Application of Monte Carlo Markov Chains for Solar Absorption Line Fitting

Antoine Paradis,[1]★ Shu Zhang,[1]†
[1]Department of Physics, McGill University, Montreal, QC, Canada

15 April 2024

ABSTRACT

Using advanced Monte Carlo Markov Chain (MCMC) fitting techniques, we aimed to identify and measure the position of the H α Balmer and O_2 Fraunhofer B line in the Sun's spectrum. The spectral data was collected in March 2023 and consists of 15 spectra collected directly from the moon's reflection and from the sky (N=30). We find that both the H α Balmer and O_2 Fraunhofer B line results agree with their theoretical values within 1 standard error. Further, we find support for using MCMC as a statistical spectra analysis tool in place of the simpler average/minimum value method used in Bordeniuk et al. (2023) with the same data. Finally, we discussed potential improvements to the experiment, notably regarding the spectrometer used for the data collection.

Key words: Sun: radiation – ground-based spectra – Methods: data processing – filtering – MCMC

1 INTRODUCTION

The idea that energy is fundamentally quantized was first proposed by Max Planck in 1900. As electrons jump from different orbitals around their parent nucleus due to excitation and de-excitation, Planck postulated that they did so in fixed jump sizes that were directly related to the nature of the atom itself. Most importantly, this entailed that every transition released a fixed, predictable amount of energy, governed by the simple equation

$$E_{photon} = \frac{hc}{\lambda}, \tag{1}$$

where h is the planck constant, c the speed of light, and λ the wavelength of the emitted photon after the transition occurs to a lower energy level. In this paper, we investigate the H α Balmer transition at 656.2nm, as well as the O_2B transition at 686.7-688.4nm (Jenkins & White (1981)), which are the most prominent absorption features in our data set. The hydrogen feature occurs because of its strong presence in the Sun's atmosphere, while the oxygen feature is caused by Earth's atmosphere, which absorbs part of the incoming solar rays at this exact wavelength.

2 METHODS AND ANALYSIS

Since we aim to locate and measured the exact wavelength of the absorption lines from our observations, we must extract the features from the data and transform them appropriately for the MCMC algorithm. To do so, we combine the 15 spectra from the moon's reflection and from the sky, and then correct the features by subtracting a smoothed version of each one. With this correction, we subset on the features and fit a Lorentzian to the Balmer line and a Gaussian to the O_2 Fraunhofer line by computing a Monte Carlo Markov Chain of length 5000.

2.1 Absorption Line Extraction

Ideally, the moon's emission spectrum should replicate that of the Sun since it is the reflection of its light. Hence, it is reasonable to assume that in both data sets, our spectra should resemble that of a Black-body curve

$$B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{exp\left(\frac{h\nu}{kT}\right) - 1}, \tag{2}$$

where ν is the frequency of the emitted photons, h the planck constant, and k the Boltzmann constant. The Sun's effective surface temperature is $T = 5500K$. However, since our data comes from ground-based observations, there is significant loss in the ultraviolet and infrared regions due to atmospheric absorption. Additionally, the spectrometer that was used performs badly in high and low frequency regions, hence deviating from the model further. With the sun itself deviating from a perfect black-body, and with the aforementioned factors, the collected spectra cannot be accurately modelled as black-body curves.

To extract the two emission lines of interest, a smoothed version of the spectrum was obtain using the Savitzky-Golay filter and then subtracted from the raw spectrum. The filter works, in the process of convolution, by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares. It has the advantage of smoothing the data without distorting the

★ E-mail: shu.zhang4@mail.mcgill.ca
† E-mail: antoinepparadis@gmail.com

signal tendency. Further, we use it as a built-in function from the *Scipy* library and choose the size of the window to be $w = 167$, dividing our data into approximately 100 zones. At this resolution, no information is lost about the behaviour of the spectral lines since the size of the window is larger than any peak or dip. Subtracting a best-fit Lorentzian was also investigated, but ultimately discarded as the correction was marginally worst. Finally, no noise data was available, at least with our apparatus, hence it could not be subtracted from the data.

