Unsupervised Spectral Unmixing For Telluric Correction Using A Neural Network Autoencoder

Rune D. Kjærsgaard
DTU Compute
rdokj@dtu.dk

Aaron Bello-Arufe
DTU Space
aarb@space.dtu.dk

Alexander D. Rathcke
DTU Space
rathcke@space.dtu.dk

Lars A. Buchhave
DTU Compute
buchhave@space.dtu.dk

Line K. H. Clemmensen
DTU Compute
lkhc@dtu.dk

Abstract

The absorption of light by molecules in the atmosphere of Earth is a complication for ground-based observations of astrophysical objects. Comprehensive information on various molecular species is required to correct for this so called telluric absorption. We present a neural network autoencoder approach for extracting a telluric transmission spectrum from a large set of high-precision observed solar spectra from the HARPS-N radial velocity spectrograph. We accomplish this by reducing the data into a compressed representation, which allows us to unveil the underlying solar spectrum and simultaneously uncover the different modes of variation in the observed spectra relating to the absorption of H_2O and O_2 in the atmosphere of Earth. We demonstrate how the extracted components can be used to remove H_2O and O_2 tellurics in a validation observation with similar accuracy and at less computational expense than a synthetic approach with `molecfit`.

1 Introduction

Absorption of light in the atmosphere of Earth, called telluric absorption, can hinder astrophysical observations by partially obscuring the object of interest. Various methods have been introduced to remove the effects of this absorption from observed spectra. One such acknowledged method, called `molecfit` [15, 12], relies on computing a synthetic transmission spectrum of the atmosphere of Earth. Synthetic methods are inherently reliant on external factors to an observation, such as atmospheric measurements and molecular line lists. Another realm of methods take a data-driven approach attempting to exploit the modes of variation in a number of observed spectra to uncover the underlying components. By analysing such variation, the telluric absorption can be modelled without relying on external factors to an observation. One such approach, based on principal component analysis (PCA), has been explored in the literature [2]. PCA methods are however ineffective on very large data sets, where the entire data can not be stored in memory. Another approach called wobble [3] uses a linear model with a convex objective to model the telluric component of observed spectra.

We present a new data-driven approach using a neural network autoencoder. Autoencoders have seen use in the literature for decades [4, 10] and have long been known to discover effective compressed data representations through dimensionality reduction [9]. To demonstrate the approach we analyse 1257 observed solar spectra [7, 5] from the high-precision spectrograph HARPS-N [6] with the aim of disentangling the observed spectra into an underlying solar component and high accuracy telluric components from H_2O and O_2. The extracted components could aid in the detection of radial velocity signals of planetary systems by quickly and accurately removing tellurics from observed spectra, leading to an increase in observation quality and hereby a reduction in observing time and cost.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

2 Proposed approach

Our approach has roots in spectral unmixing, which seeks to unmix distinct endmember spectra and their weights from an observed spectral image by constructing a mixing model of the problem. Endmember unmixing from spectral data is a rich discipline with many existing approaches [16, 8]. We consider the solar spectrum and telluric spectrum as endmember spectra with associated abundance weights and use these components to construct a linear mixing model in log-space [17] describing the observed spectra:

$$\boldsymbol{x}_n = \sum_{r=1}^{R} w_{r,n}\boldsymbol{m}_r = \boldsymbol{M}\boldsymbol{w}_n + \boldsymbol{\epsilon}_n, \tag{1}$$

where $\boldsymbol{x}_n$ is the n^{th} observed spectrum from a finite set of N observed spectra, $\boldsymbol{m}_r$ is the r^{th} endmember spectrum of R endmembers with individual endmembers $r = 1, ..., R$. Furthermore, $w_{r,n}$ is the abundance of endmember r for observation n, $\boldsymbol{M}$ is the endmember matrix having endmembers as columns, $\boldsymbol{w}_n$ is the abundance vector of the n^{th} observation and $\boldsymbol{\epsilon}_n$ is an error term. The HARPS-N spectra cover the optical wavelength range and for this reason we consider the combined telluric spectrum to be comprised of the two strongest absorbing molecules in this region, namely H_2O and O_2. From this we get $R = 3$ with $r = 1$ representing the solar endmember, $r = 2$ representing the H_2O endmember and $r = 3$ representing the O_2 endmember.

