# Diagnostics of Spectral Feature in Mg II line from the Chromoshpere

AstroML - Project

Sambit Kumar Giri

May 9, 2016

## 1 Introduction

With a lot of data coming in, it has become cumbersome to analyze them manually. A better way to do it is "Machine Learning". This project focuses on predicting the atmosphere of solar chromosphere from selected spectral features in the Mg II line. In previous literature, the line has been analyzed by forward modeling with MHD simulations (Gudiksen et al., 2011; Leenaarts et al., 2013b; Carlsson et al., 2016).

NASA's Interface Region Imaging Spectrograph (IRIS) small explorer mission is studying the process in which the solar atmosphere is energized. IRIS contains an imaging spectrograph that covers the Mg II h&k lines as well as a slit-jaw imager centered at Mg II k. Understanding these line features is of scientific importance. The forward modeling is computationally expensive method to analyze the data. The use of 'machine learning' to learn from previous MHD simulation and predict the output for the new data can be a way to make the analysis faster.

## 2 Data

The data used in this project is the Mg II line. This line is affected by the chromosphere of the sun. The Mg II contain h&k resonance doublet at 280.27 and 279.55 nm respectively that are strongest and most valuable for diagnosis of the solar atmosphere. In figure 1, the k line feature is shown. It contains two asymmetric peaks known as $k_{2v}$ and $k_{2r}$. The trough in between these two peaks is called $k_3$. The h line feature appears very similar to the k-feature.

The average temperature of the chromopshere is the output parameter that is to be predicted. In order to train the machine learning algorithms, the outputs from the radiative transfer computations have been used. It is a non-LTE radiative transfer computations using the four-level plus continuum model atom from Leenaarts et al. (2013a) with two different codes. The first is a version of RH by Uitenbroek (2001) and the second code is *Multi3d* (Leenaarts and Carlsson, 2009).

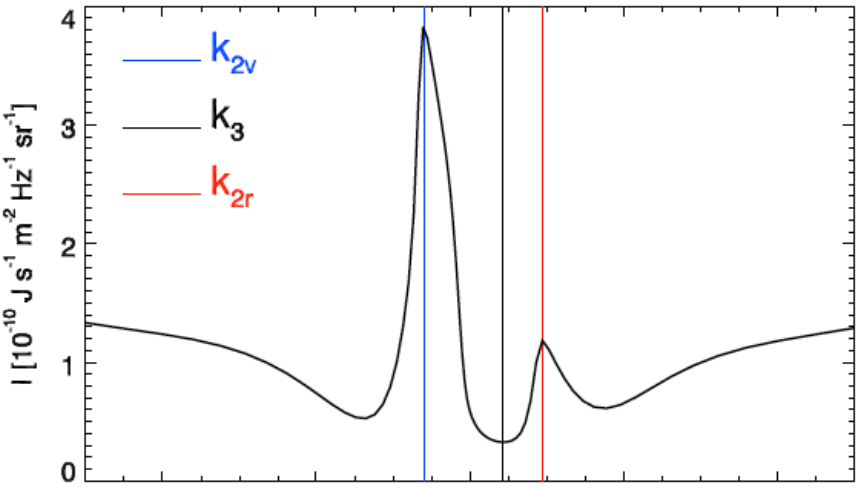


Figure 1: The plot of the k line feature of Mg II has been taken from Leenaarts et al. (2013b)

### 2.1 Correlation Test

The flux of the line that has been observed comes from a different layer of the atmosphere. However, the selected features are very close to each other. Therefore, these features are expected to affect the average temperature of the region. In order to quantify this idea, the selected line features should be correlated or anti-correlated with the average temperature of the atmosphere. The Spearman coefficient is calculated for the features that is used as input (Zwillinger and Kokoska, 2000).

In order to select the features that would be most likely to affect the temperature of the solar atmosphere, the correlation of the selected line feature has been done with the temperature that we have from the simulations. In figure 3, most of the features show good relation with the output.

