuous Fourier transform of such a function, we first need to have a notion of integrability.

Definition 1. Let $g : \mathbb{R} \to \mathbb{C}$. We say that g is *integrable* if it is Borel measurable and the Lebesgue integral $\int_{-\infty}^{\infty} |g(t)| \mathrm{d}t$ is finite [15].

Definition 2. Let $g : \mathbb{R} \to \mathbb{C}$ be an integrable function. The *continuous*

Fourier transform $\hat{g} : \mathbb{R} \to \mathbb{C}$ of g is then given by

$$\hat{g}(\omega) := \int_{-\infty}^{\infty} g(t)e^{i\omega t}\mathrm{d}t. \tag{3.1}$$

Indeed, this is well defined, since $t \mapsto e^{i\omega t}$ is continuous and $|e^{i\omega t}| = 1$ for all $t \in \mathbb{R}$, so for every integrable $g : \mathbb{R} \to \mathbb{C}$ the function $t \mapsto g(t)e^{i\omega t}$ is also integrable for every $\omega \in \mathbb{R}$.

Depending on what kind of function g is and what the (physical) meaning of this function is, the Fourier transform $\hat{g}$ has different interpretations. In our case, as g is a function of time t, $\hat{g}$ can be interpreted as a function of angular frequency ω. It then denotes to what extend each angular frequency is present in the data. Of course, the Fourier transform is complex valued. Thus, to get the right interpretation we take the absolute value squared of the Fourier transform, which gives us the resulting power of each angular frequency within the original time domain signal. The resulting spectrum is called the periodogram.

Consider the Fourier transform $\mathcal{F} : g \mapsto \hat{g}$ with domain given by the set of all integrable functions $g : \mathbb{R} \to \mathbb{C}$, which is a vector space. The function $\mathcal{F}$ has a couple of useful properties, such as linearity, invertability and shifting, which we shortly discuss now as we will use them later. (Note that both the function $\mathcal{F} : g \mapsto \hat{g}$ and $\hat{g}$ itself are called the Fourier transform, but $\hat{g}$ is specifically the Fourier transform of the function g, while $\mathcal{F}$ is the function which sends any integrable function to its Fourier transform.)

Theorem 3 (Linearity)**.** *The Fourier transform $\mathcal{F} : g \mapsto \hat{g}$ is linear, meaning that for two integrable functions $g, h : \mathbb{R} \to \mathbb{C}$ and constants $\lambda, \gamma \in \mathbb{R}$ we have that*

$$\lambda\mathcal{F}(g) + \gamma\mathcal{F}(h) = \mathcal{F}(\lambda g + \gamma h). \tag{3.2}$$

Proof. This follows immediately from the definition of the Fourier transform. □

Theorem 4 (Shifting)**.** *Let $g : \mathbb{R} \to \mathbb{C}$, $t \mapsto g(t)$ be an integrable function of t. Consider the shifted version of g, denoted $g_a : \mathbb{R} \to \mathbb{C}$, $t \mapsto g(t - a)$, which is thus integrable as well. Then*

$$\mathcal{F}(g_a) = e^{i\omega a}\mathcal{F}(g). \tag{3.3}$$

Proof.

$$\begin{aligned}\mathcal{F}(g_a) &= \int_{-\infty}^{\infty} g_a(t)e^{i\omega t}\mathrm{d}t = \int_{-\infty}^{\infty} g(t-a)e^{i\omega t}\mathrm{d}t \\ &= \int_{-\infty}^{\infty} g(\tau)e^{i\omega(\tau+a)}\mathrm{d}\tau = e^{i\omega a}\int_{-\infty}^{\infty} g(\tau)e^{i\omega\tau}\mathrm{d}\tau \\ &= e^{i\omega a}\mathcal{F}(g).\end{aligned}$$

□

With other words, the Fourier transform of a shifted function is the same as the Fourier transform of the original unshifted function, except for a phase shift. As we are in the end interested in the squared absolute value (power) of the Fourier transform, shifting a function does not influence the determined main frequency. This is what we would expect, as the frequency decomposition of a signal does not change as we shift the signal in time.

Theorem 5 (Inverse)**.** *The Fourier transform is invertible. Let $\hat{g} : \mathbb{R} \to \mathbb{C}, \omega \mapsto g(\omega)$ be an integrable function. The inverse Fourier transform $g : \mathbb{R} \to \mathbb{C}, t \mapsto g(t)$ of $\hat{g}$ is then given by*

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{g}(\omega)e^{-i\omega t}d\omega = \frac{1}{2\pi}\hat{\hat{g}}(-t) \text{ almost everywhere.} \tag{3.4}$$

We can also write $g = \mathcal{F}^{-1}(\hat{g})$ where $F^{-1} : \hat{g} \mapsto g$ is the inverse of $\mathcal{F} : g \mapsto \hat{g}$.

Proof. See [12] page 80. □

Thus, a function and its Fourier transform are uniquely coupled to each other. Because of this, the Fourier transform of a function theoretically contains all information of the original function, so we can always recover the original function from its Fourier transform.

To get some more insight into what the Fourier transform does, we will now look at some examples. Firstly, we look at the Fourier transform of the delta "function", which comes back often in Fourier analysis. There is a problem here: the delta "function" is not a function, but a distribution. Because of this, we are not able to apply our notion of the Fourier transform. A similar problem will arise later when we want to apply the Fourier transform on the sin function, which is not integrable. Both these problems can be solved by using the Fourier transform for distributions. This theory is explained and derived in Appendix A, where some examples of

mathematically correct calculations are also given. However, those rigorous calculations are more time consuming and harder to follow, which is why we use a more intuitive notation in the rest of this chapter. Note that all calculations should be interpreted in the sense of distributions.

Example 6. Let $\delta(t) \approx \begin{cases} \infty \text{ for } t = 0, \\ 0 \text{ else,} \end{cases}$ so that $\int_{-\infty}^{\infty} \delta(t)\mathrm{d}t = 1$. An important property of this delta "function" is that $\int_{-\infty}^{\infty} \delta(t)v(t)\mathrm{d}t = v(0)$ for well behaved functions v. The Fourier transform of δ, in the sense of distributions, is given by

$$\hat{\delta}(\omega) = \int_{-\infty}^{\infty} \delta(t)e^{i\omega t}\mathrm{d}t = e^{i\omega 0} = e^0 = 1. \tag{3.5}$$

From the shifting rule, we find for the shifted delta function $\delta(t-a)$ that

$$\mathcal{F}(\delta(t-a)) = e^{i\omega a}\mathcal{F}(\delta(t)) = e^{i\omega a}. \tag{3.6}$$

Because of this and the inverse Fourier theorem, we now know that

$$\delta(t-a) = \mathcal{F}^{-1}(e^{i\omega a}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega a}e^{-i\omega t}\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(a-t)}\mathrm{d}\omega. \tag{3.7}$$

This is a useful expression of the delta function that we will see more often.

From this example, we can now easily derive the Fourier transform of a constant function.

Example 7. Let $g(t) = C$ be a constant function. Then

$$\hat{g}(\omega) = \int_{-\infty}^{\infty} g(t)e^{i\omega t}\mathrm{d}t = C\int_{-\infty}^{\infty} e^{i\omega t}\mathrm{d}t = 2\pi C\delta(\omega). \tag{3.8}$$

In our research, we look at a signal of which we assume it has one main frequency. A common example of such a signal is a sine wave, which we also use later in the simulation. This kind of periodic (harmonic) signal is one that occurs often in nature and in variable sources as well. Its Fourier transform is as we expect:

Example 8. Let $g(t) = C + A\sin(Bt)$. Then

$$\begin{aligned}\hat{g}(\omega) &= \mathcal{F}(g(t)) = \mathcal{F}(C) + A\mathcal{F}(\sin(Bt)) \\ &= 2\pi C\delta(\omega) + A\int_{-\infty}^{\infty} \sin(Bt)e^{i\omega t}\mathrm{d}t \\ &= 2\pi C\delta(\omega) + A\int_{-\infty}^{\infty} \frac{e^{iBt} - e^{-iBt}}{2i}e^{i\omega t}\mathrm{d}t \\ &= 2\pi C\delta(\omega) + \frac{A}{2i}\int_{-\infty}^{\infty} e^{i(\omega+B)t} - e^{i(\omega-B)t}\mathrm{d}t \\ &= 2\pi C\delta(\omega) + \frac{A\pi}{i}(\delta(\omega+B) - \delta(\omega-B)). \end{aligned} \tag{3.9}$$

Indeed, the Fourier transform gives us peaks at the angular frequencies B and $-B$ of the sine wave and one at frequency 0, which corresponds to the constant shift (in flux) C of the sine wave. If C is too big, the peak at zero will be higher than all other peaks, resulting in this peak dominating the frequency spectrum as the highest peak. However, in our situation, this is not the peak we are looking for, as we look for the peaks corresponding with the frequency of the signal. Therefore, we often subtract the mean from the signal before we compute the Fourier transform, so that the peak at zero is not so dominating anymore. Of course, subtracting the mean does not have any effect on the frequencies of the signal and, as the Fourier transform is linear, it indeed also does not change the other peaks present in the frequency spectrum.

However, we do not have data of an infinite sine wave at our disposal, as all measurements lie within some finite time interval. We can account for this by saying that our original infinite signal is now pointwise multiplied by some box function, often referred to as a rectangular window, which takes on the value 1 within the observation time interval and the value 0 outside this range. To understand what this pointwise multiplication of two functions does with the corresponding Fourier transforms, we need to introduce another concept, called the convolution.

Definition 9 (Convolution)**.** Let $g, h : \mathbb{R} \to \mathbb{C}$ be integrable functions. The convolution $g \star h$ of g and h is defined by

$$(g \star h)(t) := \int_{-\infty}^{\infty} g(\tau)h(t-\tau)\mathrm{d}\tau. \tag{3.10}$$

This is well defined, since $g(\tau)h(t-\tau)$ is integrable for all t as g and h are integrable.

Theorem 10 (Convolution theorem)**.** *Let $g, h : \mathbb{R} \to \mathbb{C}$ be integrable functions. Then*

$$\mathcal{F}(g \star h) = \mathcal{F}(g) \cdot \mathcal{F}(h), \tag{3.11}$$

where $\cdot$ denotes pointwise multiplication.

Proof. See [12] page 86. □

However, in our case we want to compute the Fourier transform of the pointwise multiplication of two functions, which is just slightly different from the convolution theorem. Luckily, there is a similar rule for this situation.

Lemma 11. *Let $g, h : \mathbb{R} \to \mathbb{C}$ be integrable functions such that $g \cdot h$ is also integrable. Then*

$$\mathcal{F}(g \cdot h) = \mathcal{F}(g) \star \mathcal{F}(h). \tag{3.12}$$

With this statement, we can investigate what the effect of the pointwise multiplication of the rectangular window with the sine function is on its Fourier transform.

Example 12. Let the rectangular window function $W(t)$ be given by

$$W(t) := \begin{cases} \frac{1}{T} \text{ for } t \in \left[-\frac{T}{2}, \frac{T}{2}\right], \\ 0 \text{ else.} \end{cases} \tag{3.13}$$

The Fourier transform of this window function is

$$\begin{aligned} \hat{W}(\omega) &= \int_{-\infty}^{\infty} W(t) e^{i\omega t} \mathrm{d}t = \int_{-\frac{T}{2}}^{\frac{T}{2}} \frac{1}{T} e^{i\omega t} \mathrm{d}t = \left[\frac{1}{i\omega T} e^{i\omega t} \right]_{-\frac{T}{2}}^{\frac{T}{2}} \\ &= \frac{e^{\frac{i\omega T}{2}} - e^{\frac{-i\omega T}{2}}}{i\omega T} = \frac{2}{\omega T} \frac{e^{\frac{i\omega T}{2}} - e^{\frac{-i\omega T}{2}}}{2i} = \frac{2}{\omega T} \sin\left(\frac{\omega T}{2}\right) \\ &= \operatorname{sinc}\left(\frac{\omega T}{2\pi}\right). \end{aligned} \tag{3.14}$$

With Lemma 11 we can now compute the Fourier transform of the signal $s(t) := g(t)W(t)$ that results from the pointwise multiplication of the rectangular window with the original infinite signal $g(t) = C + A\sin(Bt)$ of Example 8.