The goal is to extract the endmember matrix $\boldsymbol{M}$ and abundance vector $\boldsymbol{w}_n$ by training the neural network on a set of preprocessed observed solar spectra, with the purpose of applying the extracted telluric spectrum to non-solar observations. The endmember matrix $\boldsymbol{M}$ is extracted for training regions with P pixels. Figure 1 shows a graphical representation of the approach.

In [14] they present an autoencoder for blind unmixing of hyperspectral images (HSI). We build on this idea by adapting the network architecture to the domain of astrophysical spectral data. This requires various structural changes and the introduction of several new constraints on the network.

The training data consists of 1257 observations of the solar spectrum all split into 69 apertures with $P = 4096$ pixels each. This gives a total of 69×4096 pixels per observation. We use solar data for training since this data has a very high signal to noise ratio and does not take away observing time from night time observations. The observations are from 2020 spanning a month of observations from October 22 through November 19. We filter out low flux and high airmass observations. Subsequently we interpolate all observations to a common wavelength grid, apply the natural logarithm and continuum normalise the spectra. The described procedures leave $N = 838$ spectra of which 50% are reserved for training and 50% are used for validation.

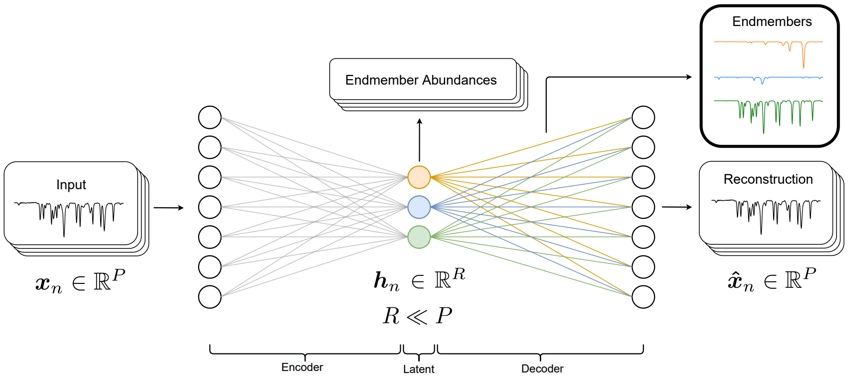
Endmembers
Endmember Abundances
Input
Reconstruction
$\boldsymbol{x}_n \in \mathbb{R}^P$
$\boldsymbol{h}_n \in \mathbb{R}^R$
$R \ll P$
$\hat{\boldsymbol{x}}_n \in \mathbb{R}^P$
Encoder
Latent
Decoder

Figure 1: The figure shows the architecture of the autoencoder. Observed spectra $\boldsymbol{x}_n$ are given as input and passed through the encoder into a lower dimensional latent space, which is subsequently decoded into the reconstruction $\hat{\boldsymbol{x}}_n$. After training by minimising the reconstruction error through a gradient descent algorithm, the endmember matrix $\boldsymbol{M}$ is extracted as the weights of the decoder and the abundance vector $\boldsymbol{w}_n$ is extracted as the latent representation $\boldsymbol{h}_n$. P is the number of pixels for each aperture in the observed spectrum. The network is illustrated for $R = 3$ endmembers representing the solar (orange, top), H_2O (blue, middle) and O_2 (green, bottom) endmembers.