## 3 Machine Learning Methods

The output that is to be predicted using machine learning depends on the whole line profile. In order to predict the nearest possible values, the input variables have to be given in such a way that it includes some of the extra physical ideas. Some of the these extra features can be be the asymmetry of the bumps and relative frequency at which they form Leenaarts et al. (2013b).

### 3.1 Features and labels

The features used are shown in figure 4. The data contains 252x252 images of intensity peaks and doppler shift at $k_{2v}$, $k_3$, $k_{2r}$, $h_{2v}$, $h_3$ and $h_{2r}$. The radiative transfer computation gives temperature values at each of these pixels. Therefore, we can relate each temperature pixel to the respective pixels of the

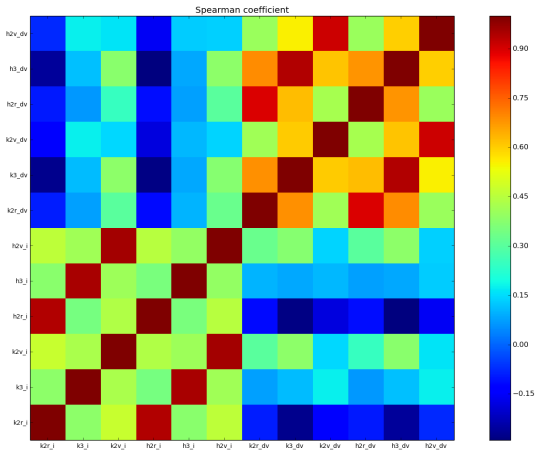


Figure 2: The Spearman coefficient for the line parameters is shown.

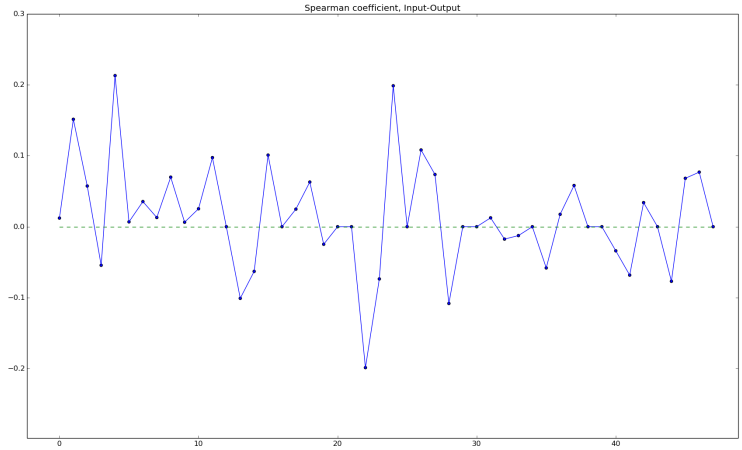


Figure 3: The Spearman coefficient for the line parameters with the average temperature is shown.

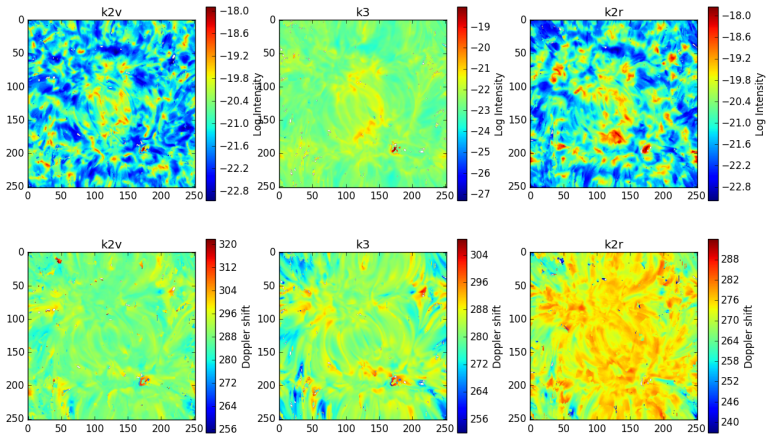


Figure 4: *Top Panel:* The intensity images of the k-features is shown in the log-scale. *Bottom Panel:* The doppler shift images of k-feature is shown.