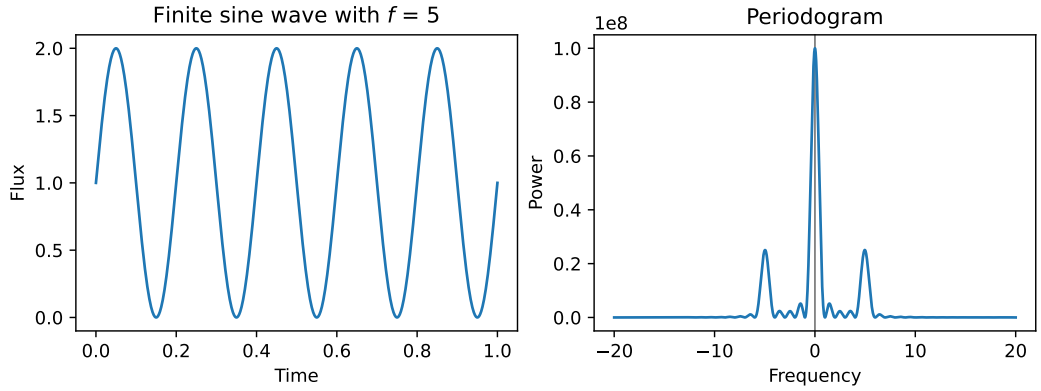

Figure 3.1: *An example of the effect of the finiteness of a continuous sine wave (left) on the corresponding periodogram (right), resulting in spectral leakage. There are peaks at frequencies 0 and ± 5, which is as computed for a sine wave with nonzero equilibrium value and frequency 5. The peaks are sinc formed and have a width at half maximum equal to $\frac{1}{T}$ where T is the total observation time, in this case both T and the width of the peaks are thus equal to 1. Between these main peaks smaller side lobes are visible. (Note that the used frequency is not the angular frequency ω but $f = \frac{\omega}{2\pi}$.)*

Example 13.

$$\begin{aligned}
&\mathcal{F}(s(t)) = \mathcal{F}(g(t)W(t)) = \mathcal{F}(g(t)) \star \mathcal{F}(W(t)) \\
&= \int_{-\infty}^{\infty} \hat{g}(x)\hat{W}(\omega - x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(2\pi C\delta(x) + \frac{A\pi}{i}(\delta(x+B) - \delta(x-B))\right) \mathrm{sinc}\left(\frac{(\omega - x)T}{2\pi}\right) \mathrm{d}x \\
&= 2\pi C \mathrm{sinc}\left(\frac{\omega T}{2\pi}\right) + \frac{A\pi}{i}\left(\mathrm{sinc}\left(\frac{(\omega + B)T}{2\pi}\right) - \mathrm{sinc}\left(\frac{(\omega - B)T}{2\pi}\right)\right)
\end{aligned} \tag{3.15}$$

Instead of sharp peaks at the relevant frequencies (at zero and at the frequency of the sine wave $\pm B$), we now find peaks at the same frequencies that have a width at half maximum equal to $\frac{1}{T}$ and smaller side lobes. This effect is often called *spectral leakage*, as some of the power at the frequencies of the spectrum seems to leak to nearby frequencies. Because of this, the power at a certain frequency is no longer completely independent from the powers of other frequencies. Furthermore, the widening of the true peaks influences the spectral resolution of the periodogram, as this effect can cause peaks at nearby frequencies to overlap, making it impossible to distinguish them. These effects thus make the interpretation of the

periodogram much less straightforward. An example of spectral leakage caused by the finiteness of the signal is given in Figure 3.1.

From the computation of the Fourier transform, we can also see that the greater T (so the wider the window, meaning the greater part of the original infinite signal is observed), the smaller the width of the peaks in the spectrum, because the sinc function has a width at half maximum equal to $\frac{1}{T}$. This is exactly what we would expect, as in the limit $T \to \infty$ we recover the complete infinite signal which has infinitely narrow peaks.

We have now converted the original infinite signal to a finite signal using a rectangular window function. However, our observations do not cover an entire finite time interval, but are a discrete sampling of times within a finite time range. We will now look into how we can describe such a discrete signal and how this affects the corresponding periodogram.

3.2 Discrete Fourier transform

Let us start with the assumption that our observations are uniformly spaced throughout time, so at times $n\Delta T$ for $n \in \mathbb{Z}$ and $\Delta T \in \mathbb{R}_{>0}$ the time interval between two measurements. In this case, we can describe our observations by multiplying the previously described finite signal by a Dirac comb function, introduced in Example 14.

Example 14. Let the Dirac comb function be defined by

$$\text{Ш}_{\Delta T}(t) := \sum_{n=-\infty}^{\infty} \delta(t - n\Delta T), \tag{3.16}$$

so that it resembles an infinite series of delta peaks that go through zero and are spaced with steps of size ΔT. The Fourier transform of this Dirac comb function, in the sense of distributions, is given by

$$\begin{aligned}
\widehat{\text{Ш}}_{\Delta T}(\omega) &= \int_{-\infty}^{\infty} \text{Ш}_{\Delta T}(t) e^{i\omega t} \mathrm{d}t = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(t - n\Delta T) e^{i\omega t} \mathrm{d}t \\
&= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t - n\Delta T) e^{i\omega t} \mathrm{d}t = \sum_{n=-\infty}^{\infty} e^{i\omega n \Delta T} \\
&\text{taking the limit to infinity of the Dirichlet kernel we get} \\
&= \frac{1}{\Delta T} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{\Delta T}\right) = \frac{1}{\Delta T} \sum_{n=-\infty}^{\infty} \delta\left(\frac{\omega}{2\pi} - \frac{n}{\Delta T}\right) \\
&= \frac{1}{\Delta T} \text{Ш}_{\frac{1}{\Delta T}}\left(\frac{\omega}{2\pi}\right).
\end{aligned} \tag{3.17}$$

Firstly, we look at the effect of discreteness on our original infinite signal, so that we can analyse the result without the presence of spectral leakage.

Example 15. Using Lemma 11, our original infinite signal $g(t)$ from Example 8 and the discrete window given by the Dirac comb function $\text{Ш}_{\Delta T}(t)$ from Example 14, we find for the discrete infinite signal $s(t) := g(t)\text{Ш}_{\Delta T}(t)$ that

$$\begin{aligned}
&\mathcal{F}(s(t)) = \mathcal{F}(g(t)\text{Ш}_{\Delta T}(t)) = \mathcal{F}(g(t)) \star \mathcal{F}(\text{Ш}_{\Delta T}(t)) \\
&= \int_{-\infty}^{\infty} \hat{g}(x)\hat{\text{Ш}}_{\Delta T}(\omega - x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(2\pi C\delta(x) + \frac{A\pi}{i}(\delta(x+B) - \delta(x-B))\right) \frac{1}{\Delta T}\text{Ш}_{\frac{1}{\Delta T}}\left(\frac{\omega - x}{2\pi}\right)\mathrm{d}x \\
&= \frac{1}{\Delta T}\left(2\pi C\text{Ш}_{\frac{1}{\Delta T}}\left(\frac{\omega}{2\pi}\right) + \frac{A\pi}{i}\left(\text{Ш}_{\frac{1}{\Delta T}}\left(\frac{\omega + B}{2\pi}\right) - \text{Ш}_{\frac{1}{\Delta T}}\left(\frac{\omega - B}{2\pi}\right)\right)\right).
\end{aligned} \tag{3.18}$$

Instead of just peaks at frequencies 0, B and $-B$ we now find three infinite series of peaks that each cause a peak at one of the frequencies 0, B and $-B$ and are linearly spaced with steps of size $\frac{2\pi}{\Delta T}$. The heights of the peaks are the same within each infinite series of peaks. Therefore, it is impossible to distinguish between for example the peak at B and the peak at $B + \frac{2\pi}{\Delta T}$. This is caused by the fact that we do not have any information about the signal between the discrete measurements, which results in these frequencies to match the measured signal equally well. Thus, we would need more information about the signal, by for example taking more measurements of the signal at different time steps, in order to be able to distinguish between these frequencies. This effect of getting multiple (infinitely many) peaks in the periodogram caused by just one frequency that is present in the underlying signal is called *aliasing*. The peaks at frequencies that are not the actual frequency of the underlying signal are called aliases of the original peak. An example of aliasing caused by the uniform discreteness of the signal is shown in Figure 3.2.

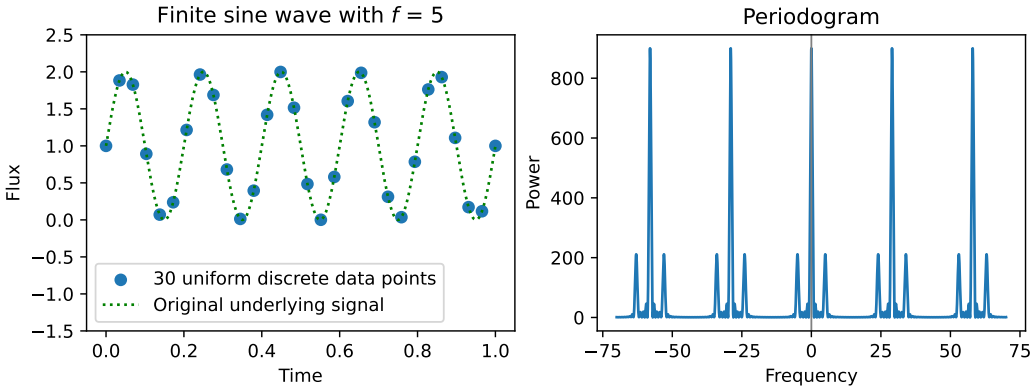

Figure 3.2: *An example of the effect of uniform discreteness of a finite sine wave signal (left) on the corresponding periodogram (right), resulting in aliasing. There are peaks at the original frequencies of the signal: at 0 and* ± 5*. These three peaks are repeated periodically with a step size of* $\frac{1}{\Delta T} = \frac{1}{\frac{T}{\#\,Data\ points-1}} = \frac{\#\,Data\ points-1}{T} = \frac{30-1}{1} = 29$ *in between. The finiteness of the signal results in the nonzero width of the peaks and the side lobes in between, as in Figure 3.1. (Note that the used frequency is not the angular frequency* ω *but* $f = \frac{\omega}{2\pi}$*.)*

Now that we have analysed the effect of discreteness on the original infinite signal and we know what the effect is of the convolution of the Fourier transform of a finite rectangular window with a sum of delta functions, we can derive what the Fourier transform of the resulting uniform discrete finite sine signal $s(t) := g(t)\text{Ш}_{\Delta T}(t)W(t)$ is: three infinite series of alias peaks around the frequencies 0, B and $-B$ with step sizes of $\frac{2\pi}{\Delta T}$, where each individual peak has the form of a sinc function with full width equal to $\frac{1}{T}$ at half maximum. Note that the step size of $\frac{2\pi}{\Delta T}$ is true when considering the frequency spectrum as function of angular frequency ω. When analysing the spectrum as function of frequency $f = \frac{\omega}{2\pi}$ we find the corresponding step size to be $\frac{1}{\Delta T}$. All described above is also illustrated in Figure 3.2.