2.2 MCMC Fitting

With the corrected features, we were able to perform MCMC fitting to locate the exact wavelength of the Balmer and Fraunhofer line. Ideal functions to fit the dips, which itself is a spectral profile, include the Lorentzian, Gaussian, and Voigt profile. All three functions allow to find the optimal parameters for the absorption line position, maximum absorption amplitude, and spread of the feature (width of the function). Considering this, we choose a Gaussian function

$$G(x,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right), \quad (3)$$

for the wider O_2 Fraunhofer feature and a Lorentzian function

$$L(x,\sigma) = \frac{k}{1+\left(\frac{x-p_0}{\sigma/2}\right)}, \quad (4)$$

for the sharper H α Balmer feature, where p_0 is the position of absorption line, k the amplitude, and σ the half-width of the function. We take these to be the model in the likelihood function, where we assume Gaussian errors on the intensity values, and assume a uniform prior with reasonable bounds observed from the data. We then run an MCMC with 5000 steps and extract the optimal wavelength parameters and compare them to the known literature values.

3 DISCUSSION

We successfully identify 2 absorption lines, H_αC and O_2B, with high precision in both data sets, as shown in table 1 below. The Balmer line values from both data sets do not agree within their standard error to the expected theoretical value of 656.2 nm, while both Fraunhofer line wavelengths agree to the theoretical range of 686.7-688.4nm. However, accounting for the systematic instrument error, the Balmer line measurement agrees in both case. We also note that errors are dramatically reduced using the MCMC fitting method compared to Bordeniuk et al. (2023), which obtain random errors on the order of 0.7 nm.

Further, we can observe the usefulness of the Savitzky-Golay filter in figure 1 and 2 in order to extract the Balmer feature. Running the MCMC on this feature with an inverse Lorentzian model, we plot a subset of the 5000 parameter sets in figure 3.

Table 1. Fitted values for H α Balmer and O_2 Fraunhofer B for Moon (1) and Sky (2) data in nanometers. Note that there is an additional systematic error of 0.5nm from the spectrometer, not included here for reasons discussed in Sec 3.2. Literature values refer to results from Jenkins & White (1981).

Absorption line	Fitting result (nm)	Literature value (nm)
α Balmer 1	656.4±0.1	656.2
α Balmer 2	656.66±0.07	
O_2 Fraunhofer B 1	688.33±0.06	686.7-688.4
O_2 Fraunhofer B 2	688.46±0.06	
Average error	0.07	

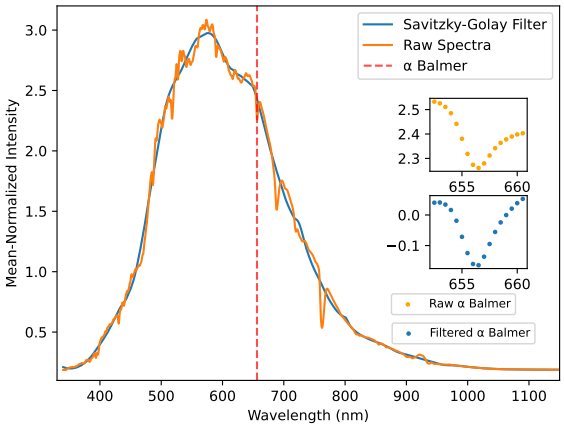

Figure 1. Savitzky-Golay Filter applied on the Moon spectrum, with isolated H α Balmer feature in the zoom-in subplot. The original spectrum, the orange line, shows an imperfect Maxwell-Boltzmann distribution shape. After smoothing the spectrum and subtracting it from the original, the peak signal (zoom-in subplot) shows an inverse Lorentzian distribution. Before the process, the signal shows no trending distribution, thus making it very hard to fit either a normal or Lorenzian distribution to it.