2.1 Neural network autoencoder

The network uses fully connected layers and consists of an encoder function $\boldsymbol{h}_n = f(\boldsymbol{x}_n)$, which maps the input data $\boldsymbol{x}_n \in \mathbb{R}^P$ to an internal latent representation $\boldsymbol{h}_n \in \mathbb{R}^R$. This representation is then passed through the decoder function $g(\boldsymbol{h}_n) = \hat{\boldsymbol{x}}_n$, which seeks to reconstruct the input data $\boldsymbol{x}_n$ with the reconstruction $\hat{\boldsymbol{x}}_n \in \mathbb{R}^P$. The decoder is constructed without bias terms and performs the following affine transformation:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{h}_n, \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{P\times R}$ are the weights of the decoder, which are extracted as the endmember spectra $\boldsymbol{M}$, and $\boldsymbol{h}_n \in \mathbb{R}^R$ is the latent representation, which can be extracted as the endmember abundances $\boldsymbol{w}_n$.

The network features a utility layer responsible for normalisation of endmember abundance variation, which ensures a fixed abundance of the solar component, and a utility layer clamping the solar decoder weights to interval $[0, 1]$ and the telluric decoder weights to interval $[-1, 0]$. Moreover, the network contains a utility layer Doppler shifting the solar decoder weights to account for the spectral shift caused by the rotation and elliptical orbit of the Earth. We perform this shift using the barycentric Earth radial velocity of each observation. The network also features a batch normalisation layer [13, 11]. We use the validation set to determine hyperparameters based on a Tree-structured Parzen Estimator approach carried out with `optuna` [1]. Training is performed with stochastic gradient descent to minimise the mean squared error (MSE). The network is trained on non-stitched spectra for the 69 apertures separately to retain the high fidelity of the observed spectra and to avoid the complications involved in stitching spectra. Training on all apertures takes approximately 3 hours and 15 minutes on an Intel 6 core i7, UHD 630 CPU laptop.

3 Results

The extracted endmembers for aperture 60 are shown in Figure 2.

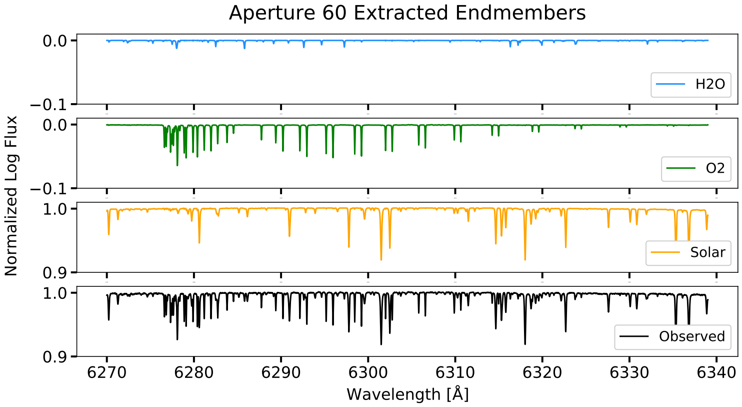

Figure 2: The figure illustrates extracted endmembers in addition to an observed solar spectrum for aperture 60. The extracted endmembers represent from top to bottom the H_2O (blue), O_2 (green) and solar (orange) components of the observed spectrum (black, bottom).

To validate extracted endmembers we perform telluric correction using our autoencoder tellurics and compare with a `molecfit` synthetic telluric spectrum. Telluric correction aims to remove tellurics by dividing the observed spectrum with a telluric transmission spectrum. We perform the comparison by correcting a solar observation with strong tellurics from HARPS-N, which the autoencoder has not been trained on. We compute the `molecfit` telluric spectrum on a HPC cluster using a 10 core Intel Xeon E5-2660v3, Huawei XH620 V3 node and utilise atmospheric measurements from the time of the observation in addition to a fit to the stitched version of the observation from the HARPS-N pipeline. We compute the autoencoder correction using the extracted telluric components, which have been converted back from log-space to represent standard transmission spectra. Autoencoder telluric abundance weights are found using a least squares fit to known telluric lines in the spectrum. We interpolate the observed spectrum to the telluric wavelength axes of `molecfit` and the autoencoder before the corrections. The comparison can be seen in Figure 3.