In conclusion, we find that the Fourier transform of the uniform discrete finite signal results in a periodogram that is, although not perfect to retrieve the main frequency, well understood. The effects of aliasing and spectral leakage cause the periodogram to show peaks at frequencies that are not present in the original underlying signal, but this is unavoidable because of the missing information about the signal between the uniform discrete measurements.

Our data, however, does not consist of uniform discrete data, but of nonuniform discrete data. Therefore, the next section is dedicated to the effects that this change imposes on the resulting Fourier transform.

3.3 Nonuniform discrete Fourier transform

We have now considered two types of observation windows: a rectangular window, describing a continuous signal being measured continuously throughout a finite time interval, and a finite Dirac comb window (which can be regarded as an infinite comb combined with a rectangular window), describing a uniform discrete signal being measured throughout a finite time interval. In the case of a nonuniform discrete finite signal, we can think of the observation window as a sequence of nonuniformly spaced delta functions, peaking at the times at which the signal is measured.

Mathematically, if $N \in \mathbb{N}$ is the number of observations and $\{T_n\}$ for $n = 1, \ldots, N$ are the observation times, then we can describe the window function as

$$W_{\{T_n\}}(t) := \sum_{n=1}^{N} \delta(t - T_n). \tag{3.19}$$

Again, if we now want to compute the Fourier transform of the observed signal $s(t) := g(t)W_{\{T_n\}}(t)$ we would compute the Fourier transform of the window function $W_{\{T_n\}}(t)$ and of the original underlying sine wave signal $g(t)$ and convolve these to find the Fourier transform of $s(t)$. However, the analytical computation of $\mathcal{F}(W_{\{T_n\}})$ has now become more complicated because the symmetry of the Dirac comb function is broken and the analytical computation of $\mathcal{F}(s(t))$ will quickly become very tedious.

Therefore, we illustrate the effect of nonuniformity on the resulting periodogram by an example shown in Figure 3.3. In the upper figures we see that the periodogram of the nonuniform discrete finite sine wave becomes very 'noisy' with respect to the periodograms found in Figures 3.1 and 3.2. With noisy we mean that there are many seemingly random peaks present in the periodogram. When analysing the periodogram of the observation window, as shown in the lower figures, we see that the same kind of random noise is present here. In general, this is what we would expect: the randomness of the random observation times should in some way impose randomness on the resulting Fourier transform and thus on the periodogram.

Furthermore, we see in the periodogram in the upper right figure that the structured aliasing of the peaks at the original frequencies from the uniform discrete case has completely disappeared. This is also what we expected, as this was imposed by the regularity of the delta peaks in the Dirac comb function, causing the Fourier transform of the window function to be a regular comb function itself. In this case, this results in the possibility to retrieve the frequency of the underlying signal with certainty:

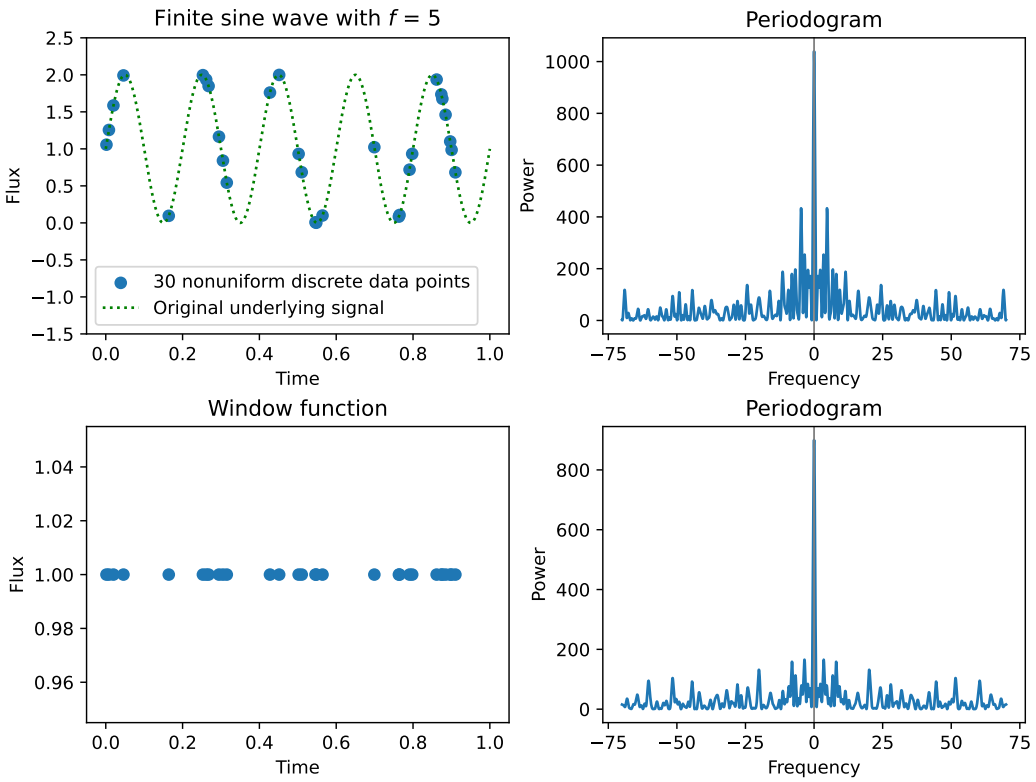

Figure 3.3: *An example of a nonuniform discrete finite sine wave signal (upper left) and its corresponding periodogram (upper right). The observation times are chosen randomly between 0 and 1. There are significant peaks visible at the frequencies 0, 5 and -5, but there are also a lot of noisy peaks at other frequencies. The observation window corresponding to this data set and its periodogram are shown in the lower figures. Noise similar to the noise in the upper periodogram is visible here.*

there are two significant peaks at frequencies 5 and -5 (ignoring the one at 0) and no other peaks that reach powers similar to theirs.

One might wonder how this is possible, as we still have the problem of missing information between the discrete observations. However, with the uniform sampling, the problem was actually the missing information about specific parts of the phase of the signal, allowing sine waves with alias frequencies to also fit the data. With the random nonuniform sampling this problem disappears, as all parts of the phase of the original signal can now be observed by a random observation. In other words, a nonuniform discrete sampling in general catches more information about the underlying signal than a uniform discrete sampling, which systematically skips parts of the signal. Therefore, using a nonuniform sampling can

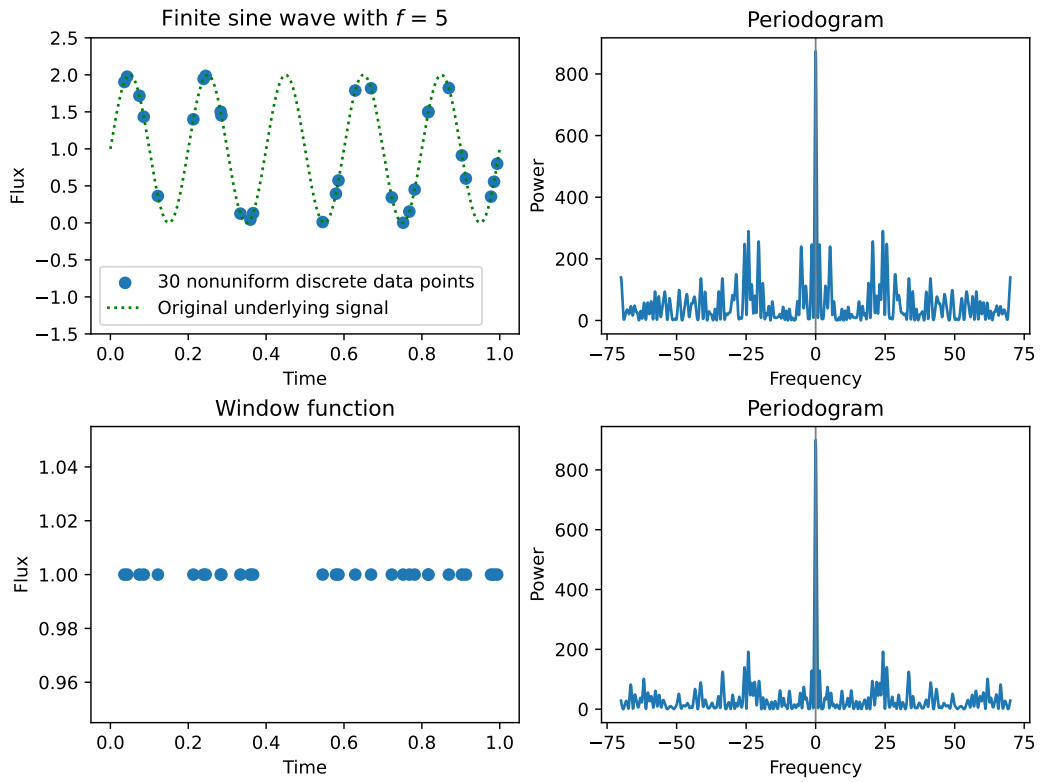

Figure 3.4: *An example of a nonuniform discrete finite sine wave signal (upper left) and its corresponding periodogram (upper right). The observation times are chosen randomly between 0 and 1. The observation window corresponding to this data set and its periodogram are shown in the lower figures. The noise in the periodogram caused by the observation window results in peaks with powers higher than the powers of the peaks at the main frequencies* ± 5.

be a good way to increase the possibility of retrieving the main frequency of the underlying signal, but it also imposes noise on the periodogram which can overshadow the main peak. An example of this is given in Figure 3.4, where the peaks around ± 25 have powers higher than the powers of the main frequencies at ± 5. The periodogram of the window function also shows peaks with higher powers around frequency ± 25, but these do not have the exact same form as those in the periodogram of the sampled signal as they are convolved with the Fourier transform of the underlying signal.

In conclusion, the understanding of the periodogram of a nonuniform discrete finite sine wave signal is much less straightforward than for a uniform discrete signal. This is caused by the irregularities within the time sampling which impose noise on the resulting periodogram. This

noise can result in peaks with powers higher than the power of the main frequency, making it more difficult to retrieve the correct frequency. However, there is no systematic aliasing as for the uniform discrete case, which does make it possible to retrieve the main frequency with certainty if the noise is low enough.

We have now discussed random nonuniform discrete samples. In reality, it is difficult to achieve pure randomness, as there are often some regularities in the way observations are made. For Gaia data we have already discussed in Section 2.2 that the time series seem random, but do actually contain regularities that are caused by the scanning law. We found that this causes spurious periods to appear more often with maximal power in the resulting periodograms. This can be explained by the effect of the semi-random time series, imposing semi-random noise on the periodogram which results in higher noise peaks at the spurious periods. As the form of the noise depends on the observation times, it is possible to identify spurious period candidates caused by the window function before analysing the signal.

3.4 Deconvolution

One might think, after all this talk of convolutions, that it could be possible to use some sort of inverse convolution, or deconvolution, from the Fourier transform of the observed signal with the Fourier transform of the observation window in order to recover the true periodogram of the underlying signal. However, because the window function is typically equal to zero for large parts of its domain, especially in the case of discrete sampling, the inverse of the convolution is not well defined. Thus, any sort of deconvolution is no solution to our problem and we have to deal with the periodogram of the observed signal.

Chapter 4

Methods

In this chapter we elaborate on the methods that we used to investigate which frequency in the NUFFT-periodogram corresponds to the true period of the underlying signal. Before we can investigate the periodogram, we of course have to compute it. Therefore, we explain in the first section how this is done. In the second section we clarify how we have simulated the data that is later used by the research methods. These methods, which consist of taking the frequency with maximal power and examining whether this is the true frequency and later determining the false alarm probability of these frequencies, are discussed in sections three and four.

4.1 Periodogram computation

In this research, the NUFFT-periodogram of the data is computed using the FINUFFT Python package. This package includes multiple functions to handle nonuniform data and its Fourier transform. In [5] they explain what each of the implemented functions does.

