MNRAS 000, 1–4 (2018) Preprint 23 January 2022 Compiled using MNRAS LATEX style file v3.0

Using Machine Learning to Predict the Correlation of Spectra Using SDSS Colour Magnitudes as an Improvement to the Locus Algorithm

Thomas O'Flynn [1] Kevin Nolan [1] Oisin Creaner [2] Eugene Hickey ⋆ [1]
[1] *Technological University Dublin, Tallaght, D24 FKT9, Dublin, Ireland*
[2] *Dublin Institute for Advanced Studies, 10 Burlington Rd, Dublin, D04 C932, Ireland*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The Locus Algorithm is a new technique to improve the quality of differential photometry by optimising the choices of reference stars. At the heart of this algorithm is a routine to assess how good each potential reference star is by comparing its sdss magnitude values to those of the target star. In this way, the difference in effect of the Earth's atmospheric scattering between target and reference can be minimised. This paper sets out a new way to calculate the calculate the quality of each reference star using machine learning. A random subset of stars from sdss with spectra was chosen. For each one, a suitable reference star was chosen, also with a spectrum. The correlation between the two spectra was taken to be the gold-standard measure of how well they match up for differential photometry. The five sdss magnitude values for each of these stars were used as predictors. A gradient boosting model was constructed on a training set of the stars and was evaluated on a testing set. The dataset used, the model construction, and performance evaluation is presented here.

Key words: Differential Photometry – Locus Algorithm – Machine Learning – SDSS

1 INTRODUCTION

The Locus Algorithm (Creaner *et al.* (2022)) has been used to create catalogues of pointings suitable for differential photometry on astromonical targets based on a novel technique of choosing appropriate reference stars (Creaner *et al.* (2020) and Creaner EXO's). The algorithm no longer places the target in the centre of the field of view but in general repositions it so as to include the best set of reference stars. Assessment of each reference star is performed by referring to the sdss catalogue and the colour band magnitudes therein. These magnitudes can be used to infer the overall shape of the star's spectrum. Stars that have similar spectra with be effected by scattering from the Earth's atmosphere to a more comparable degree that stars with dissimilar spectra. The original Locus Algorithm used a rational, but ad-hoc, method to estimate the correlation of stellar spectra based on differences between their g, r, and i sdss colour magnitudes (Creaner [Thesis] (2017)). This was necessary for computational efficiency. The work presented here presents a more rigorous technique to estimate the correlation of stellar spectra based on machine learning. The subset of stars in sdss that have their spectra measured are used. These stars are paired off such that each pair has similar colour magnitude differences and are thus potentially a good match for differential photometry. The correlation between each pairs spectra is calculated. This forms the basis of a goodness-of-fit between the two spectra. The sdss magnitudes (u, g, r, i, and z for both stars in the pair) are then used to train a machine learning algorithm to predict this goodness-of-fit. The model produced is then applied to other pairs of stars, the test set, to evaluate its performance. The results show a significant improvement over the original ad-hoc Locus Algorithm routine, this model will be incorporated to future generations of the Locus Algorithm.

⋆ eugene.hickey@tudublin.ie

2 DATA

This work uses 5591 stellar spectra from the SDSS SEGUE and BOSS observations and their physical parameters from the 13th SDSS data release (Aguado *et al.* (2018)). The spectra are clipped to just the wavelengths contained in the sdss r band (between 550nm and 700nm). Stars are paired off based on their sdss colour magnitudes so that both stars in a pair are of similar colour. Specifically, both $(g_1 - r_1) - (g_2 - r_2)$ and $(r_1 - i_1) - (r_2 - i_2)$ will be between 0 and 0.1. This ensures that these stars would be realistic matches for differential photometry. In addition, stars were chosen that had r colour magnitude values between 15 and 20. The SQL queries used to download physical parameters and the spectra are given in the supplementary materials for this paper. Correlations between spectra are calculated using the usual Pearson Correlation formula, equation (1).

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}. \quad (1)$$

where x_i refer to the flux from the first star in units of $erg/cm^2/s$, y_i the flux from the second star. Figure 1 shows some pairs of stars along with their correlations. The first pair, A and B, are representative of the sample, the second pair, C and D, were chosen to have an unusually low correlation for this sample set.