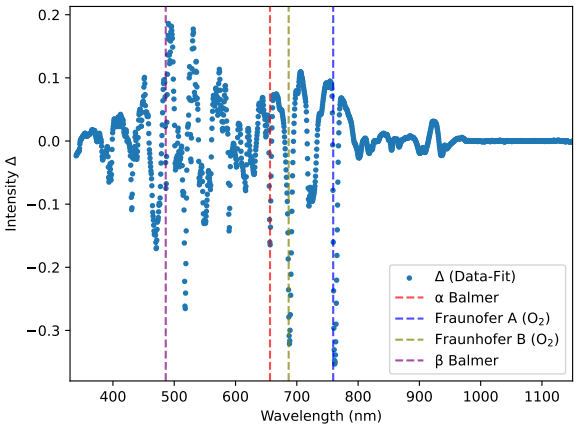

Figure 2. The subtracted spectrum from Fig.1. The graph now clearly shows the peaks of the absorption lines. Note that there are still some widely-spread peaks since the filter we deployed is very strong and keeps such fluctuation in the spectrum. The filter uses a wide window to prevent any loss in the absorption lines. We are able to visually identify our two lines of interest, H_αC and O_2B, as well as two others that suggest H_βC and O_2A.

3.1 Lorentzian, Gaussian, and other models

Naively, since the absorption line is the result of interactions between photons and atomic quantum states, it could be though that electrons will only "capture" certain photons with designated energy

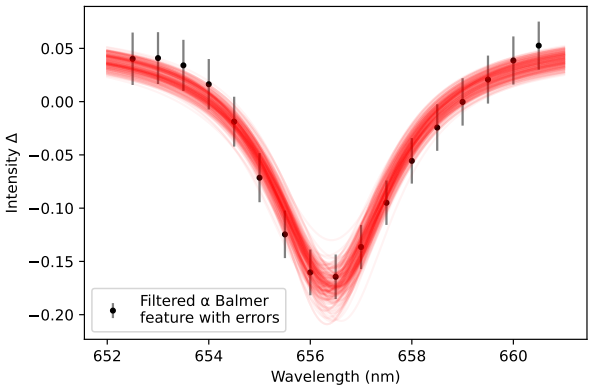

Figure 3. Subset of fitted Lorentzians from MCMC fitting. Notice the convergence of the curves as well as some fainter curves that briefly deviate in parameter space. The data demonstrate consistency with the Lorenzian model and acceptable error size compared to the amplitude of the spike.

to make it jump to the next quantized energy stage. This process is rather unmistakable because electrons won't initiate the jump for lesser energy than the threshold energy, but will for higher energy. As an example, in Fig.3, the dips are steeper on the side of lower frequency as $E = hc/\lambda$. This shape will immediately suggest a Maxwell-Boltzmann distribution. However, the Maxwell-Boltzmann distribution extends to infinity and diverges, which is not a behaviour of absorption lines. Thus it is not the ideal fitting function.

Further, absorption line distributions are rather pathological since electrons don't have uncertainty when "capturing" the photons, suggesting only the location of the peaks is distinct and any other parameters are undefined. This makes the Lorentzian distribution a better fit since the Gaussian fit also contains other well-defined parameters like the uncertainty σ. Hence, a Lorentzian function is our primary choice of model for the MCMC likelihood. The advantage of the Lorentzian distribution is more obvious for H_α line because hydrogen is a much simpler model with only one electron in its vicinity. It should therefore follow the behaviour mentioned above. However, the situation is more chaotic and probabilistic for oxygen due to the excess electrons, hence a boarder distribution is expected. This is directly reflected in the range of theoretical values for the O_2B wavelength of 686.7 to 688.4nm (Jenkins & White (1981)). We therefore fit a wider Gaussian distribution for the O_2B line.

3.2 Error Behaviour of Fit and Measurements

The uncertainty on our results comes from two main sources: the error from the fitting parameter, and the instrumental error from the equipment used during the data collection. The former is very small since our data is sampled across a finite wavelength range with small spacing. The peaks and dips are sharp and well-defined in our data, in which case we could easily locate the absorption line with high precision. Therefore, when fitted with the MCMC method, the returned error on the wavelength is very small.

On the other hand, the StellarNet spectrometer that was used for observations has a instrumental error of 0.5 nm reported by the manufacturer. Such an error is systematic and impossible to reduce by increasing the sample size. Traditionally, both factors matter, and when combined with a proper formula for error propagation, should produce the total error. However, the error from the MCMC fit is smaller than the systematic error from the camera by as much as 1 order of magnitude in some cases. Hence, if the 2 types of error would be combined, the camera error would dominate and drown out the random error from the MCMC fit. Thus, we regard the camera error as an offset on the data, and discard it for error calculation to reflect the true random error from the fitted and combined data.