In this research we use their Type 3 transform, which is specifically designed to evaluate the Fourier transform of nonuniformly spaced data at arbitrary (nonuniform) target frequencies. The corresponding function is as follows `finufft.nufft1d3(x, c, s)`, where the 1d stands for the 1 dimensional nature of the data and the 3 refers to the Type 3 transform. The parameter x is called the 'nonuniform source points', which in our case are the times at which the flux is measured. Parameter c is called the 'source strengths', which are the corresponding measured fluxes. As we saw in Section 3.1 a constant shift of the fluxes of the signal causes a peak at frequency zero, which can result in having the highest power in the peri-

odogram, while this is not the frequency corresponding to the underlying signal. Therefore, we subtract the mean flux from all measured fluxes and use the result for c. Lastly, parameter s denotes the angular frequencies ω at which we want to evaluate the Fourier transform (recall that $f = 2\pi\omega$). These Fourier transform values are returned in an array with size equal to that of s. As the Fourier transform is defined for functions $g : \mathbb{C} \to \mathbb{C}$, both c and the returned Fourier transform values are complex valued. We turn the Fourier transform values into periodogram powers by taking the absolute values and squaring these.

4.1.1 Evaluation frequencies

We choose the evaluation frequencies f of s such that they are linearly spaced between 0 and 100 [1/d], because these are the frequencies that could realistically be reached by signals of variable sources and are still detectable by the Gaia measurements. Note that there are variables with timescales of minutes or even seconds, meaning that we would also have to look at frequencies of tens of thousands per day. However, as the shortest possible sample interval for Gaia data is 106.5 minutes, these short periods are very difficult to retrieve anyway, so we focus on longer timescales. For the spacing between two evaluation frequencies, we use an approach similar to that explained in [3]. As we saw in Chapter 3, the width of the periodogram sinc peaks becomes $\frac{1}{T}$, because of the rectangular window with width T. As we do not want to miss any of the peaks in the periodogram, the step width between two evaluation frequencies is chosen as $\Delta f = \frac{1}{n_0 T}$, where $n_0 = 10$ and T is the time between the first and the last observation. In this way, every peak is covered by about n_0 evaluations.

4.2 Simulated photometric data

In order to investigate whether a frequency that we find from a periodogram is actually the frequency that we are looking for, we need to know the original frequency underlying the signal. To achieve this, we simulate our own photometric data in such a way that it resembles real photometric Gaia data.

4.2.1 Time series

The time samplings for the simulation are taken from real Gaia data. As explained in Section 2.2, the Gaia time series are nonuniform, but do in-

clude some specific regularities caused by the Gaia scanning law. Furthermore, the number of time samplings and their spacing are also influenced by the position of the source on the sky, as explained in Section 2.2 and shown in Figure 2.6. As the number of samplings and their spacing might influence the periodogram and therefore the ease with which the right frequency is found, we do not want to bias our research to certain fixed values of these parameters. Therefore, we chose to take actual time samplings from Gaia data, from 35 sources with different ecliptic coordinates, reaching from longitude 0 to 90 and latitude 0 to 60 with steps of about 15 in between. In this way, we try to reach many different kinds of time samplings of Gaia data, with different numbers of data points and visibility periods. Figure 2.6 was used to help choose different longitudes and latitudes that would result in a diverse set of time series. Furthermore, the sources were selected on having photometric data available (has_epoch_photometry = 'True') and on being actual variable sources (phot_variable_flag = 'VARIABLE'), such that the mean flux and amplitude of their signal could be used as realistic values for these parameters.

4.2.2 Original signal

To construct photometric time series of variable sources, we start by simulating the original signal that would come from this variable source. As we assume in this research that all variable source candidates have signals consisting of one main frequency, we choose to model the original signal as a pure sine wave. This method is similar to the approach used in [16]. We denote the amplitude of the sine function by A, its frequency by f and its equilibrium value by C, which results in the function given by

$$g(t) = C + A\sin(2\pi f t). \tag{4.1}$$

Whether a frequency can easily be found from the periodogram might depend on the frequency itself. For example, low frequencies can result in not even one full phase being present in the data, while high frequencies can easily be undersampled. Therefore, we chose to use 20 different frequencies for our simulations, ranging uniformly between 0.002 and 65.34 1/d. We choose to take the simulation amplitude and equilibrium value from one of the variable sources that we also used to take the time series from. In this way, we assured that these values are realistic for variable sources. Furthermore, by keeping these values the same for all simulated data sets, we can more easily compare the results and ensure that the signal to noise ratio, which we define by the amplitude divided by the measurement uncertainty on one flux measurement, does not influence the

results, as this is now constant. The equilibrium value is thus given by the mean of all fluxes of the chosen source and the amplitude is given by the maximum flux for this source minus the found equilibrium value. This results in the parameter values given in Table 4.1.

Parameter	Value	Unit
C	5194	electron/s
A	123	electron/s
f	20 linearly spaced between 0.002 and 65.34	1/d
σ_F	25	electron/s
S/N	5	-

Table 4.1: *The parameter values used for all simulated light curves. The equilibrium value of the sine wave of Equation* (4.1) *is denoted by C, the amplitude by A, the frequencies by f. The standard deviation used for all simulated data points is denoted by σ_F and the resulting signal to noise given by $\frac{A}{\sigma_F}$ is denoted as S/N.*

4.2.3 Flux

The simulated original signal and time series are combined to find the corresponding fluxes. The flux that would be measured from the original signal at a certain measurement time from the time series is given by the sine function $g(t)$ from equation (4.1) evaluated at that measurement time. However, in reality there is always a measurement error on the flux. For Gaia data, this measurement error is mainly dependent on the flux itself. The python library PyGaia by Gaia DPAC includes the typical measurement errors on the magnitude, as a function of the actual magnitude and the photometric band. (In this research we only consider the G-band flux.) As we want our simulation data to represent Gaia data as realistically as possible, we add this error to our fluxes.

The measurement uncertainty on the flux is computed as follows:

Firstly, we determine the magnitude m corresponding to the flux F using the expression

$$m = -2.5 \log_{10}(F) + z \tag{4.2}$$

where $z = 25.6873668671$ is the (Vega) magnitude zero point in the G-band for Gaia DR3. Secondly we use the PyGaia function `magunc` to find the magnitude uncertainty σ_m for this magnitude m in the G-band. This magnitude uncertainty is defined as the uncertainty on the average magnitude, given in mmag. Therefore, we have to convert this to mag by division by 1000 and we also have to convert it to the magnitude uncertainty

per data point, which is done by multiplication by $\sqrt{N}$ with N the number of data points. This uncertainty is then converted to the flux uncertainty σ_F by error propagation, giving that

$$\sigma_m = \sqrt{\left(\frac{\partial m}{\partial F}\sigma_F\right)^2} = \left|\frac{-2.5}{\ln(10)F}\sigma_F\right| = \frac{2.5}{\ln(10)}\frac{\sigma_F}{F}, \tag{4.3}$$

so

$$\sigma_F = \sigma_m F\frac{\ln(10)}{2.5}. \tag{4.4}$$

This uncertainty σ_F can now be used to simulate the observed flux F_{obs} by assuming the flux on each of the data points follows a Gaussian distribution with mean F and standard deviation σ_F. Therefore, we get a realistic F_{obs} by drawing a random sample from such a Gaussian distribution.

We could compute a separate standard deviation for each of the data points to compute the final flux for the data points. However, as the different fluxes vary only quite little (given that the equilibrium of the simulated sine wave signal is much bigger than the amplitude), the resulting σ_F would also vary only a little bit. As we then use this σ_F to compute a random value from a Gaussian distribution to compute the resulting F_{obs}, the effect of this small difference in σ_F really becomes quite negligible for the fluxes. Therefore, we decided to compute σ_F once for all data points of a certain signal, by using the σ_F on the equilibrium flux of the sine wave. It is important to note here again that σ_F depends on the number of data points as well, so the σ_F would in reality differ per time series. However, in order to keep the signal to noise S/N constant for all data sets, we chose to compute the σ_F for one data set and use this for all others as well. Both the resulting values of this σ_F and S/N are shown in Table 4.1.

4.3 Maximal power frequency correctness

As explained in Chapter 3, from the interpretation of the periodogram we generally expect the frequency with maximal power to be the main frequency of the underlying signal that we are looking for. However, when considering discrete signals we found that this is not always the case, because of aliasing, caused by the fact that we do not have any information about the signal between the measurements. We also found in Section 2.2 that there are many spurious periods occurring more often with maximal power in the periodograms of Gaia DR3 time series. Therefore, we are

interested in to what extent this is a problem and whether we can distinguish between situations in which this method gives us either the correct or an incorrect frequency.

Now that we have 700 simulated light curves (35 time series $\times$ 20 frequencies), we compute the periodogram for each using the method described in Section 4.1 and determine which frequency is the frequency with maximal power. Thereafter, we look into which determined frequencies are equal to the simulated frequencies. The periodogram is evaluated at a finite number of frequencies, as explained in Section 4.1.1, with steps of $\Delta f = \frac{1}{10T}$ where T is about 950 days. So we find that $\Delta f = \frac{1}{9500} \approx 0.0001$ 1/d. Therefore, a difference between the simulated frequency and determined frequency of about 0.0001 1/d can be explained by this step size and thus any determined frequencies that lie within the range 'simulated frequency ± 0.0001' are labelled as correctly determined. The other determined frequencies are labelled as being incorrect.

Note that the difference between the determined frequency and simulated frequency for incorrect determinations is not actually of interest, as a completely different peak has maximal power in the periodogram, so this difference does not give us any information about the error on the simulated frequency. Thus we are only interested in whether the determined frequency is correct or incorrect.

4.4 False alarm probability computation

After computing the frequency with maximal power for the periodograms of the simulated light curves and analysing which are correctly and incorrectly determined, we try to find a way to decide which frequencies might be incorrectly determined.

To this end, we tested the idea of computing a false alarm probability (FAP) as function of the frequency and the power of this frequency in the periodogram of the data. This is done by taking bootstrap samples of the data (explained in Section 4.4.1) and computing the periodogram power for the given frequency for each of these bootstrap samples, after which a cumulative distribution function (CDF) of the random variable Z describing the found powers is formed. The resulting false alarm probability for the frequency f with power z in the original periodogram is then given by $\text{FAP}(z) = \mathbb{P}(Z \geq z) = 1 - \mathbb{P}(Z \leq z)$. In other words, the FAP is given by the number of found bootstrap powers that are greater than the power in the original periodogram divided by the total number of bootstrap samples.

This value describes the probability that the power at the given frequency can be reached by similar data without the presence of an underlying signal. Thus, when this value is relatively high for an incorrectly determined frequency (where the actual range of 'relatively high' is still to be determined), this could be an indication of the incorrectness of the determined frequency, because the observed power has a relatively high probability of being caused by factors other than the underlying signal.

Note that this computation and interpretation of a false alarm probability is different from the Baluev FAP used for Figure 2.7 and also different from other FAPs from literature [3] which are similar to the Baluev FAP. Recall that the Baluev FAP denotes the probability that the given frequency occurs with maximal power in random samples without underlying signal. Thus, an example of the computation of the Baluev FAP is by using bootstrap samples similar to the ones used in this research and considering the distribution of the maximal power frequencies in the corresponding periodograms.

4.4.1 Bootstrap method

To compute the FAP for a certain frequency and power, we use a bootstrap method with 500 samples. Each of these bootstrap samples is constructed as follows: The time series from the original data are kept the same and a permutation of the flux measurements is taken. Thus, for each of the measurement times one of the flux measurements is taken randomly and with repetition (meaning that a certain flux measurement can be chosen multiple times for different measurement times). In this way, the signal underlying the data is removed while the describing parameters of the data set, like its equilibrium value, amplitude and window function (time series) remain the same. This ensures that the scaling of the powers in the resulting bootstrap periodograms is similar to that of the periodogram of the original data.