Correlation is usually bounded by -1 and 1. And because these are

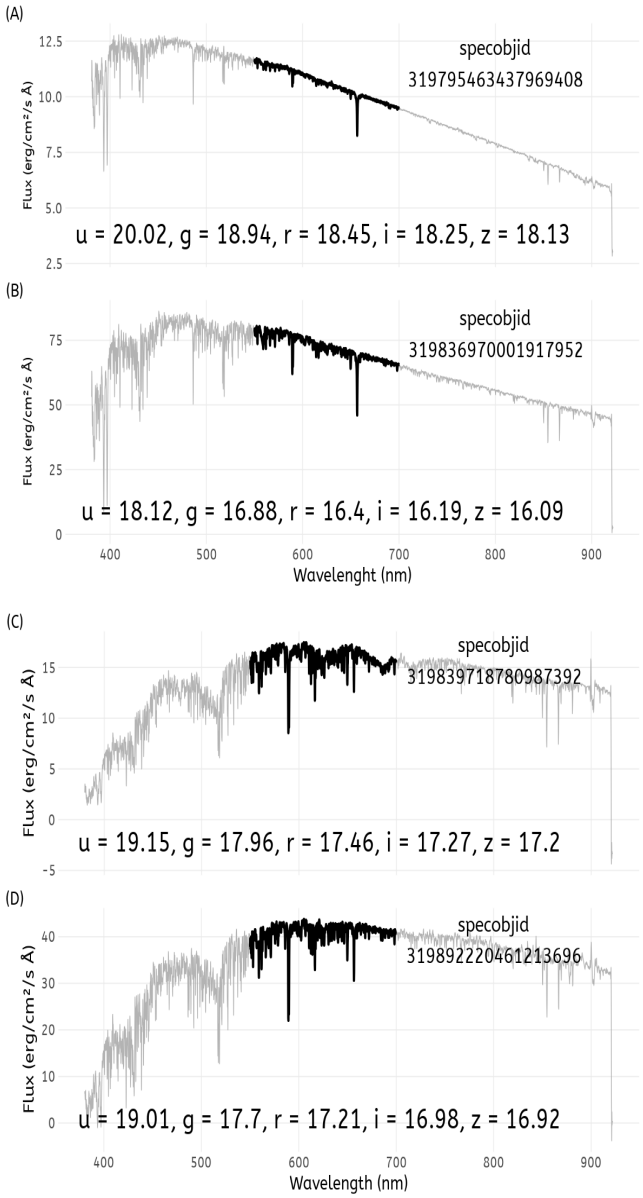

Figure 1. Two pairs of spectra downloaded from SDSS. The ugriz colour magnitudes for each star is given below its spectrum. The darkened area of the spectral line corresponds to the r-band wavelengths. The correlation between spectra A and C is 0.96. That between spectra B and D is 0.75.

spectra from stars and they have similar colour magnitudes, the correlations tend to be clustered near this higher end, see the histogram in figure 2A below. Machine learning algorithms work better with normally distributed values (need reference) and this is especially true when it comes to analysing model performance (another reference), so the correlation values were transformed. First of all by a logit transformation (2):

$$\text{logit}(x) = \ln\left(\frac{1+x}{1-x}\right) \tag{2}$$

And then by scaling and normalising the values to have a mean of 0 and a standard deviation of 1. The resulting transformed values are shown in figure 2B.

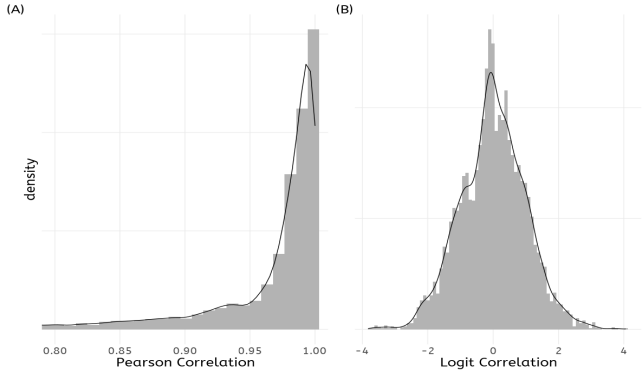

Figure 2. (A) Histogram of Pearson correlation values between r-band spectra between pairs of matched stars. (B) Values in (A) transformed by a logit function.

The data is split into test and training sets, with 70% of the data (3915 samples) in the training set and the remainder in the test set. Each set has a representative sample of correlation values, to do this the original sample of 5591 pairs is split into five groups based on percentiles of the correlation and test and train get a commensurate proportion of each group.