4 FUTURE IMPROVEMENT

By fitting more than 2 absorption lines, we find that our results are always either an overestimation, or on top of a given range, compared to the literature values. The entire spectrum is stretched over the range of wavelengths, causing the shift in the resulting wavelength to be proportional to the wavelength itself. Since this error is present in both the sky and moon data, we think this error might be a result of instrumental limitations. Spectrometers collect analogue signals, and as such, when converting to digital ones for computers, there might be miscommunication. It would be important in a future study to determine the source of error and eliminate it. Thus, a higher-precision spectrometer camera would be essential.

We also assumed the errors on the intensity values to be normally distributed, which was not an assumption that was fully supported by our data. We noticed an overall correlation with wavelength, with the largest errors close to the spectra peaks. Hence, a better probability distribution for the intensity errors could improve the likelihood function and reduce even further the random errors on our fitted wavelength values.

5 CONCLUSIONS

We successfully identify at least 2 absorption wavelengths with our method. With high precision and averaged from both the sky and moon spectra, we identify the H_α Balmer line to be $656.53 \pm 0.06nm$, and the O_2B Fraunhofer line to be $688.40 \pm 0.05nm$. These values both agree to theoretical expectations when considering both systematic and random errors, and we conclude that the systematic overestimations are caused by the low precision of the spectrometer camera and the built-in ADC algorithm in the camera's driver. We propose future improvements by using better equipment to resolve the existing systematic errors and by adjusting the likelihood function to better reflect the errors in the data.

ACKNOWLEDGEMENTS

Thank you to Bordeniuk B., Pereira O., and Walton-Knight B. for their help collecting the data used in this study. We also appreciate Georgia Mraz's help regarding feature extraction methods for our analysis.

REFERENCES

Bordeniuk B., Paradis A., Pereira O., Walton-Knight B., 2023, The Moon Does Not Twinkle: Comparing the Blue Sky Solar Spectrum with Moonlight, McGill Physics Journal

Jenkins F. A., White H. E., 1981, Fundamentals of Optics (4th ed.), McGraw-Hill. p. 744. ISBN 978-0-07-256191-3

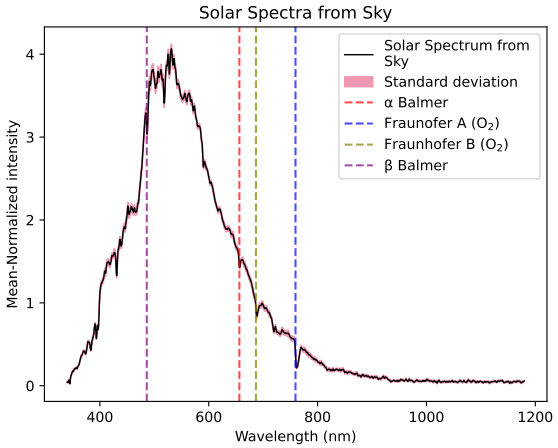

Figure A1. Raw sky data with mean-normalized intensity as a function of wavelength.

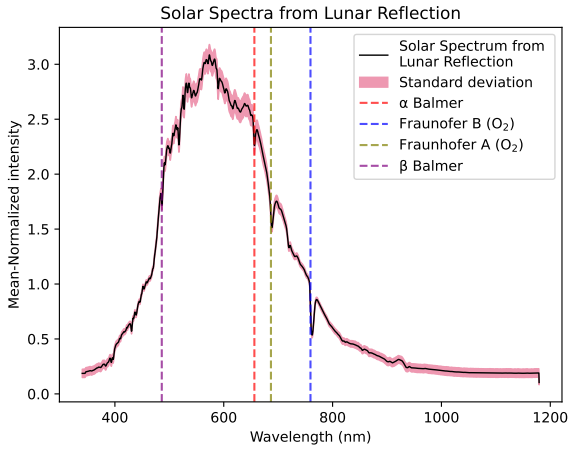

Figure A2. Raw Moon data with mean-normalized intensity as a function of wavelength.

APPENDIX A: COMBINED MOON AND SKY SPECTRUM

See above figure A1 and A2.

APPENDIX B: SAMPLE CODE AND DATASET

Sample code for the H_α Balmer line MCMC fit is available here, along with the data used in this study.

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