Version of July 6, 2023– Created July 6, 2023 - 19:17

Chapter 5

Results

5.1 Maximal power frequency analysis

In this section we discuss the results of the method from Section 4.3 applied to the simulated data sets described in Section 4.2, in which the frequency with maximal power is retrieved from the periodograms of the data as this frequency should have a high probability of being the one corresponding to the underlying signal that we are looking for. An example of a simulated data set with its corresponding periodogram is given in Figure 5.1. The determined and simulated frequencies are also indicated here and it is clear that the determined frequency is incorrect.

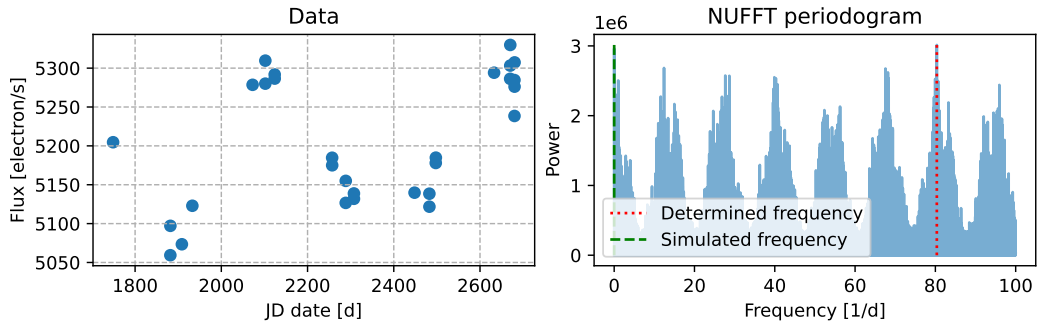

Figure 5.1: *Left: Simulated light curve with simulated frequency of 0.002 1/d. Right: Periodogram corresponding to the simulated data set. The red dotted line marks the determined frequency with maximal power of about 80.36 1/d and the green dashed line marks the simulated frequency of 0.002 1/d. This determined frequency is thus incorrect.*

Using the method on all simulated data sets results in the differences

between the simulated and determined frequencies shown in Figure 5.2.

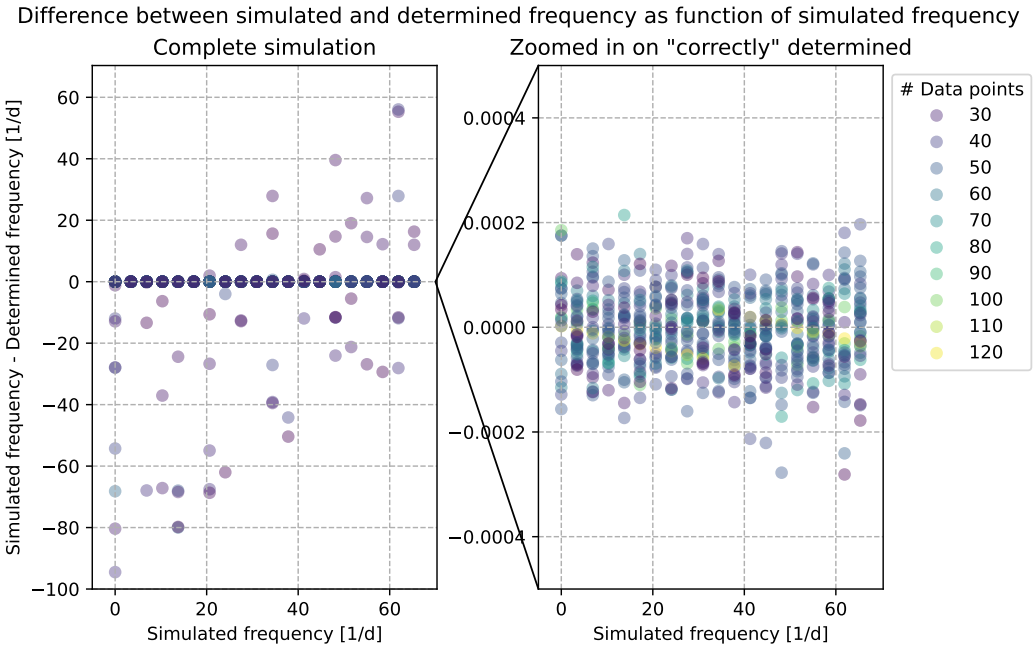

Figure 5.2: *Left: Difference between simulated and determined frequency as function of simulated frequency for the complete simulation. This includes 35 different time series and 20 simulated frequencies, so 700 data sets in total. Right: Zoomed in version of the figure on the left on the region around a difference of zero. The colours represent the number of data points, so flux measurements, for each data set.*

As we can see in the left panel of this figure, for every of the 20 simulated frequencies there are time series for which the determined frequency is similar to the simulated frequency, resulting in the difference of about zero. For most simulated frequencies there are also time series for which the determined frequency is different from the simulated frequency, resulting in all other dots with an absolute difference much greater than zero. Thus there is no simulated frequency for which all determinations are incorrect. Furthermore, we see that the simulated - determined difference is almost always negative for the low simulated frequencies, while it is more often positive for the high simulated frequencies. Besides, we see a lower and upper bound trend for the found differences. This is caused by the fact that the periodogram is evaluated at frequencies between 0 and 100 1/d, thus low simulated frequencies have a high chance of the determined frequency being found at frequencies higher than the simulated frequency, while high simulated frequencies have a higher chance of the determined

frequencies being lower than the simulated frequency, because more frequencies are evaluated in this range. The bounds are caused by the fact that the periodogram is not evaluated at frequencies outside the bounds: for simulated frequency of around 50 1/d, the absolute difference cannot be greater than 50 1/d, because the periodogram frequencies range between 0 and 100 1/d. Moreover, we see that the incorrect determinations all have colours within the range of about 30 to 50 data points. We will further investigate this later. Lastly, we see that the dots around a difference of 0 are very darkly coloured, indicating a dense region of dots.

Zooming in on this "correctly" determined region we see in the right panel that there are indeed a lot of data sets situated here, with a wider range of colours. The dots are clearly centred around zero with a spread reaching mainly between -0.0001 and +0.0001. However, we also see quite some dots in the regions towards ± 0.0002 and even some outliers towards ± 0.0003. These differences cannot be explained only by the step size of about 0.0001 between the evaluation frequencies in the periodogram, but they could be explained by a slight shift of the peak in the periodogram caused by the measurement errors on the fluxes. Considering also that the width of each peak is of the order of $\frac{1}{T} \approx 0.001$ 1/d, so that the range of dots around zero still falls within the range of the width of the peak (meaning that they do not correspond to a different peak at another frequency), we choose to regard all dots with a frequency difference between ± 0.0004 1/d as being correctly determined.

This results in 625 of the 700 determined frequencies being correct, thus about 90% of the determinations is correct. The incorrect determinations are 'completely' wrong in the sense that a completely different peak in the periodogram at a frequency that is not the simulated frequency is found having maximal power.

Now that we have decided which determined frequencies are correct and incorrect, we can analyse the situations in which this method gives us either correct or incorrect frequencies. In Figure 5.3 we see that the fraction of correctly determined frequencies does not show any clear correlation with the simulated frequencies.

We expect the probability to retrieve the right frequency from the periodogram to be higher the more information we have about the signal. Therefore, we expect the fraction of correctly determined frequencies to be higher for time series with more data points. As explained in Section 2.2, data points can be clustered which means they do not provide any new information about the signal. Therefore, the number of these clusters, called the visibility periods, is a better measure of the effect number of data points. To analyse any relations present, we firstly look at the re-

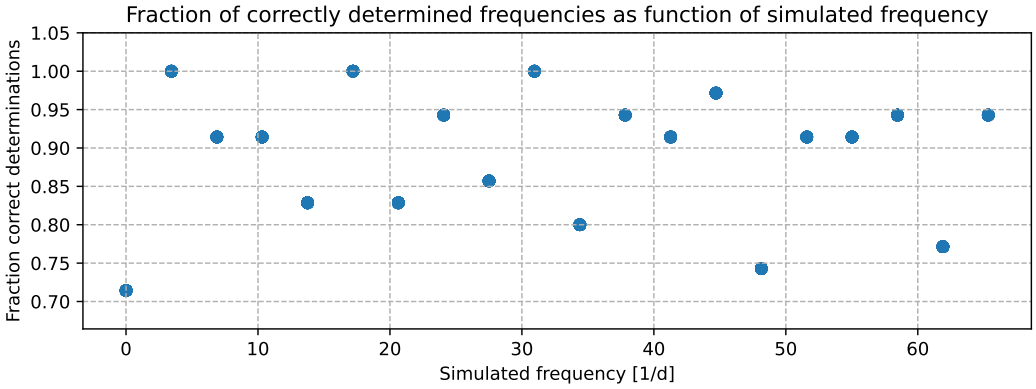

Figure 5.3: *The fraction of correctly determined frequencies is shown for each simulated frequency. Note that the y-axis does not start at zero.*

lation between the number of data points and the number of visibility periods. We see in Figure 5.4 that there is a correlation between the number of data points and the number of visibility periods: generally we see that the more data points, the more visibility periods. However, when looking closely we see that this correlation is mainly true for low visibility periods and low numbers of data points, while the correlation lessens for higher numbers of data points. This is to be expected as high numbers of data points can also specifically be caused by quick successive measurements causing for more clusters of measurements and thus less visibility periods. Note also that there are many visibility periods that come back more often for different numbers of data points. As this correlation between the number of data points and the number of visibility periods is not perfect, we will take them both into account when analysing the relation with the fraction of correctly determined frequencies.

In Figure 5.5 we see that there is a correlation between the number of data points and the fraction of correctly determined frequencies: the more data points the higher the correct fraction, as expected. We do see, however, that there is a slight spread present, so the relation is not perfect. Furthermore, we see that from about 60 data points and higher, all frequencies are correctly determined. Lastly, we also see in this figure that for low numbers of data points the number of visibility periods is also relatively low, while for higher numbers of data points the number of visibility periods can be either high or low.