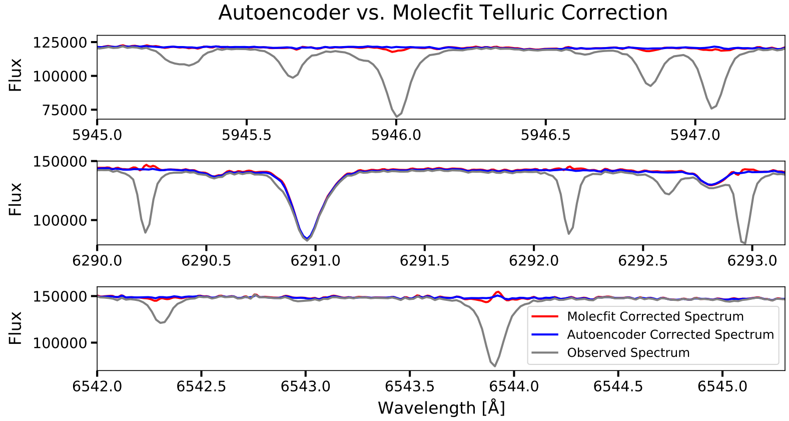

Figure 3: The figure displays a comparison of corrected spectra computed either using autoencoder extracted tellurics or `molecfit` tellurics for H_2O and O_2 lines in three spectral regions of high interest to the study of exoplanetary atmospheres. For pure telluric lines, an ideal correction would result in a flat line. In the middle plot, the autoencoder and `molecfit` agree on the presence of solar features around 6290.5 Å, 6291 Å and 6292.75 Å, which remain in the corrected spectra.

4 Discussion

High-resolution spectroscopy is limited to ground-based observations, and consequently high-resolution ground truth solar spectra are not available. This makes evaluating the performance difference between the two approaches complicated. A natural area of comparison however lies around corrected tellurics, where an accurate correction will leave no trace of the telluric absorption. As illustrated in Figure 3, the autoencoder correction removes tellurics in the observed spectrum to continuum level, while `molecfit` leaves slight traces of the correction. The sinusoidal shape of the `molecfit` correction around 6544 Å in Figure 3 could indicate imprecision on the exact wavelength location of the telluric line centre used in `molecfit`. The `molecfit` correction takes approximately 30 minutes to compute, while the autoencoder correction takes about 0.2 seconds. This difference in compute time is significant and makes the autoencoder much more feasible for correction of multiple spectra. Additionally, the autoencoder approach can be used as a complementary data-driven validation tool to inspect the accuracy of synthetic approaches like `molecfit`.

While the results show that the autoencoder correction of the validation observation is performed with similar accuracy to `molecfit`, the question of how well the extracted endmembers generalise to observations dissimilar to the training data still remains. Future work includes exploring this potential limitation by performing corrections on numerous non-solar observations and inspecting the impact on radial velocity extraction, as well as the consistency of retrieved exoplanetary atmosphere signals.

This paper demonstrates the approach applied to the HARPS-N spectrograph, but the telluric autoencoder is designed as a general tool, which can be trained on solar data from any spectrograph and wavelength range and subsequently perform corrections on new observations (including non-solar observations) from the given spectrograph. This is an advantage as solar data is already gathered and ready to train on for many spectrographs.

5 Conclusion

We have demonstrated an approach for computing a compressed representation of the solar data from HARPS-N with a constrained neural network autoencoder. This representation can be used to extract endmembers that directly relate to the solar spectrum as well as the transmission spectra of H_2O and O_2. After the autoencoder representation has been computed for a given spectrograph, the extracted components can be used to perform very quick and accurate telluric correction on any observations from the same spectrograph. The autoencoder approach and the detailed extracted telluric spectrum could aid in the detection of faint radial velocity signals and atmospheric features of Earth analogue exoplanets observed from ground-based telescopes.

References

[1] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. [2019], 'Optuna: A next-generation hyperparameter optimization framework [arxiv]', *Arxiv* p. 10 pp.

[2] Artigau, É., Astudillo-Defru, N., Delfosse, X., Bouchy, F., Bonfils, X., Lovis, C., Pepe, F., Moutou, C., Donati, J.-F., Doyon, R. et al. [2014], Telluric-line subtraction in high-accuracy velocimetry: a pca-based approach, *in* 'Observatory Operations: Strategies, Processes, and Systems V', Vol. 9149, International Society for Optics and Photonics, p. 914905.

