ESCAPE Task 4.3

Progress Meeting, December 16-17, 2019

Notes on the Deep Learning analysis of HARPS spectra

History:
2020-01-10, M. Romaniello: creation
2020-01-21, M. Romaniello: incorporation of feedback from participants

Table of Contents

1. Preamble: scope of the document

These notes aim at capturing the status on the analysis of the HARPS spectra, as presented and discussed during the ESCAPE Task 4.3 progress meeting held at ESO HQ on December 16th and 17th, 2019.

Meeting attendance

Mark Allen (CDS), Jon Carrick (ESO/Lancaster), Antonio D'Isanto (HITS), Francois-Xavier Pineau (CDS, remote), Kai Polsterer (HITS), Martino Romaniello (ESO), Nima Sedaghat (ESO), Michael Sterzik (ESO), Felix Stoehr (ESO).

2. The tools

The networks

Two network designs, one developed at HITS and one at ESO, were presented. They analyze with Deep Learning techniques the HARPS processed spectra as retrieved from the ESO Science Archive. The dataset consists of about 275,000 individual spectra of about 7,000 individual targets.

Unique targets

The main science case of HARPS is to discover and follow-up extra-terrestrial planets via the periodic radial velocity shift they imprint on the host star. This requires repeated observations of the same targets. Both HITS and ESO developed tools to group the individual spectra by target. They are based on clustering of spatial coordinates and, to a lesser extent, target names, both as recorded in the file headers.

Both methods return very similar results, with a list of about 7,000 unique targets. This provides a very important sample to cross-check the results of the analysis results.

Visualization

HITS developed a prototype to visualize and interact with the output results from their network. A movie illustrating its main capabilities is available at: https://we.tl/t-G9LYPxDyhe. The prototype currently consists of python scripts. It will be evolved to a Jupyter notebook by HITS/CDS to allow for convenient communal access, which was identified as an important feature for progress.

3. The HITS network

The spectra are pre-processed to remove the overall signal DC level (constant) and sub-sampled to 11,000 spectral channels (out of the original 300,000). The spectra are then fed to an autoencoder network with a 2-dimensional latent space. The autoencoder is a fully-connected unsupervised neural network, which is meant to projects the input spectra to the lower dimension latent space, and then to reconstruct the inputs to their initial shape.

The architecture is kept small and agile in order to preserve speed and interactivity. For this reason, only three hidden layers, with a limited number of neurons, are used.

Similarly, the main driver to choose 2 dimensions is to allow for easy and intuitive visualization of the results and their exploration. Also, the reduced spectra sampling makes for extremely fast execution times, allowing on-the-fly re-training, e.g. with different loss functions. In its current status, the prototype can be trained and used for data visualization and inspection on a common laptop with an Intel I7 CPU. In this case, a randomly selected spectrum per source is used. The model shown during the meeting instead was based on a pre-trained version of the architecture, performed using all the available sub-sampled spectra, on a Nvidia Pascal P40 GPU. The training takes few seconds on the unique target catalog, and less than a hour for the entire catalog on GPU.

Both features are exploited in the visualization prototype, which convincingly demonstrates the potential of connecting different aspects of the data.

Furthermore, other dimensionality reduction models are currently available in the prototype, in order to compare their performance with the autoencoder, namely PCA and GPLVM.

Considerations

A cursory look at the reconstructed spectra indicates that the overall shape of the continuum and the broad absorption lines of the Balmer series are well reproduced in the reconstructed spectra, while the narrow lines are not. The points in the latent space define a rather tight relationship. In many (the majority?) of cases, the repeated observations of the same object lie along this sequence, but their locations display a large spread.

Overall, it is not clear what are the features in the spectra that drive the similarity in the latent space. The fact that repeated observations of the same object do not generally fall close to each another in the latent space indicates that the similarity may not be mainly based on the intrinsic characteristics of the sources (there can, of course, be sources that legitimately wander around in the latent space because of bona-fide intrinsic variability, but they are reasonably to be expected to be a minority). For example, the shape of the continuum is determined by the superposition of the intrinsic stellar spectrum to extrinsic nuisance effects (interstellar absorption along the line of sight, absorption due to the Earth atmosphere, telescope and instrument response curves and differential loss of light at the entrance of the HARPS fiber). In its current status, these latter seem to play an important role in the determination of the similarity.