Looking at Figure 5.6 we see a very similar relation between the number of visibility periods and the correct fraction. As there are multiple time series that have the same number of visibility periods, we have not

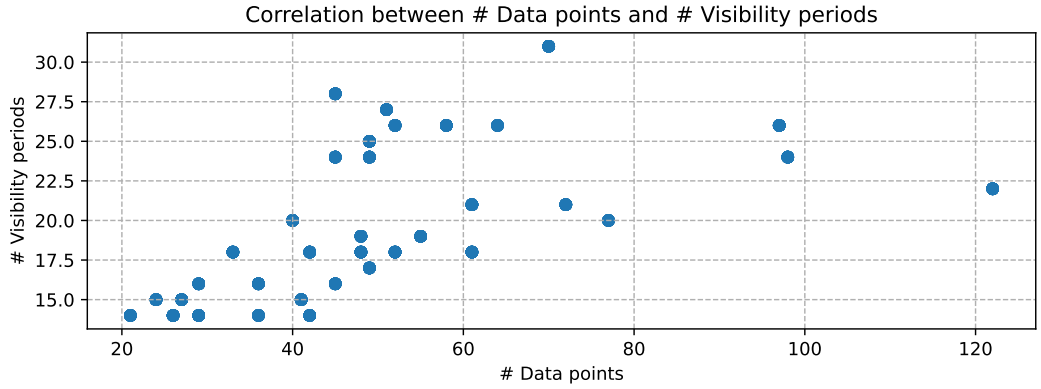

Figure 5.4: *The number of visibility periods is plotted as function of the number of data points for the time series used in the simulation.*

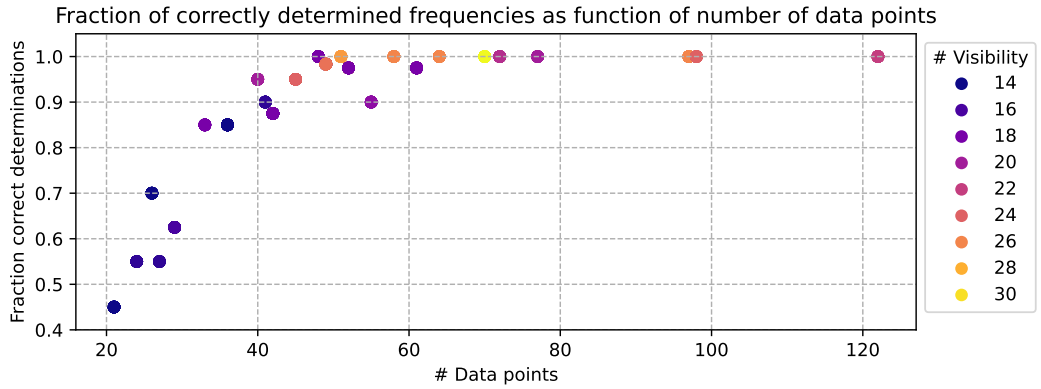

Figure 5.5: *The fraction of correctly determined frequencies as function of the number of data points within the 35 time series is shown. The colours depict the number of visibility periods for these time series. Note that the y-axis does not start at zero.*

colour coded the dots by number of data points, as multiple numbers of data points can correspond to a certain number of visibility periods. This is also why there are less dots present in this figure then there are in Figure 5.5. The relation seems to be somewhat neater, so with less spread, than for the number of data points. This, however, can be a visual effect caused by the lower number of dots. On the other hand, almost all possible numbers of visibility periods from 14 to 28 are represented. For 21 visibility periods or more all frequency determinations are correct.

Now that we have some idea about when the determinations are mostly correct, we look at what happens when the determination is incorrect. In Figure 5.7 the distribution of incorrectly determined frequencies is shown.

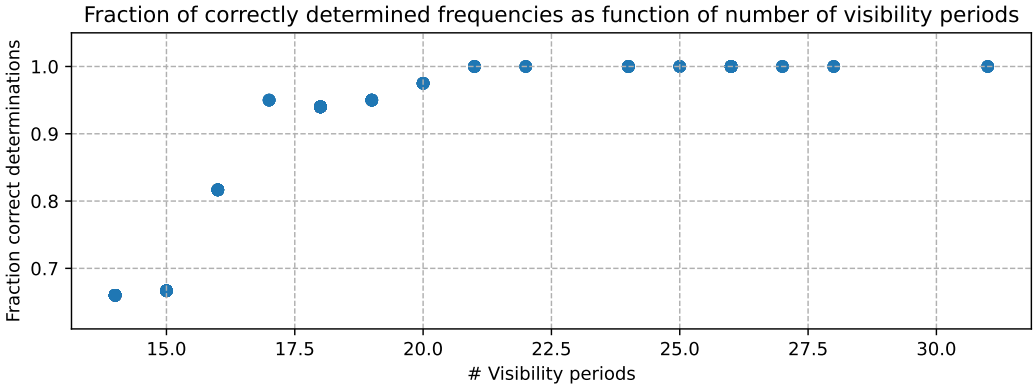

Figure 5.6: *The fraction of correctly determined frequencies as function of the number of visibility periods is shown. Note that the y-axis does not start at zero.*

There are some ranges of frequencies with higher peaks, indicating that multiple incorrectly determined frequencies lie within that range.

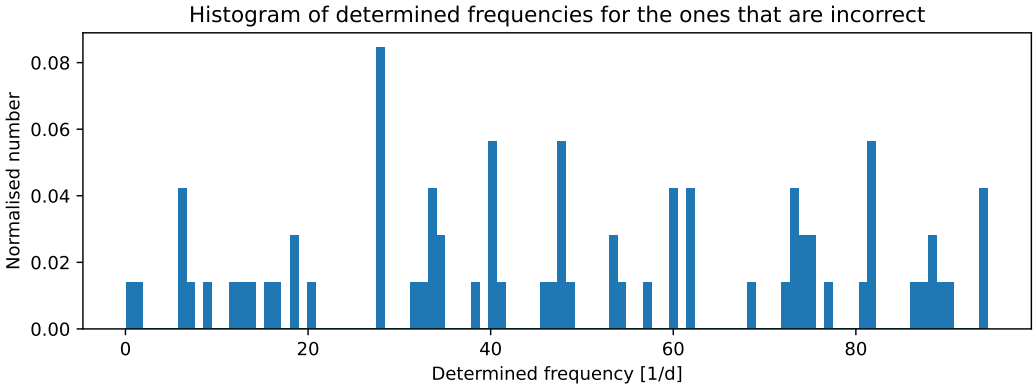

Figure 5.7: *Histogram showing the distribution of incorrectly determined frequencies. There are 100 bins, each with width equal to 1.*

5.2 FAP for simulated and determined frequency

We compute a false alarm probability for all determined and simulated frequencies using the method described in Section 4.4. In this section we analyse the corresponding results.

In Figure 5.8, 3 of the 500 bootstrap samples and their corresponding periodograms used to compute the CDFs of the determined and simulated frequencies of the simulated data set shown in Figure 5.1 are shown.

The examples of the bootstrap samples in the left panels illustrate that the bootstrap samples indeed all have the same time series with a random permutation (with repetition) of the flux measurements of the original data set from Figure 5.1. The corresponding periodograms in the right panels illustrate the diversity of possible periodograms for random data (without signal) with the same time series. While the periodogram from bootstrap sample 2 seems to consist mainly of noise without any clearly recognisable peaks, as one would expect from random data, the periodograms from bootstrap samples 1 and 3 contain more structure with recurring patterns of higher and lower power regions. Furthermore, the maximal power in the periodogram is much higher for periodograms 1 and 3 with respect to periodogram 2. It seems that the more structure there is in the periodogram (and thus also in the random data set), the higher the powers in the periodogram. Moreover, the incorrectly determined frequency from the original data set from Figure 5.1, indicated with the red dotted lines, also gives a peak in periodogram 3 (although not as high as other peaks), but it lies between peaks in periodograms 1 and 2. Lastly, when we compare the bootstrap periodograms with the periodogram from the original data set in Figure 5.1, we see that the powers of the peaks in the original periodogram are much higher than the powers of the peaks in the bootstrap periodograms. However, analysing the periodicity and overall structure of the periodograms it is remarkable how much bootstrap periodogram 3 resembles the original periodogram, with the same width and number of peak regions (7 from frequency 0 until 80) and both a peak at the determined frequency.

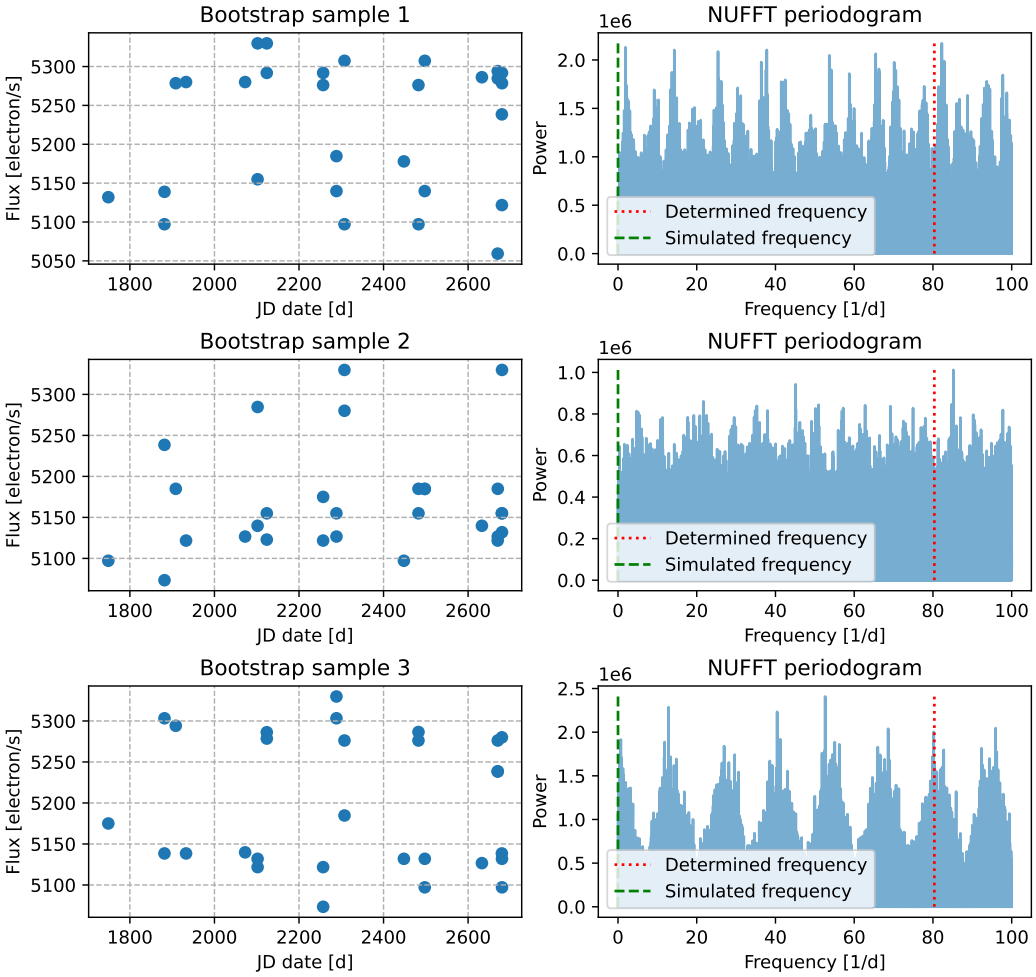

Figure 5.8: *Left: Three bootstrap samples for the computation of the FAP of the simulated and determined frequency of the simulated light curve of Figure 5.1. Right: Periodograms corresponding to each of the bootstrap samples. The lines indicate the frequencies at which the periodogram is evaluated to find the powers with which the CDFs for these frequencies are computed. The red dotted line marks the determined frequency of the original light curve of about 80.36 1/d and the green dashed line marks the simulated frequency of the original light curve of 0.002 1/d.*

Combining the powers in the periodograms of all 500 bootstrap samples, corresponding to the original data set of Figure 5.1, at the determined frequency of 80.36 1/d and the simulated frequency 0.002 1/d, we get the CDFs shown in Figure 5.9. In both panels of this figure it is clear that the power at the given frequency in the original periodogram is much higher than the powers at this frequency in the periodograms of any of the boot-

strap samples. Thus, the resulting FAP for both the simulated and determined frequency of the simulated light curve of Figure 5.1 is equal to zero. (The probability that the power at the given frequency in the original periodogram is reached in the periodograms of bootstrap samples without any signal is zero.)

