DSB

MiniProject D

Aarhus Institue of Technology

Morten Høgsberg: 201704542
Sune Dyrbye: 201205948
Yehven Parolia: 20112870

Date: December 21, 2021

Mean, variance and effect

Considering some dataset, it might at times be useful to do statistical analysis on some dataset to see if it mean values are reasonable in accordance with the theoretically predicted data.

it might also be useful for determining if some data set is pure noice of if it might contain some statistically significant event.

The mean value of a dataset is given by the simple equation

$$\mu = \frac{1}{N}\sum_{i=0}^{N} x_i$$

which comes out to 5062 using matlabs sum and length functions. Nothing interesting can be said for the mean value by itself but the standard deviation and variance might give some clues as to how much the data deviates from the mean.

both the standard deviation, σ and the variance σ^2 are defined in the following ways

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(X_i - \mu)^2} \tag{1}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(X_i - \mu)^2$$

For both the standard deviation and the variance, using matlabs std and var functions we get $\sigma = 38.4$ and $\sigma^2 = 1.480$. The statistical significance of these figures however are unclear in the context of looking for exoplanets as is the case with this worked upon dataset.

Taking a look at the signals effect which is defined similarly to the variance

$$s_{effect} = \frac{1}{N}\sum_{i=0}^{N} X_i^2 \tag{2}$$

the effect of the signal is defined as the mean square of the dataset. When we compare (2) and (1) we can see that the variance tells us the mean distance that the signals effect has to from the mean value of the signal.

lastly we can plot the amplitude of the dataset as a histogram to get an idea of the most commonly occuring signal amplitudes.