As mentioned above, the driver for a 2D latent space dimensionality is data exploration, not the desire to capture the complexity within the data. As such, it does not allow the network to account for and disentangle the different possible effects. In this setup, this is to be done by pre-processing the data before they are fed to the network. HITS will explore this by extending the pre-processing to, e.g., correcting for atmospheric absorption, or normalizing out the continuum shape altogether.

In general, care and insight have to be exercised in order to ensure that the information that is captured by the network is relevant for the intended purpose.

HITS will then move on to feed their network with UVES spectra, which they have already retrieved from the ESO archive.

4. The ESO network

Original spectra, with the full spectral sampling of about 300,000 wavelength channels are used. The latest version normalizes the spectra by the median of the flux values; however, this is not a requirement as networks trained with non-normalized spectra have shown to work equally well. Unsuitable spectra are detected and removed from the original dataset in advance. Such samples are identified based on a) non-numeric data points, b) non-stable spectra, e.g. showing extreme variations in flux, and c) extremely low SNR estimates.

The motive has been to allow the network to show us the highest degree of compression we can apply, while still being able to reconstruct the spectra to a satisfactory level. Therefore, we explored various dimensionalities for the latent representation, from 4 to 8192. The idea is that learning such an efficient feature representation can be used to identify meaningful similarities.

Visualization of such a high-dimensional space is not trivial. This, however, is not the main focus of ESO's approach, where the main objective has been to come up with a minimal representation which preserves all the "useful" information.

The networks use a combination of Convolutional and Fully-connected layers, allowing for *deep* feature extraction: more than 30 layers. Obviously, the approach is rather computer-intensive, with about one week required to train one instance of the network on the current ESO hardware (Nvidia TITAN RTX + >32 GB RAM).

Considerations

Networks with latent space dimensionalities of 8 or above reconstruct the large majority of common spectral features. In practice, the shape of the continuum and individual absorption features, large and small, are overwhelmingly reproduced. At the same time, the surprising observation is that in addition to rejecting noise, the network naturally rejects tellurics lines too (details further below).

The next steps include trying to make (some) sense of the latent space e.g. by disentangling it and figuring out which dimensions carry useful information for the intended scope, which should help to alleviate the visualization of the results, at least partially. The extent to which this is even possible is an important piece of information in how to present the results to users in ways that they can understand and use for their indented purposes. Also, ESO will complete a prototype proximity service based on the output of its network, which was presented to the meeting in a very preliminary form.

Telluric lines

An early observation triggered by the inspection of the output of the ESO network is that, while the majority of spectral features are well reconstructed by the network, there is a fair number that is not. Their visual inspection indicated that they are likely largely be telluric absorption features.

This first ansatz was confirmed by a detailed analysis based on the molecfit suite (http://www.eso.org/sci/software/pipelines/skytools/molecfit), which provides an accurate, independent expectation for the telluric imprint on each individual spectrum. It does so by modelling the telluric features using detailed molecular physics coupled with the profile of the physical conditions of the Earth atmosphere at the time and place of the observations, and solving the corresponding radiative transport equations.

The molecfit results for each individual spectrum were compared to the ones from the network, confirming that the latter is overwhelmingly not able to reconstruct the telluric lines (while doing so for the stellar ones). A metric is being finalized to quantify this effect in an objective way and, e.g. cross-correlate with the characteristics of the network. A publication is also being worked on. It is led by Nima and includes all the participants as authors.

The reasons why the ESO network is able to distinguish between the two sets of features, intrinsic and telluric, were intensely debated, and probably will continue to be.

5. Simulations

In order to quantitatively explore the limits to which the networks are able to characterize the data, ESO proposed to carry out accurate simulations of the data. The steps involved for HARPS are illustrated in Figure 1.