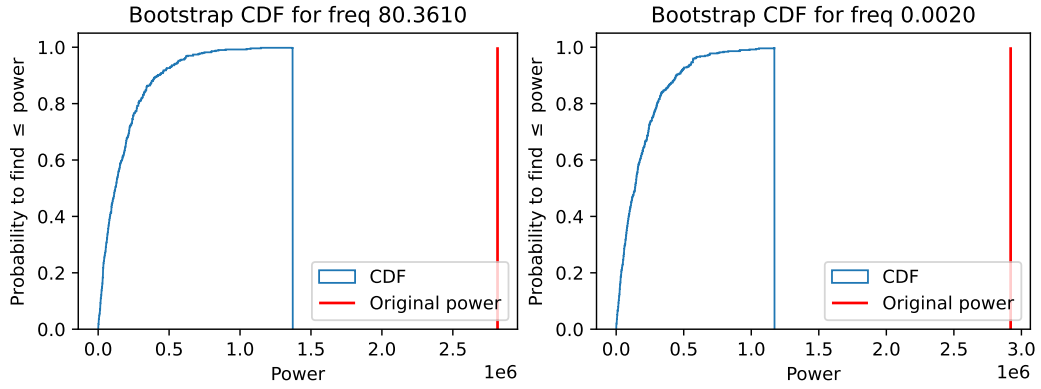

Figure 5.9: *Left: In blue the CDF for the determined frequency of 80.36 1/d of the simulated light curve of Figure 5.1 is shown (this is computed with the bootstrap method from Section 4.4.1). The red line indicates the power of this frequency in the periodogram of the original light curve. Right: Same picture as on the left but now for the simulated frequency of 0.002 1/d.*

When computing the FAP for all simulated and determined frequencies of all simulated light curves, we find that only a few of the simulated frequencies have a FAP greater than zero and that none of the determined frequencies have a FAP greater than zero, as shown in Figure 5.10.

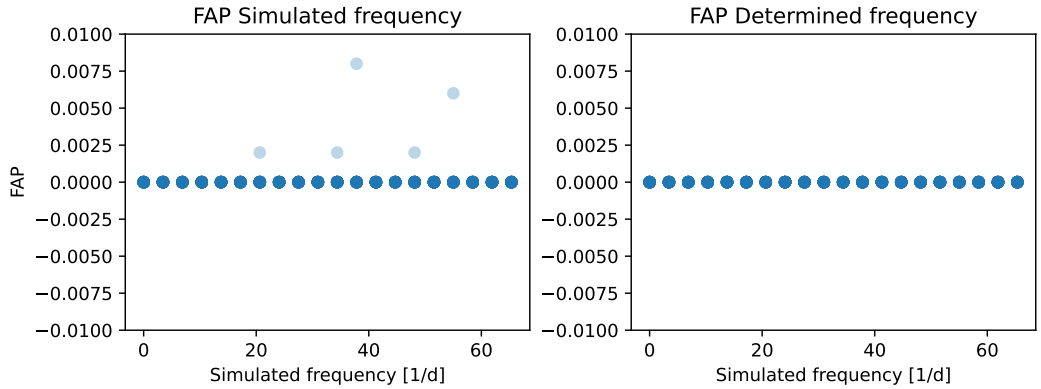

Figure 5.10: *Computed false alarm probabilities for all simulated (left) and determined (right) frequencies as a function of the simulated frequency.*

Chapter 6

Discussion

In this chapter we discuss our findings. In the first three sections the results and some limitations of the simulation and used methods are discussed. In the fourth section general limitations for period determinations of Gaia data that we found are pointed out. Lastly, the fifth section is dedicated to ideas for further research in the direction of main frequency search with the use of the NUFFT periodogram.

6.1 Simulated data

6.1.1 Limitations

In our simulations we simulate light curves by sine waves. Although many realistic light curves do look much like sine waves or sine waves with slight alterations, as we have seen in Section 2.1, there are also many periodic signals that cannot realistically be resembled by a sine wave, like the light curves of eclipsing systems. Therefore, the results from the simulations cannot be translated directly to apply on actual Gaia data.

Furthermore, only very specific values for many of the simulation parameters have been chosen, like the fixed equilibrium value, amplitude, derived flux error and resulting signal to noise. This approach is chosen to ensure that these parameters do not influence the results, so that the focus lies on the differences in simulated frequency and numbers of data points and visibility periods between the different simulated data sets. However, different values of the signal to noise, for example, likely affect the results, meaning that for a full understanding of the effectiveness of the methods, one would also need to analyse the results for different values of these

parameters.

6.1.2 Reproducibility

It is important to note that the reproducibility of the simulation is not perfect, because of the random flux errors on the data points within each simulated light curve. We noticed that when running the simulation multiple times, the results would slightly deviate each time. The figures shown in this report are all from one of the runs of the simulation, to make sure that they correspond to the same underlying simulated data. We have seen when running multiple simulations that the average results for the 700 simulated light curves would stay about the same (for example, at one of the simulations the number of correctly determined frequencies would be 625, while for a successive run of the simulation it would be 626). Therefore, we still regard those results as reliable. However, for individual simulated light curves we found that, while the time series and simulated frequency where the same, the resulting determined frequency could differ greatly caused by the different flux errors on the data points (so a completely different peak has maximal power). This shows that the random flux errors do have a significant affect on the periodogram. To analyse this effect one could analyse the distribution of determined frequencies for different runs (thus different random flux errors) of one simulated light curve.

6.2 Determined frequency analysis

In the maximal power frequency analysis in Section 5.1 we looked for situations in which the determined frequency is either correct or incorrect.

We found that in general about 90% of the determinations were correct. Thus for the specific parameters of our simulation the method of choosing the frequency with maximal power in the periodogram as the main frequency underlying the signal turned out to be quite good. However, it is important to note here that our simulation is quite specific, including only sine wave like light curves with very specific values for the amplitude, equilibrium value, frequencies, flux error and resulting signal to noise. The result can therefore not be regarded as generally true for sine wave light curves on which the NUFFT is used to find the main underlying frequency of the signal.

When analysing the fraction of correctly determined frequencies as a function of the simulated frequency we found that there was no correla-

tion between these variables. However, we have chosen our simulation frequencies as linearly spaced between two realistic values and have only used 20 different frequencies. Therefore, it is possible that there are specific simulation frequencies for which the fraction of correct determinations is either relatively high or low compared to the fraction for the average other simulated frequency. This could for example be true for simulated frequencies that lie close to a spurious period from Gaia DR3, causing them to be correctly determined more often. In order to investigate this hypothesis, one could use more simulation frequencies and choose simulation frequencies close to spurious periods. However, it is questionable how useful this information would be, as it would remain difficult to tell whether the determined frequency corresponds to the signal or to the time sampling.

The analyses of the fraction of correctly determined frequencies as function of the number of data points and visibility periods showed that there is a correlation here. In general we find that the more data points, the higher the fraction of correctly determined frequencies. The same is true for the correlation with visibility periods. This is as expected, as more data points and/or visibility periods means more information about the signal and thus a higher chance to recover its frequency correctly. We also saw that the correlation with number of data points showed a larger spread then the correlation with the number of visibility periods. This could be explained by the fact that there is a higher diversity of numbers of data points, causing this spread to be visible, while for the number of visibility periods more simulated samples are taken together in one point as they have the same number of visibility periods. For the correlation with number of visibility periods there are still some point that seem to be outliers, in the sense that they do not follow the trend of the other points, at 15 and 17 visibility periods. This spread could be caused by the distribution of the visibility periods over the observation time interval, which, as explained in Section 3.3, also influences the resulting periodogram. As this distribution can be different for all our used time series and is thus not fixed, this variable can result in the correlation with number of visibility periods to be spread out.

Lastly, we looked at the distribution of incorrectly determined frequencies. We found that there are some frequency ranges with width of 1 1/d within which multiple incorrectly determined frequencies are found. This could be a coincidence, but it could also be that these specific determined frequencies correspond to spurious periods of Gaia DR3, causing them to occur more often with maximal power in the periodograms. Whether this is the case has not been investigated yet, but this would be interesting to analyse, since this could give more information about whether the in-

correctly determined frequencies are spurious periods or aliases and thus about whether we can find a way to separate the correctly and incorrectly determined frequencies. If one would want to analyse this, however, it could be useful to use a larger simulation, so that there is a larger distribution of incorrectly determined frequencies to analyse (as there are only 75 cases of incorrectly determined frequencies in our simulation).

6.3 FAPs

In Section 5.2 we found that the computed false alarm probabilities turned out to be equal to 0 for almost all determined and simulated frequencies. Of course we already expected the FAP of the simulated frequency to always be relatively low, possibly 0, as removing the signal with the simulated frequency from the data should generally lower the power at the simulated frequency in the bootstrap samples with respect to the power at this frequency from the original signal, (the computation of this FAP was always more of a sanity check), but we did not expect this for the FAP of the determined frequency. We did not expect this beforehand, as we specifically used the bootstrap method in order to keep all other parameters describing the data the same, so that the values of the powers within the bootstrap periodogram would be similar to those in the periodogram of the original data set. We expected that we could then compare the powers in the bootstraps with the power in the original data and see what the probability is to find a certain power without the signal being present. There are a few possible explanations for this discrepancy.

Firstly, it could be that removing the signal from the data results in all powers in the bootstrap periodograms generally being lower than they are in the periodogram of data with the signal. We see that this is true for the three bootstrap sample examples given in Figure 5.8, even for example 3 which has a periodogram that looks very much like the periodogram of the original data set, except for the values of the powers, which are all lower for the bootstrap samples. We do not know, however, whether this is the case for all bootstrap samples for all simulated data sets. This could be investigated by, for example, computing the maximal power in each of the bootstrap periodograms and comparing these with the maximal power in the periodograms of the original data sets. If the maximal powers in the bootstrap periodograms are all lower than the maximal powers in the original periodograms, this hypothesis would be confirmed.

Secondly, it could be that the powers at the determined and simulated frequency specifically are lower in all bootstrap periodogram than

they are at the original periodograms, because the heights of their powers in the original periodograms are caused specifically by the signal (or the combination of the signal with the time series) and thus can never be reached in the periodograms of the bootstrap samples, where the signal is not present. This is what we already expected to be the case for the simulated frequency (for which we expected to always get a FAP of 0), but for the determined frequency we thought there would be a possibility of the power reaching its height in the original periodogram because of that frequency being for example a spurious period, so that it would not need the signal to reach that height. If this case is true, then the determined frequencies from our simulation also need the signal to reach the height of their powers in the original periodograms, indicating that the determined frequencies are all some kind of alias of the underlying signal.

It is also possible for a combination of the two cases above to be the actual explanation of the discrepancy.

In either of the cases we can conclude that our FAP did not give us the desired result and the computation of this version of the FAP does not give any useful information about the correctness of the determined frequency.

6.4 General limitations for period determination of Gaia data

Our research is a first step towards finding a way to automatically determine the main frequency of variable sources with Gaia data. In other words, we are working towards an unsupervised algorithm which determines the main frequency. There are a few important limitations for any possible algorithm of this kind.

Firstly, we choose periodogram evaluation frequencies between 0 and 100 1/d. However, in reality it is also possible for variable sources to have frequencies up to tens of thousands per day. As mentioned in Section 4.1 these frequencies would be very hard or even impossible to retrieve from the data because of the shortest possible time sampling in the Gaia data. Therefore, no algorithm would be able to retrieve these frequencies from the data with certainty. This in itself is not really a problem, because there simply is not enough information about the signal to retrieve the right frequency, so we cannot expect the algorithm to be able to do so. However, it can form a problem if the algorithm is looking for the right frequency within the range of 0 to 100 1/d, while the actual main frequency lies outside this range, as the algorithm will then appoint an incorrect frequency

as the main frequency. It would be best if the method is such that the algorithm would not appoint any frequency as main frequency in this case. Either way, the algorithm will never be perfect because of this, so we have to take this limitation into account when interpreting the outcomes of the algorithm.

Secondly, we have found that spurious periods form a problem when looking for the main frequency. Although we have tried to find a way to determine whether a determined frequency might correspond to a spurious period, this answer will always remain probabilistic. Thus we will never be able to say with certainty whether a possible spurious period is or is not equal to the main frequency. For example, there can be situations in which the main frequency of the signal is very close to a spurious period.