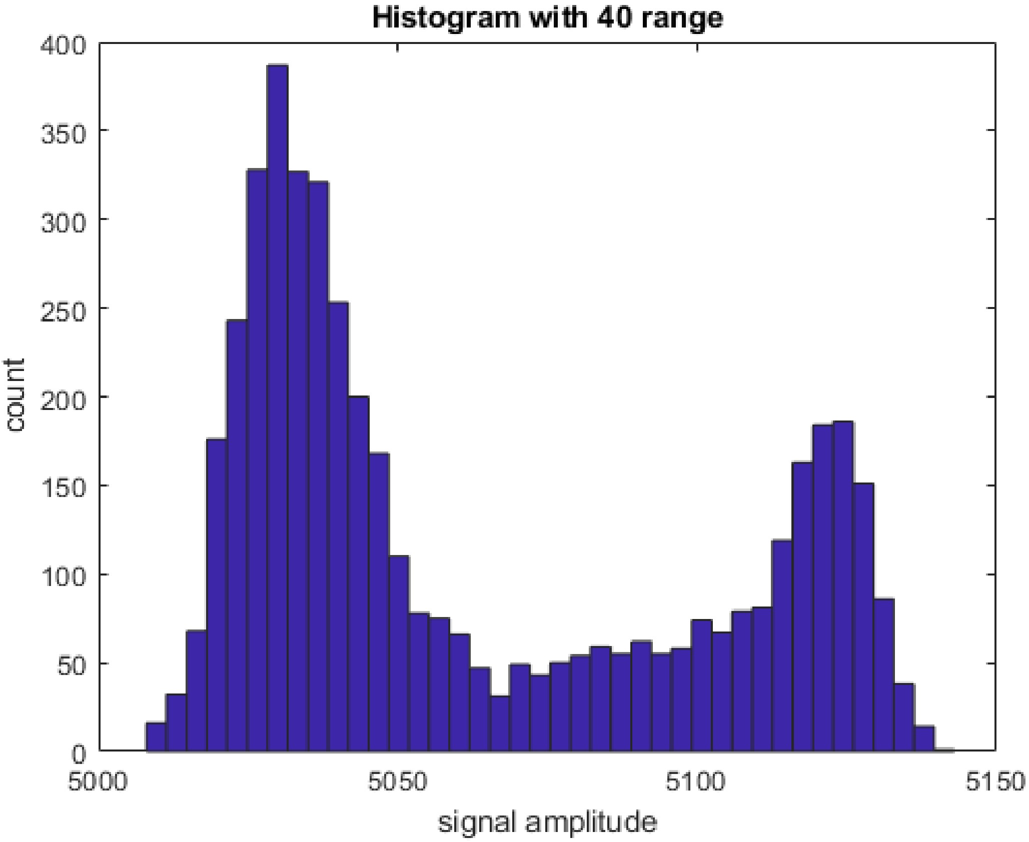

Figure 1: amplitude histogram, for dataset of exoplanets

Correlation of data

One way to detect exoplanets in a time series, such as the one being analysed here, is to use an autocorrelatin function. This is possible because the decrease in luminousity caused by the exoplanet passing in front of the star will happen with a set periode (the periode of the exoplanets orbit). It is therefore expected that theese dips wil show up in the correlation as lags where the fit is better than usual. In order to get a usable result the timeseries has to be ‘flattened’, removing the features caused by the drift in the telescope calibration. This is done by using a moving mean filter with a binsize of 50 (chosen to be big enough that the interesting features are not removed, while still being small enough to remove all the unwanted features) and subtracting this from the raw data, leaving only the flat data.

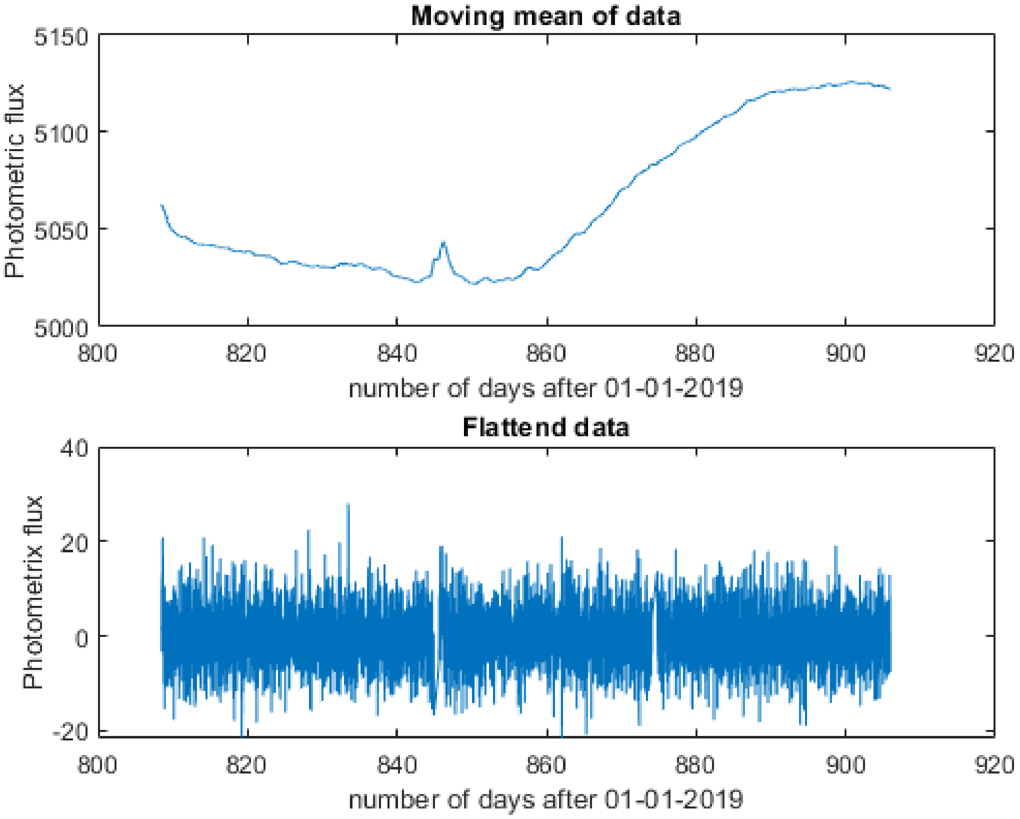

Figure 2: Moving mean and flattened dataseries

The flattened data and the moving mean are shown in Figure 2.

It is now possible to autocorrelate the data to see if there are any periodic behaviour. This is shown in Figure 3. The center peak is of cource the complete overlap of the data, and goes up to one. It is also clear that there is a handfull of lags with a decent overlap. The larges of these are suspected to be the gaps in the data linig up. There are a series of smaller peaks which appear somewhat evenly spaced out, at around 400–450 steps. This could very well be an exoplanet, though it might also be some other feature of the star.