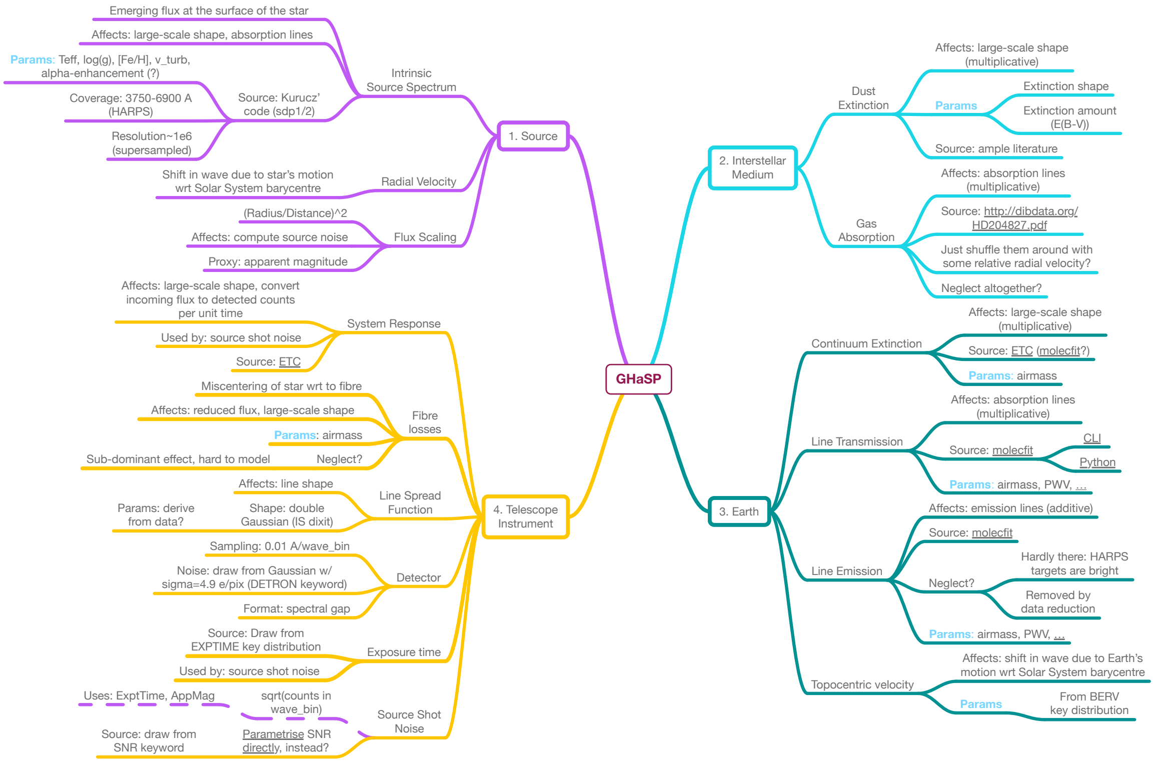
GHaSP
1. Source
Intrinsic Source Spectrum
Emerging flux at the surface of the star
Affects: large-scale shape, absorption lines
Source: Kurucz' code (sdp1/2)
Params: Teff, log(g), [Fe/H], v_turb, alpha-enhancement (?)
Coverage: 3750-6900 A (HARPS)
Resolution~1e6 (supersampled)
Radial Velocity
Shift in wave due to star's motion wrt Solar System barycentre
Flux Scaling
(Radius/Distance)^2
Affects: compute source noise
Proxy: apparent magnitude
2. Interstellar Medium
Dust Extinction
Affects: large-scale shape (multiplicative)
Params
Extinction shape
Extinction amount (E(B-V))
Source: ample literature
Gas Absorption
Affects: absorption lines (multiplicative)
Source: http://dibdata.org/HD204827.pdf
Just shuffle them around with some relative radial velocity?
Neglect altogether?
3. Earth
Continuum Extinction
Affects: large-scale shape (multiplicative)
Source: ETC (molecfit?)
Params: airmass
Line Transmission
Affects: absorption lines (multiplicative)
Source: molecfit
CLI
Python
Params: airmass, PWV, ...
Line Emission
Affects: emission lines (additive)
Source: molecfit
Neglect?
Hardly there: HARPS targets are bright
Removed by data reduction
Params: airmass, PWV, ...
Topocentric velocity
Affects: shift in wave due to Earth's motion wrt Solar System barycentre
Params
From BERV key distribution
4. Telescope Instrument
System Response
Affects: large-scale shape, convert incoming flux to detected counts per unit time
Used by: source shot noise
Source: ETC
Fibre losses
Miscentering of star wrt to fibre
Affects: reduced flux, large-scale shape
Params: airmass
Neglect?
Sub-dominant effect, hard to model
Line Spread Function
Affects: line shape
Shape: double Gaussian (IS dixit)
Params: derive from data?
Detector
Sampling: 0.01 A/wave_bin
Noise: draw from Gaussian w/ sigma=4.9 e/pix (DETRON keyword)
Format: spectral gap
Exposure time
Source: Draw from EXPTIME key distribution
Used by: source shot noise
Source Shot Noise
sqrt(counts in wave_bin)
Uses: ExptTime, AppMag
Parametrise SNR directly, instead?
Source: draw from SNR keyword

Figure 1 Schematics of the Great HARPS Simulation Project.

It was agreed that this was a valuable thing to do and that ESO will proceed to do so.