Lastly, we found that too few data points will result in aliasing, because we do not have information about all parts of the phase of the signal. Because of this there can also be frequencies that fit (almost) equally well to the data, without being spurious periods in the sense of Gaia data in general, making it impossible to distinguish between them. In such situations more information about the signal and thus more data points will be necessary to determine the correct main frequency.

6.5 Ideas for further research

6.5.1 Samples with different noise realisations

Instead of computing a false alarm probability using a bootstrap method and thus a permutation of the data points, one could try using samples that each have the same time series and average flux per data point, but with for every sample a different realisation of the noise added to the individual fluxes. This method would not be useful to compute a kind of false alarm probability, as the signal would still be present in all samples, meaning that the probability to get a certain period with maximal power from the sample periodograms has nothing to do with a false alarm. In fact, both the actual underlying frequency and spurious periods would still be expected to occur with high power in the samples, as both the signal and the time series are still present. However, this method could be used to analyse whether the determined frequencies with maximal power also have high (similar) powers for all other samples. If this is not the case and the power fluctuates much between the different samples, this could indicate that the high power at this frequency is caused by noise and not by the signal.

6.5.2 Maximal power significance

Something else that could provide useful information about the correctness or probability of correctness of the determined frequency is the significance of the maximal power with respect to the powers of other peaks in the periodogram. For example, if there is a periodic repetition of peaks that all have similar power, causing the maximal power to be similar to these other powers, this could be an indication of aliasing and thus of indistinguishability of these peaks. On the other hand, if the maximal power is significantly higher than all other powers in the periodogram, this could indicate that the determined frequency is the correct frequency that we are looking for. However, these are only hypotheses, so more research in this direction with the help of simulations would be necessary to find out whether this could really help determining the correctness of the maximal power frequency.

6.5.3 Improvement of the NUFFT periodogram

In this research we have worked with the FINUFFT periodogram. We found in Section 3 that spectral leakage induced by the rectangular observation window causes a kind of noise in the periodogram in the form of side lobes. Recently, a new way to compute a NUFFT periodogram has been introduced in [17], which uses a NUFFT in combination with different forms of tapers (which are similar to the idea of the rectangular window that we discussed) which are mathematically independent of one another, after which they take the average over the resulting NUFFTs. They claim that this method results in a final periodogram where spectral leakage caused by the observation window is reduced. Although we have not encountered any negative effects on the frequency determination caused mainly by spectral leakage, as the effects of aliasing and the spurious periods were more significant, it would be interesting to investigate whether the so called mtNUFFT (multitaper NUFFT) helps achieve better results with respect to the classical NUFFT periodogram.

Chapter 7

Conclusions

Our main research questions were whether the method of taking the frequency with maximal power from the NUFFT periodogram gives us the main frequency of the underlying signal of the variable source and whether it would be possible to distinguish between in this way correctly and incorrectly determined frequencies.

We found for our simulations of sine wave signals using 35 different Gaia DR3 time series and 20 different simulation frequencies with fixed signal to noise of about 5 that about 90% of the determined frequencies were correct. In general we found that the more data points and / or visibility periods, the more information about the signal and thus the higher the probability to retrieve the correct frequency from the periodogram by choosing the one with maximal power. Thus, the method of regarding the frequency with maximal power in the periodogram as the main frequency of the underlying signal is correct in many cases covered by our simulation (taking different values of for example the signal to noise will likely influence this result), but the problems of aliasing and spurious periods caused by the time sampling of the data points can influence the periodogram powers such that the determined frequency is incorrect. Moreover, aliasing can make it impossible to distinguish between multiple frequencies as they all fit the data almost equally well.

Our method to compute a false alarm probability resulted in almost all false alarm probabilities of both the determined and simulated frequencies being equal to zero. Therefore, this measure does not provide us with any useful information about the correctness of the determined frequencies. Other methods with the aim to distinguish between correctly and incorrectly determined frequencies have to be further investigated in order to conclude whether this is possible.

Acknowledgements

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

I want to thank Anthony Brown and Onno van Gaans for their knowledge, insights, feedback, guidance and moral support, without which this research could not have been successfully completed.

Bibliography

[1] J. R. Percy, *Understanding variable stars*, Cambridge University Press, 2007.

[2] L. Eyer and N. Mowlavi, *Variable stars across the observational HR diagram*, Journal of Physics: Conference Series **118**, 012010 (2008).

[3] J. T. VanderPlas, *Understanding the Lomb-Scargle periodogram*, The Astrophysical Journal Supplement Series **236**, 16 (2018).

[4] L. Eyer et al., *Gaia Data Release 3. Summary of the variability processing and analysis*, Astronomy & Astrophysics (2022).

[5] A. H. Barnett, J. Magland, and L. af Klinteberg, *A parallel nonuniform fast Fourier transform library based on an "exponential of semicircle" kernel*, SIAM Journal on Scientific Computing **41** (2019).

[6] D. Ruiz-AntolÃn and A. Townsend, *A nonuniform fast Fourier transform based on low rank approximation*, SIAM Journal on Scientific Computing **40** (2018).

[7] S. Deb and H. Singh, *Light curve analysis of variable stars using Fourier decomposition and Principal component analysis*, Astronomy and Astrophysics **507** (2009).

[8] A. Vallenari, A. G. A. Brown, T. Prusti, and et al., *Gaia data release 3. Summary of the content and survey properties*, Astronomy & Astrophysics (2022).

[9] T. Prusti et al., *The Gaia mission*, Astronomy & Astrophysics **595**, A1 (2016).

[10] B. Holl, C. Fabricius, J. Portell, L. Lindegren, P. Panuzzo, M. Bernet, J. Castaneda, G. Jevardat de Fombelle, M. Audard, C. Ducourant, and et al., *Gaia data release 3. Gaia scan-angle-dependent signals and spurious periods*, Astronomy & Astrophysics (2023).

[11] R. V. Baluev, *Assessing the statistical significance of Periodogram Peaks*, Monthly Notices of the Royal Astronomical Society **385**, 1279 (2008).

[12] A. v. Rooij, *Fouriertheorie: van reeks tot integraal*, Epsilon, 1988.

[13] E. M. Stein and R. Shakarchi, *Fourier analysis: An introduction*, Princeton University Press, 2003.

[14] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, Prentice-Hall, Inc., 2 edition, 1999.

[15] D. L. Cohn, *Measure theory*, BirkhaÌuser / Springer, 2013.

[16] L. Eyer and F. Mignard, *Rate of correct detection of periodic signal with the Gaia satellite*, Monthly Notices of the Royal Astronomical Society **361**, 1136 (2005).

[17] A. A. Patil, G. M. Eadie, J. S. Speagle, and D. J. Thomson, *Improving power spectral estimation using multitapering: Precise asteroseismic modeling of stars, exoplanets, and beyond*, arXiv preprint (2022).

Appendix A

Fourier transform for distributions

The definition of the Fourier transform can be extended to objects called distributions. In this Appendix we cover some of the theory behind and the derivation of this extension of the Fourier transform, as well as some examples of how this theory can be used and interpreted. The main theory is taken from [13].

Definition 16. We define the *Schwarz space* by

$$S := \{v : \mathbb{R} \to \mathbb{C} | v \text{ is } n \text{ times differentiable and } x \mapsto |x|^m |v^n(x)| \text{ is bounded } \forall n, m \in \mathbb{N}\}. \tag{A.1}$$

Clearly, S is a vector space. The elements v of S are called *test functions*.

Definition 17. A *distribution* is a linear map $\phi : S \to \mathbb{C}$. If $g : \mathbb{R} \to \mathbb{C}$ is a function such that the product gv is integrable on $\mathbb{R}$ for all $v \in S$, then

$$\phi_g(v) := \int_{-\infty}^{\infty} g(t)v(t)\mathrm{d}t \tag{A.2}$$

defines a distribution $\phi_g : S \to \mathbb{C}$. This distribution is called the *distribution induced by the function* g.

Proposition 18. *If $g : \mathbb{R} \to \mathbb{C}$ is integrable or piecewise continuous and bounded, then gv is integrable for all $v \in S$, thus g induces a distribution.*

However, not all distributions are induced by functions.

Example 19. The map $\delta : S \to \mathbb{C}, v \mapsto v(0)$ is a distribution by Definition 17, as it is a linear map from $S \to \mathbb{C}$. If $g : \mathbb{R} \to \mathbb{C}$ would exist such that

$$v(0) = \delta(v) = \int_{-\infty}^{\infty} g(t)v(t)\mathrm{d}t \text{ for all } v \in S, \tag{A.3}$$

then one can show that $g(t) = 0$ for all $t \neq 0$ while $\int_{-\infty}^{\infty} g(t)\mathrm{d}t = 1$. Thus, g would be the Dirac delta "function", which is not a function. The map $\phi = \delta$ is called the *Dirac delta distribution*.

Distributions are sometimes called *generalised functions*.

Definition 20. If $g : \mathbb{R} \to \mathbb{C}$ is integrable, then $\hat{g} : \mathbb{R} \to \mathbb{C}$ is continuous and bounded, so by Proposition 18, $\hat{g}$ induces a distribution given by

$$\begin{aligned}
\phi_{\hat{g}(v)} &= \int_{-\infty}^{\infty} \hat{g}(\omega)v(\omega)\mathrm{d}\omega \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(t)e^{i\omega t}\mathrm{d}t \right) v(\omega)\mathrm{d}\omega \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} v(\omega)e^{i\omega t}\mathrm{d}\omega \right) g(t)\mathrm{d}t \\
&= \int_{-\infty}^{\infty} \hat{v}(t)g(t)\mathrm{d}t \\
&= \phi_g(\hat{v}). \qquad \text{(A.4)}
\end{aligned}$$

Furthermore, if $v \in S$ then also $\hat{v} \in S$ and vice versa. Therefore, we can define the *Fourier transform of a distribution* ϕ by

$$\hat{\phi}(v) = \phi(\hat{v}) \text{ for all } v \in S. \qquad \text{(A.5)}$$

Example 21. Let δ be the Dirac delta distribution. For every $v \in S$ we have that

$$\hat{\delta}(v) = \delta(\hat{v}) = \hat{v}(0) = \int_{-\infty}^{\infty} v(t)e^{i0t}\mathrm{d}t = \int_{-\infty}^{\infty} 1v(t)\mathrm{d}t = \phi_1(v). \qquad \text{(A.6)}$$

So we find that $\hat{\delta}$ is the distribution induced by the constant 1 function. Loosely speaking, we say that $\hat{\delta}$ equals the constant 1 function.

Example 22. Let $g : \mathbb{R} \to \mathbb{C}, t \mapsto \sin(t) = \frac{e^{it}-e^{-it}}{2i}$. For all $v \in S$ we find that

$$\begin{aligned}
\widehat{\phi_g}(v) &= \phi_g(\hat{v}) \\
&= \int_{-\infty}^{\infty} \hat{v}(t) g(t) \mathrm{d}t \\
&= \int_{-\infty}^{\infty} \hat{v}(t) \frac{e^{it} - e^{-it}}{2i} \mathrm{d}t \\
&= \frac{1}{2i} \left(\int_{-\infty}^{\infty} \hat{v}(t) e^{it} \mathrm{d}t - \int_{-\infty}^{\infty} \hat{v}(t) e^{-it} \mathrm{d}t \right) \\
&= \frac{1}{2i} \left(\hat{\hat{v}}(1) - \hat{\hat{v}}(-1) \right) \\
&= \frac{2\pi}{2i} \left(v(-1) - v(1) \right) \text{ by Theorem 5} \\
&= \frac{\pi}{i} (\delta_{-1} - \delta_1) \text{ by Example 19.}
\end{aligned} \tag{A.7}$$

This Fourier transform of the sine wave in terms of distributions is equal to what we found with the more intuitive computations in Chapter 3.

Version of July 6, 2023– Created July 6, 2023 - 19:17