3 MODEL

A regression model was built on the training set, using the ugriz values for both stars in each pair as predictors for the logit(correlation) value. An eXtreme Gradient Boosting model (Friedman, Hastie and Tibshirani (2000), Chen and Guestrin (2021)) was used. This was chosen because of its performance and reliability (Bentéjac, Csörgő and Martínez-Muñoz (2020)). Cross validation was performed using a bootstrap method (Efron (1983)). The model was fit with the maximum number of boosting iterations set to 150, the learning rate set to 0.3, the maximum tree depth set to 3. It was set to minimise the RMSE on the training set. The final model fit was produced after 106 iterations.

One of the 150 trees produced is shown in figure 3.

4 MODEL EVALUATION

The model was then used to predict the logit correlation values from the 1676 star pairs from the test set. Figure 4 shows the resulting values of observed logit correlation values against predicted logit correlation values. Figure 5 shows the resulting values of the residuals, the observations minus the predicted values, against predicted logit correlation values. Figure 6 shows a histogram of the residual values, figure 7 shows a quantile-quantile (QQ) plot of the residuals. The shape of this last plot shows the residuals to be somewhat platykurtic which is acceptable for a machine learning fit (ref???).

The R^2 value of predicted logit correlation on the test set was found to be 71%. The RMSE on the test set was found to be 0.55. The performance of the original function used in Creaner *et al.* (2022) was worse, with an R^2 of 13%.

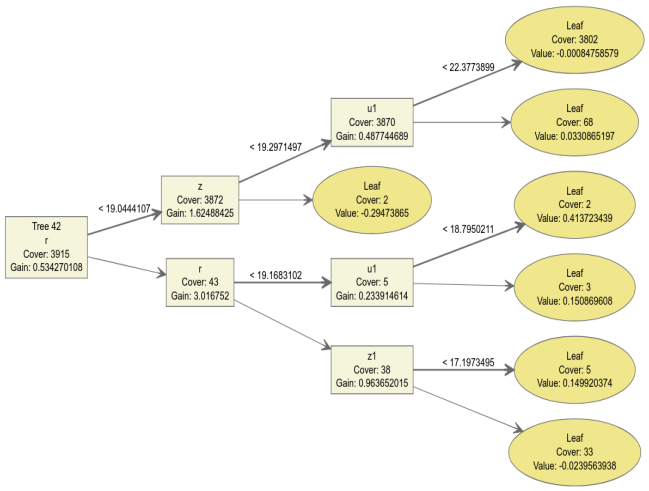

Figure 3. One of the 150 decision trees produced by the gradient boosting algorithm