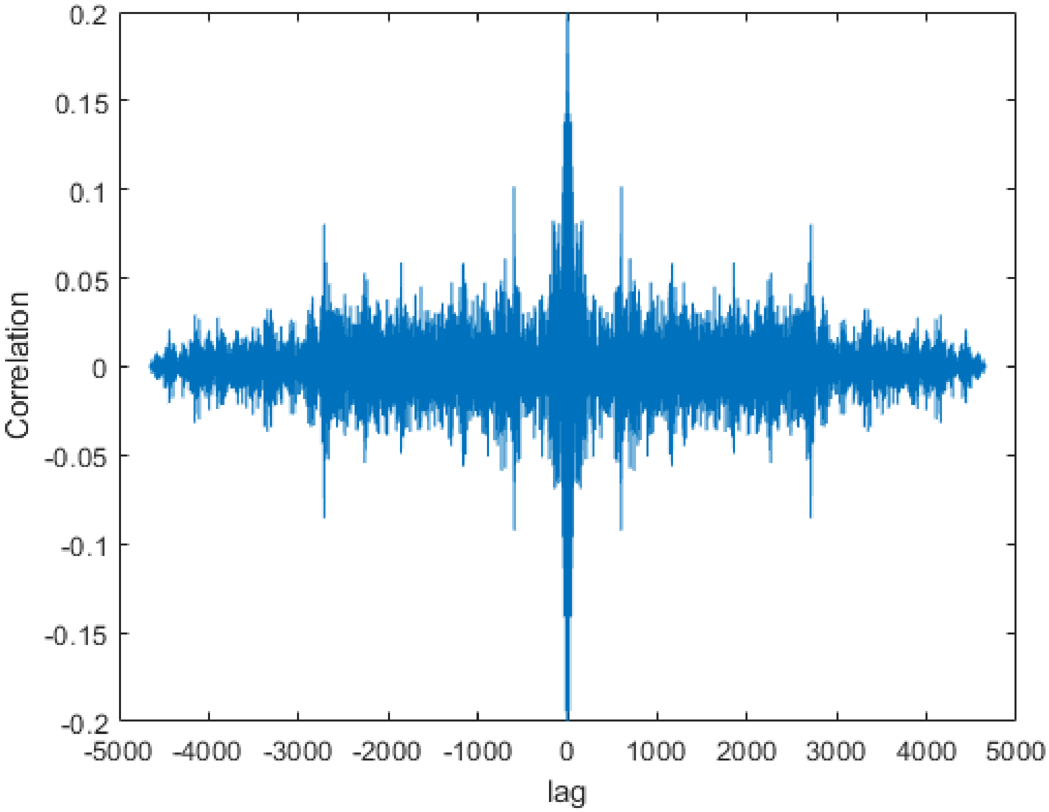

Figure 3: Autocorrelation of time series

One way to check if the feature is actually an exoplanet the timeseries if folded so the datapoints of the feature will overlap. This is shown in figure Figure 4. It is clear from the figure that there is no significant feature in the data with the found frequency.
The star meassured is KIC10001893, which have the signs of three exoplanets[1]. The features might not, however, be actual planets after all[2]. Given the questionable nature of the exoplanets surrounding KIC10001893, it is not surprising that they are not found using the relatively simple autocorrelation.

DFT spectral multiplication

We will now try to show that a there is another way doing filtering with spectral multiplication in frequency domain instead of convolution in time domain. A guiding figure 5 is shown below. We first take convolution of the signal with the filter to represent the time domain filtering. Secondly we take the filter , the data and fft transform them independently, multiply them in the frequency domain and reverse fft them back to time domain. The plot of time domain for both methods can be seen on figure 6

