Highlights of Spanish Astrophysics XII, Proceedings of the XVI Scientific Meeting of the Spanish Astronomical Society held on July 15 - 19, 2024, in Granada, Spain. M. Manteiga, F. González Galindo, A. Labiano Ortega, M. Martínez González, N. Rea, M. Romero Gómez, A. Ulla Miguel, G. Yepes, C. Rodríguez López, A. Gómez García and C. Dafonte (eds.), 2025

Gaia's XP spectra reddening removal using machine learning techniques

Pallas-Quintela, L.[1], Regueiro, Á.[1], Dafonte, C.[1], Manteiga, M.[2], Santoveña, R.[1], and Garabato, D.[1]

[1] CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Spain
[2] CIGUS CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Spain

Abstract

One of the problems that arises when dealing with stellar sources is reddening or extinction, which makes those spectra which have more energy in the blue part (hotter stars) move their energy to the red part of the spectra and potentially create confusion between the spectra of hot stars with high reddening values and cool stars. Nowadays, the Gaia mission publishes extinction values, but they are estimated just in 2 dimensions. For the more distant sources, such as galaxies and quasars, this approximation is enough. However, in case we want to remove reddening from the sources which are closer in our galaxy, it might cause reconstruction errors since extinction values for those sources are distance sensitive, so we would ideally need a 3D model capable of removing reddening for stars in our Galaxy. In this work, there are being introduced two methods that aim to remove extinction using machine learning techniques. More precisely, both denoising autoencoders and disentangling techniques are being tested. As the training dataset, we are using pairs of real spectra composed of sources with similar astrophysical parameters but reddening. In this way, we do not depend on extinction laws, and we work with spectra at different distances.

1 Introduction

Gaia is one of the most ambitious missions of the European Space Agency (ESA) by creating the most complete 3D map of our Galaxy; the Milky Way. The satellite was launched in 2013 and it is expected that it will observe around 2.5 billion stars, which is only 1% of the available objects in the Galaxy but will be translated in 1Petabyte of information. In addition to astrometry, we will have BP/RP (Blue and Red Photometers, respectively) of almost every observed source.

One of the main problems that arise when dealing with such spectra is reddening, which

is the distortion of light to the blue light of the sources due to the dust placed between the object and Gaia, in this case.

Our research group leads the Outlier Analysis (OA) [1] working package, which is devoted to performing unsupervised classification (clustering) of outlier sources, solely based on BP/RP, through Self-Organizing Maps (SOM). For such reason, it is necessary to develop a method for removing reddening to improve the quality of the classifications performed by OA. Even though Gaia already provides an estimator for reddening and extinction laws could be employed to unredden sources, such estimation is computed in 2D, which is adequate for the furthest objects, but may lead to the miss-reconstruction of nearby objects. In order to address this issue, two Machine Learning (ML) techniques were tested in an attempt to mitigate it; the so-called "denoising autoencoders" and "disentangled learning".

In the following sections, we will be explaining the dataset used to build the models and the techniques we used. Finally, we will show the results and sum up this paper with conclusions and future work.

2 Dataset and techniques

In this section, it will be explained both the methods and the dataset used for creating models that are able to remove reddening from BP/RP spectra.

2.1 Dataset

For this experiment, we are using BP/RP spectra from stars with effective temperatures between 5 000 and 10 000 K (A and B type stars), which are the most affected ones. First of all, it is necessary to build the baseline dataset, which is composed of A-B stars whose reddening value is residual. To decide which ones to use, there are selected those sources located at the fiducial line of a low-reddening Hertzsprung–Russell (HR) diagram, as the one in [2] and based on [3]. Afterwards, it is necessary to pick analogous stars but with identified reddening values. In this case, sources with similar astrophysical characteristics are selected, such as temperature or metallicity but with a different reddening value. More precisely, the selected A-B stars have reddening values in the range $(0-2]$, representing the most abundant range of reddening in our Galaxy.

Once the sources were selected, it was necessary to generate the externally calibrated spectrum of each of them. In GDR3, the community is able to use BP/RP spectra in three different ways: either directly as coefficients or as internally and externally calibrated spectra. As previously stated, there will be used using the externally calibrated ones, which were downloaded and generated using the `GaiaXPy`[4] Python package.

2.2 Techniques

Two different techniques are tested to accomplish our problem, denoising autoencoders and disentangled learning, which are explained below.

2.2.1 Denoising Autoencoders

It is a neural network technique which is able to reconstruct a noisy input to its clean version. For such, at the input, it needs pairs of <noisy,clean> data for performing the data reconstruction, as shown in Figure 1.