[3] Bedell, M., Hogg, D. W., Foreman-Mackey, D., Montet, B. T. and Luger, R. [2019], 'wobble: a data-driven analysis technique for time-series stellar spectra', *The Astronomical Journal* **158**(4), 164.

[4] Bourlard, H. and Kamp, Y. [1988], 'Auto-association by multilayer perceptrons and singular value decomposition', *Biological cybernetics* **59**(4), 291–294.

[5] Dumusque, X., Cretignier, M., Sosnowska, D., Buchschacher, N., Lovis, C., Phillips, D., Pepe, F., Alesina, F., Buchhave, L., Burnier, J. et al. [2021], 'Three years of harps-n high-resolution spectroscopy and precise radial velocity data for the sun', *Astronomy & Astrophysics* **648**, A103.

[6] *HARPS-N Instrument Page* [n.d.], `http://www.tng.iac.es/instruments/harps/`.

[7] *HARPS-N Published Solar Observations* [n.d.], `https://dace.unige.ch/dashboard/`.

[8] Heylen, R., Parente, M. and Gader, P. [2014], 'A review of nonlinear hyperspectral unmixing methods', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(6), 1844–1868.

[9] Hinton, G. E. and Salakhutdinov, R. R. [2006], 'Reducing the dimensionality of data with neural networks', *science* **313**(5786), 504–507.

[10] Hinton, G. E. and Zemel, R. S. [1994], 'Autoencoders, minimum description length, and helmholtz free energy', *Advances in neural information processing systems* **6**, 3–10.

[11] Ioffe, S. and Szegedy, C. [2015], Batch normalization: Accelerating deep network training by reducing internal covariate shift, *in* 'International conference on machine learning', PMLR, pp. 448–456.

[12] Kausch, W., Noll, S., Smette, A., Kimeswenger, S., Barden, M., Szyszka, C., Jones, A., Sana, H., Horst, H. and Kerber, F. [2015], 'Molecfit: A general tool for telluric absorption correction-ii. quantitative evaluation on eso-vlt/x-shooterspectra', *Astronomy & Astrophysics* **576**, A78.

[13] LeCun, Y. A., Bottou, L., Orr, G. B. and Müller, K.-R. [2012], Efficient backprop, *in* 'Neural networks: Tricks of the trade', Springer, pp. 9–48.

[14] Palsson, B., Sigurdsson, J., Sveinsson, J. R. and Ulfarsson, M. O. [2018], 'Hyperspectral unmixing using a neural network autoencoder', *IEEE Access* **6**, 25646–25656.

[15] Smette, A., Sana, H., Noll, S., Horst, H., Kausch, W., Kimeswenger, S., Barden, M., Szyszka, C., Jones, A., Gallenne, A. et al. [2015], 'Molecfit: A general tool for telluric absorption correction-i. method and application to eso instruments', *Astronomy & Astrophysics* **576**, A77.

[16] Somers, B., Asner, G. P., Tits, L. and Coppin, P. [2011], 'Endmember variability in spectral mixture analysis: A review', *Remote Sensing of Environment* **115**(7), 1603–1616.

[17] Zhao, H. and Zhao, X. [2019], 'Nonlinear unmixing of minerals based on the log and continuum removal model', *European Journal of Remote Sensing* **52**(1), 277–293.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our approach incurs low computational costs when performing numerous corrections. This results in a net societal benefit compared to the continued use of more computationally expensive methods like `molecfit`.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The project is a work in progress. The code and data will be published when the project is complete.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 2.1, Section 3 and Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 1.
   (b) Did you mention the license of the assets? [No] The HARPS-N solar data is public. We reference both the paper detailing the data release [5] as well as the site to download the data [7].
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] The project is a work in progress. The code and data will be published when the project is complete.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The HARPS-N solar data is public and available for everyone.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The astrophysical data does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]