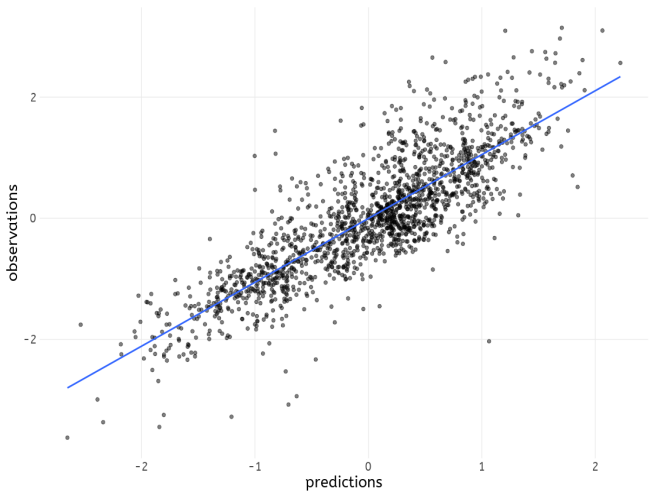

Figure 4. Observed versus predicted logit correlation values

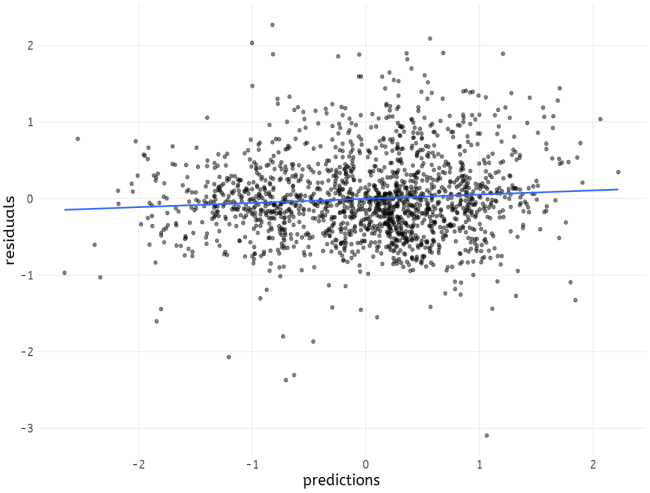

Figure 5. Observed versus predicted logit correlation values

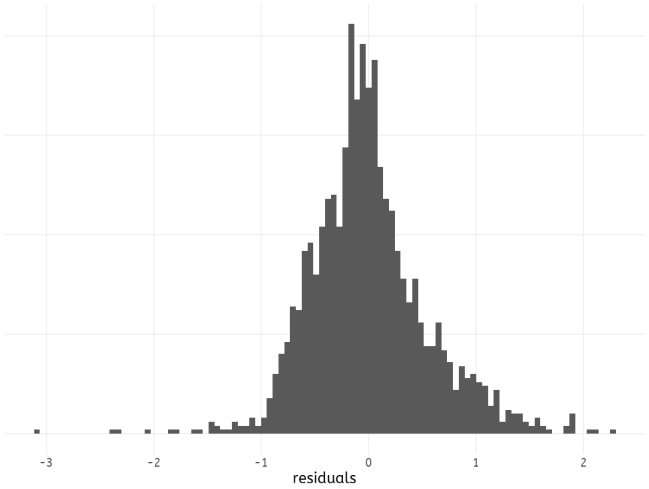

Figure 6. Observed versus predicted logit correlation values

5 CONCLUSIONS

The last numbered section should briefly summarise what has been done, and describe the final conclusions which the authors draw from their work.

REFERENCES

Aguado, D.S. *et al.* (2018) "The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA Derived Quantities, Data Visualization Tools and Stellar Library." doi:10.3847/1538-4365/aaf651.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2020) "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review 2020 54:3*, 54(3), pp. 1937–1967. doi:10.1007/S10462-020-09896-5.

Chen, T. and Guestrin, C. (2021) "Extreme Gradient Boosting [R package xgboost version 1.5.0.2]," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, pp. 785–794. doi:10.1145/2939672.2939785.

Creaner, O. *et al.* (2020) "A catalogue of Locus Algorithm pointings for optimal differential photometry for 23 779 quasars," *Monthly Notices of the Royal Astronomical Society*, 498(3), pp. 3720–3729. doi:10.1093/MNRAS/STAA2494.

Creaner, O. *et al.* (2022) "The Locus Algorithm: A novel technique for identifying optimised pointings for differential photometry," *Astronomy and Computing*, 38, p. 100537. doi:10.1016/J.ASCOM.2021.100537.

Creaner [Thesis], O. (2017) "Data Mining by Grid Computing in the Search for Extrasolar Planets," *Doctoral* [Preprint]. doi:https://doi.org/10.21427/7w45-6018.

Efron, B. (1983) "Estimating the error rate of a prediction rule: Improvement on cross-validation," *Journal of the American Statistical Association*, 78(382), pp. 316–331. doi:10.1080/01621459.1983.10477973.

Friedman, J., Hastie, T. and Tibshirani, R. (2000) "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, 28(2), pp. 337–407. doi:10.1214/AOS/1016218223.

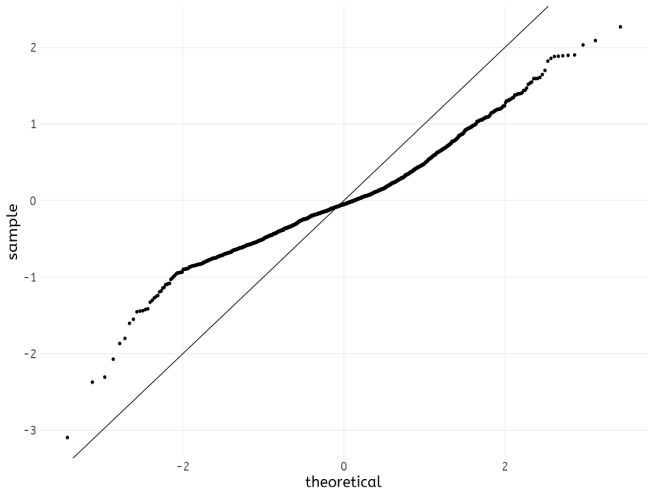

Figure 7. Observed versus predicted logit correlation values

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

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