feature images independently. This gives 63504 data points to feed the machine learning algorithm. In order to fit an even better model, the asymmetry of the peaks and relative doppler shifts of the h&k features are included. Leenaarts et al. (2013b) suggests a formula to incorporate these information.

$$R_v = \frac{I_{k2v} - I_{k2r}}{I_{k2v} + I_{k2r}} \tag{1}$$

A similar equation can be written for doppler shifts also. Therefore, there are 36 more features along with the previous 12 features. The data contains NAN at many pixels. This can lead to inaccurate fitting of models. The NAN pixels are not common for every set of data. The union of all these NAN pixels was found to be about 1% of the data points, which was removed.

## 3.2 Regression Methods

The radiative transfer computations is so complex that it is very difficult to reproduce the same temperature with these limited line information. However, some regression method can give better predictions than others. It is difficult to decide the superiority of any method for this complicated data.

After removing the NAN pixels from the data, the set has been randomly shuffled. This is done so that there is no bias due to edge of the images. The data set is split into training and test set using the *train_test_split* module from *sklearn*. Various regression methods have been applied to the data. The results predicted for the test set is compared with the true values in figure 5.

In figure 5, it can be seen that the predicted values with the linear regression have a very bad correlation with the true values. The figure also shows that some of the regression methods, such as "Decision Tree Regression", "K-Neighbors

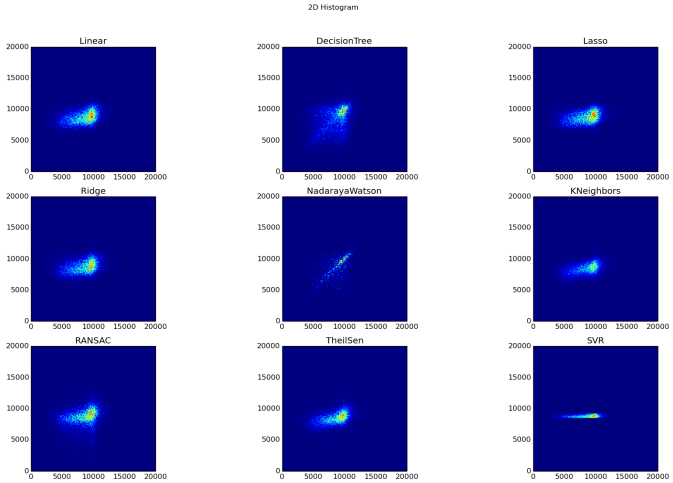


Figure 5: The 2D histogram of predicted average temperature is compared to the true average temperature for different methods. The x-axis has the true values where as the y-axis has the predicted values.

Regression" and "Nadaraya-Watson Regression", work quite well. The number of neighbors used in K-Neighbors regression is 15.

# 4 Results

After the preliminary test, it was found that the some regression methods worked better than others. The predictions with these methods are shown is figure 6. A small box was cropped out of the input images. The rest part of the box is used to train the regression method. A subset of pixels is chosen from the rest part of the box randomly while training the method. In this way, the model can be cross-validated by randomly choosing a subset of pixels again. After cross-validation, the trained methods are used to predict the average temperature for the cropped box.

In figure 6, the small box has been predicted using four different regression methods. The linear regression has failed to show any pattern that we can see in the rest part of the image. The values predicted by the linear method has a very low variance. The values predicted by the K-Neighbor regression seems to be very close to the true values. The large scale patterns in the image is visible, however, the patch looks very grainy. The Nadaraya-Watson regression gives better resolved patterns. This is not obvious because it contains a lot of white spots. These white spots are the points where it couldn't predict values. However, the pixels for which it could predict the output, it was almost accurate. The Decision Tree regression gave the best results among all the methods. The small box predicted by this method has the large scale patterns along with some small scale patterns. The method could predict for all the pixels in the small box.