[1] https://arxiv.org/abs/1409.6975
[2] https://arxiv.org/abs/1906.03321

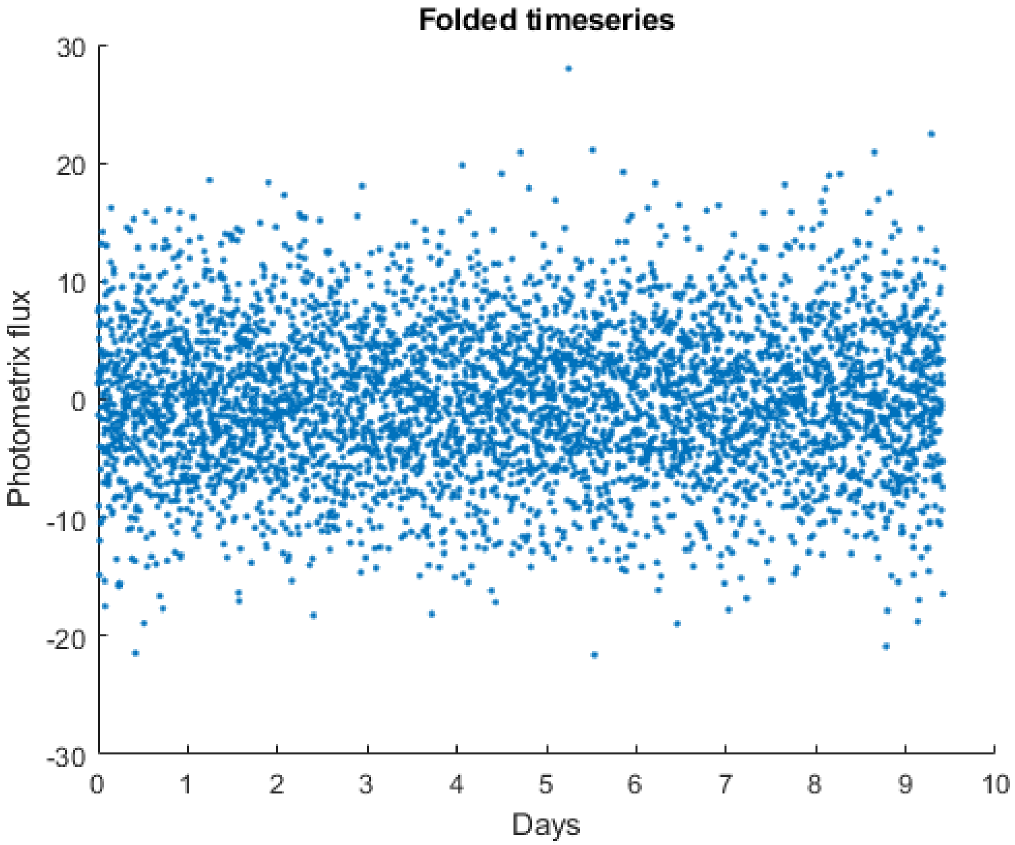

Figure 4: Time series folded to match the possible exoplanet periode

STFT of kepler data

Here we try use Matlab STFT spectrogram function to look for any prevalent frequency change over the observation time. We try to find a windows function that will give us more frequency resolution, the data is already very noisy and normal rectangular window will give a lot of articulates that will pollute the spectral plot. We go with blackman. The windows size is a trade-of between time resolution and frequency resolution, we go for something in between and as the number of samples is 4654 we go for 128.

The spectrogram plot is shown in figure 8. It is very difficult to see any change. This is due to very low data count and slow / very small changes, coupled with a lot of noise.