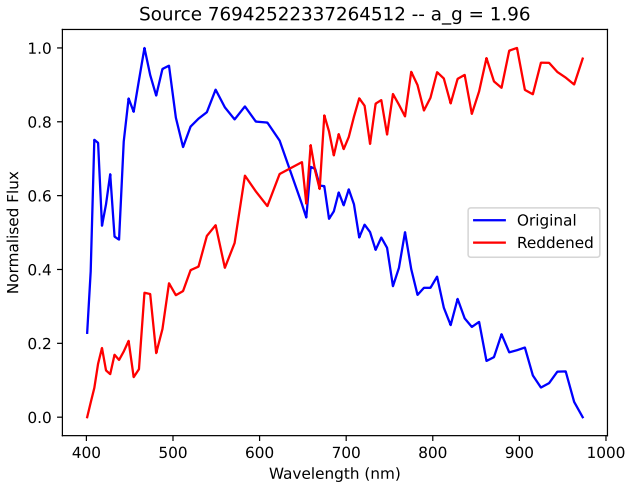

Figure 1: Example of a noisy vs clean BP/RP spectrum

Understanding the effect of interstellar dust as a factor that introduces a spurious signal into the spectrum, effectively adding "noise", it is possible to train a network so as to isolate this effect by providing training samples containing spectra both with and without reddening. In this way when the network is trained, it will learn the differences or noise between the clean and the noisy versions (spectra unaffected and affected by reddening). As a result, when the model is created and it receives noisy inputs, it will be possible for the model to generate its version without noise.

2.2.2 Disentangled Learning

Disentangled Learning is a technique based on Generative Adversarial Networks (GANs) which is able to remove properties of the input data by confronting two networks; the 1^{st} one attempts to deceive the 2^{nd} by making it believe that the generated inputs are real data.

This technique is widely used in images to add/remove glasses from people or change their hair, for example. However, with spectra, it is used to edit or remove (meaning remove generating a representation with a value equal to 0) its astrophysical properties, as in [5] and [6] (in preparation).

In this case, as an input, the network needs the noisy BP/RP spectra and the reddening value, which is the property that we need to alter from them. Even though it is possible to remove $1 \ldots n$ parameters, in this case, only extinction is taken into account.

3 Results

In this section, the results obtained with both networks will be analysed, where it is possible to observe that there are achieved similar results with any of them.

To create the models, around 1 million pairs of spectra were used. To train the disentangled learning model, it was used a Python tool named `GANDALF`[1], which was created for developing [6]. On the other hand, `PyTorch` was used to create the autoencoder model.

Figure 2 shows an example of a reconstruction of a source using autoencoders (left) and disentangled learning (right). Comparing the reconstructed spectra reveals a high degree of similarity, both in terms of the differences between the original and the reconstructed spectra and in terms of the algorithms. However, the final reconstruction is not an exact representation of the original source due to the generalization of the algorithm, which also reflects accurate training rather than a memorisation process of the training cases.

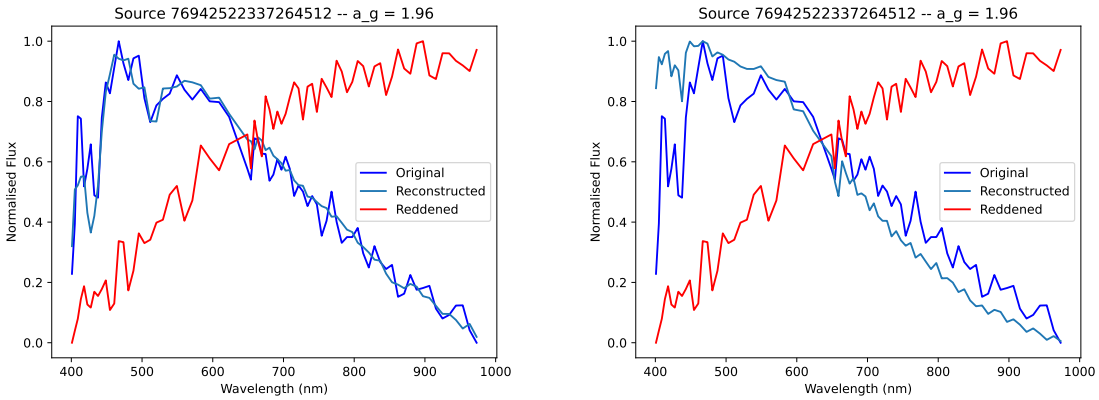

Figure 2: Result of denoising a source having $ag\_gspphot = 1.95$.
(Left) Denoising with the autoencoder and (Right) denoising with disentangling.