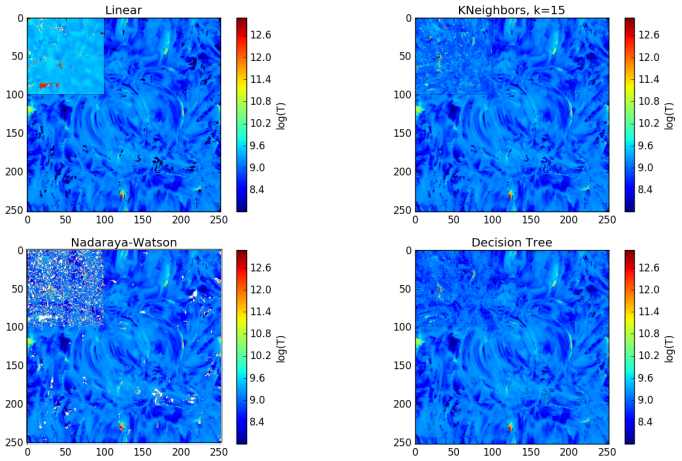


Figure 6: The upper left box in each of the subplot shows the predicted values. The rest part of the subplot is the true value that has been computed from radiative transfer calculations.

# 5 Conclusion

To understand the details of the atmosphere it is necessary to simulate the whole atmosphere since the different layers interact strongly. Therefore, it is very difficult to predict the exact atmosphere considering a few spectral features. Even though these features carry more information than the other part of the line, they are not enough.

However, the MHD modeling involves highly efficient massively parallel numerical code to solve the associated equations. These numerical codes need an initial guess value for the atmosphere. The code will converge faster if the guess values are closer to the solution. The machine learning techniques shown above can be used to get the initial values. This can reduce the run-time of the simulations.

# References

Carlsson, M., Hansteen, V. H., Gudiksen, B. V., Leenaarts, J., and De Pontieu, B. (2016). A publicly available simulation of an enhanced network region of the Sun. *A&A*, 585:A4, 1510.07581.

Gudiksen, B. V., Carlsson, M., Hansteen, V. H., Hayek, W., Leenaarts, J., and Martínez-Sykora, J. (2011). The stellar atmosphere simulation code Bifrost. Code description and validation. *A&A*, 531:A154, 1105.6306.

Leenaarts, J. and Carlsson, M. (2009). MULTI3D: A Domain-Decomposed 3D Radiative Transfer Code. In Lites, B., Cheung, M., Magara, T., Mariska,

J., and Reeves, K., editors, *The Second Hinode Science Meeting: Beyond Discovery-Toward Understanding*, volume 415 of *Astronomical Society of the Pacific Conference Series*, page 87.

Leenaarts, J., Pereira, T. M. D., Carlsson, M., Uitenbroek, H., and De Pontieu, B. (2013a). The Formation of IRIS Diagnostics. I. A Quintessential Model Atom of Mg II and General Formation Properties of the Mg II h&k Lines. *ApJ*, 772:89, 1306.0668.

Leenaarts, J., Pereira, T. M. D., Carlsson, M., Uitenbroek, H., and De Pontieu, B. (2013b). The Formation of IRIS Diagnostics. II. The Formation of the Mg II h&k Lines in the Solar Atmosphere. *ApJ*, 772:90, 1306.0671.

Uitenbroek, H. (2001). Multilevel Radiative Transfer with Partial Frequency Redistribution. *ApJ*, 557:389–398.

Zwillinger, D. and Kokoska, S. (2000). *CRC Standard Probability and Statistics Tables and Formulae*.