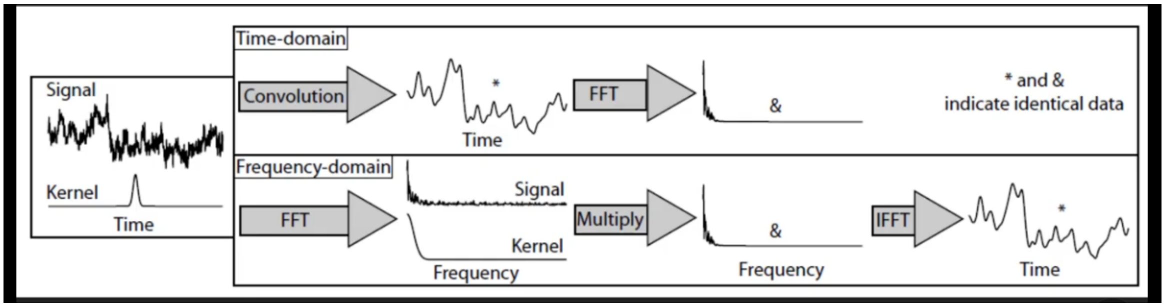

Figure 5: Time domain filtering versus frequency domain

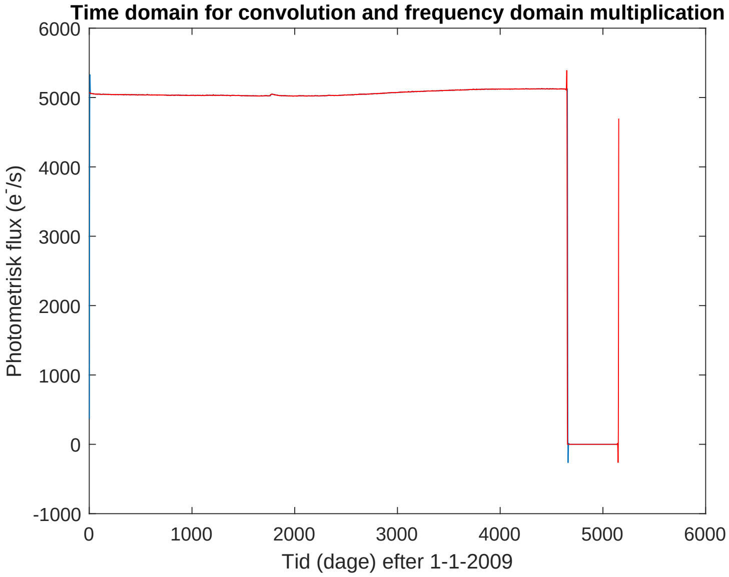

Figure 6: Time domain for both methods.

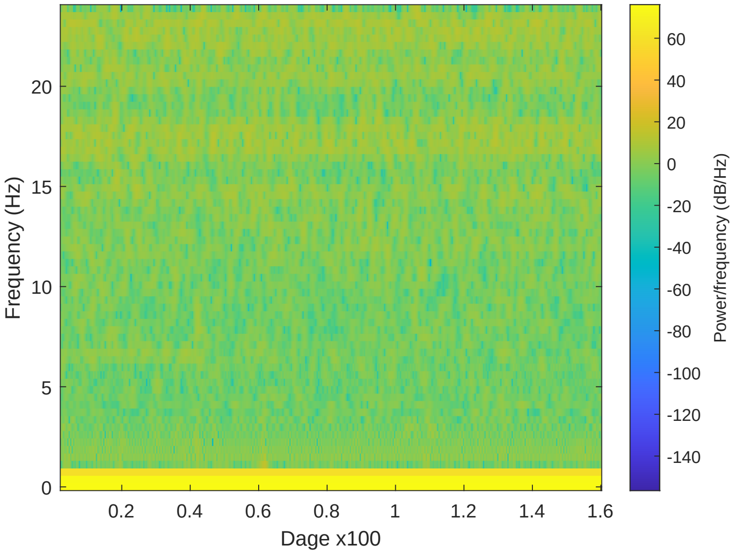

Figure 7: STFT spectrogram of the kepler data

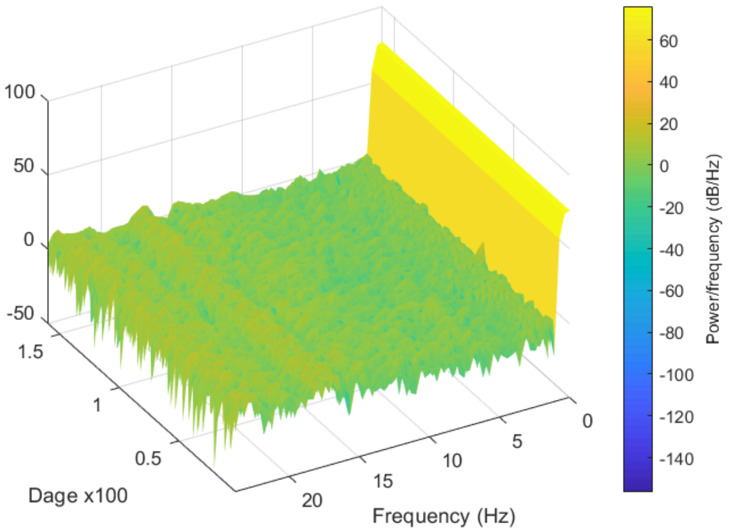

Figure 8: STFT spectrogram of the kepler data