Figure 3 provides a more detailed representation of the results through the Mean Squared Errors (MSEs) with respect to the evolving reddening values. Each one of the lines (and their corresponding errors) shows the average difference between the clean and noisy pairs of spectra employed, as well as its evolution through the reddening values that were taken into account for the experiment. In orange it is represented such difference before applying the models and in blue or grey; after. As it can be observed, before any of the models is applied both the MSE and the error bar increase exponentially whereas, once the spectra are corrected, MSE stabilizes (although some small fluctuations are found throughout the reddening evolution).

[1] `https://github.com/raul-santovena/gandalf`

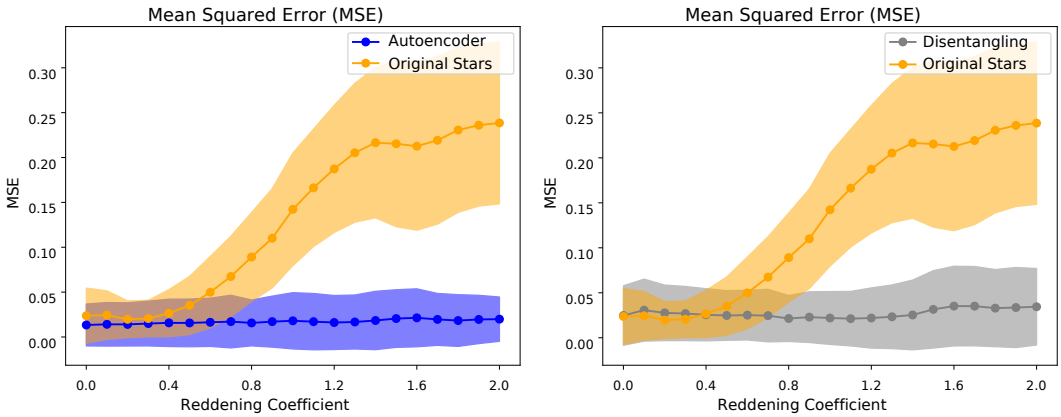

Figure 3: Evolution of the MSE before removing reddening (orange) and after (blue). (Left) Denoising Autoencoder and (Right) disentangling.

4 Conclusions and future work

In this work, there were have introduced two approaches in order to tackle the reddening problem in BP/RP spectra without using the classic extinction laws and trying to create a 3D model (sensitive to distance). Even though for creating the model using disentangled learning we are relying on the extinction values computed by Gaia (which are actually in 2D), we introduced spectra at different distances and the performance between both models is quite similar. Moreover, we were recently given access to the G-Tomo platform, which computes 3D extinction in Gaia and we can use those values for performing disentangling.

To improve our current models, on the one hand, we should compare the disentangling performance using both Gaia and G-Tomo extinction values. On the other, it is necessary to generalize the models by retraining them using any kind of stars. This is important because at OA we are using outlier data, whose astrophysical parameters are sensitive not to be correctly estimated. In this way, it will not be necessary to rely on any input temperature value to decide whether or not to remove the reddening.

Acknowledgments

This work has made use of data from the European Space Agency (ESA) Gaia mission and processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular, the institutions participating in the Gaia Multilateral Agreement.

This work has been funded by the Xunta de Galicia, through the PhD grant ED481A-2021/296, and grants ED431B 2021/36 and ED431B 2024/21. It has also been founded by the Spanish MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR through grant PID2021-122842OB-C22 and the Horizon Europe [HORIZON-CL4-2023-SPACE-01-71], SPACIOUS project funded under Grant Agreement no. 101135205.

We also acknowledge the support received from the Centro de Investigación de Galicia “CITIC”,

funded by Xunta de Galicia and the European Union (European Regional Development Fund-Galicia 2014-2020 Program), by grant ED431G 2019/01.

References

[1] Delchambre, L. et al. 2023, A&A, 674, A34

[2] Pallas, L., Garabato, D., Manteiga, M. and Dafonte, C., 2023, Highlights on Spanish Astrophysics XI

[3] Gaia Collaboration, Babusiaux, C., van Leeuwen, F., et al. 2018, A&A, 616, A10. doi:10.1051/0004-6361/201832843

[4] Ruz-Mieres, D. & zuzannakr, 2024, . gaia-dpci/GaiaXPy: GaiaXPy v2.1.2 (2.1.2). Zenodo. https://doi.org/10.5281/zenodo.11617977

[5] de Mijolla, D., Ness, M. K., Viti, S., et al. 2021, ApJ, 913, 12. doi:10.3847/1538-4357/abece1

[6] Santoveña, R., Dafonte, C. and Manteiga, M. (2023) preprint doi:10.2139/ssrn.4665534