Final Project
Adam Wheeler (ajw2207)
December 21, 2019

1 Introduction

In this work, I pursue a model that uses a high-resolution spectral survey (a set of spectra taken by the same instrument with the same reduction pipeline) to learn a low-dimensional representation, z of spectra from a low-resolution spectral survey. This scheme has the advantage that the structure of the model (detailed below) enforces that z is minimally influenced by the peculiarities of either instrument.

Such a low-dimensional representation might be useful to denoise[1] spectra, or as a basis from which to infer stellar parameters. Stellar spectra are very often used for the determination of chemical abundances (roughly: the logarithm of the fraction of atoms of a given element in a star's atmosphere), which are interesting in their own right, but also useful as the only observable that links a star to its birth conditions. *Chemical tagging*, the clustering of stars in abundances space to reconstruct dissolved star clusters, is an as-yet-unrealized desideratum in the study of the Milky Way. One reason for the difficulty of this project is large systematic uncertainty in measured stellar abundances. Abundances might be more easily inferred for from a low-dimensional space, or clustering might be done directly on a projection of it.

2 Model

Assume there are two spectroscopic surveys of stars in the Milky Way. Survey 1 has high-resolution spectra, while survey 2 has low-resolution spectra, presumably of a larger number of stars or of stars of particular interest. Let $F^1_{1:n}, F^2_{1:n}$ be the spectra from each, for only the stars observed by both. They are taken to be vectors of fluxes evaluated over a fixed set of rest-frame wavelengths. Each F^2_i has an associated error vector E^2_i. For now, I will pretent that F^1 is measured perfectly. I use a model that factorizes like this:

$$p(F^2_{1:n}, \theta_z, \theta_F | F_1) = p(\theta_z) p(\theta_F) \prod_{\text{stars } i} p(F^2_i | \theta_z, \theta_F). \tag{1}$$

The generative model is as follows: Define $\theta_x = (\beta_x, b_x)$ to be the Jacobian and intercept of a linear function. Each element of each b and β is drawn from a unit normal[2], and the likelihood function is given by

$$F^2_i \sim \mathcal{N}_m(\beta_F \beta_z F^1_i + b_F, E^2_i) \tag{2}$$

where m is the length of each F^2_i. (b_z is fixed to the zero vector because it would be redundant.)

[1] I'm not certain that I'm using this term correctly. I'm referring to approximately projecting a spectrum onto the manifold of "perfectly measured" spectra by removing both shot noise and non-astrophysical structure.

[2] I realized too late that this is a bit of a silly prior for b_F, since the spectra have flux roughly equal to one in the absence of emission or absorption features. $\mathcal{N}(1, 1)$ would make more sense, but I don't think error strongly effects anything in this the report.

Here $z_i = \beta_z F_i^1$ is not instantiated as a latent variable. Requiring each z_i (a length-c vector) to be determined exactly from each F_i^1 is what prevents it from capturing the peculiarities of survey 2. Since c is small compared the length of each F_i^1 and F_i^2, this model is linear regression constrained to estimate a matrix of constrained rank[3].

My use of a linear model is not motivated entirely by computational ease. While linear models of spectra are known to be sub-optimal for parameter estimation, many spectral features grow roughly linearly with underlying stellar parameters.

3 Data

In my case, survey 1 is APOGEE (data-release 15), an infrared survey of a few $\times 10^5$ stars. Survey 2 is LAMOST (data-release 4 v2), an optical survey of a few $\times 10^6$ stars. There is no overlap in wavelength between these surveys. They have 3591 stars with very high S/N (LAMOST `snrz` and APOGEE `SNR`) in common. I removed stars flagged by APOGEE as problematic, those missing pixels in their APOGEE spectrum, and those for which stellar parameters were not available (a sign that the spectrum is in some way pathological). This left 1109 stars.

The spectra are pseudo-continuum-normalized (the blackbody emission is removed, leaving only narrow features).

4 Results

I calculated maximum *a posteriori* (MAP) estimates of model parameters for several values of c via stochastic optimization (`pytorch`'s `AdamW` implementation). Figure 1 shows the log joint as a function of training epoch for $c = 2, 10, 50$, and Figure 2 shows the resultant log joint probability of the estimated latent parameters with held-out data. I have only optimized ocne for each c value, so much of the difference in performance between versions with small Δc is presumably due to randomized initial conditions. Nevertheless, we can see that $c \gtrsim 10$ is strongly preferred, which concords nicely with theoretical expectations, since stellar atmospheres have at least a few easily observable distinct parameters. Generalization roughly increases with c all the way up to $c = 500$, the largest value I tried. This suggests to me that a more flexible model is may be in order if the goal is for z to be small (see however §5).

Given estimates of β_F and b_F, I can investigate $p(z_i|F_i^2)$. The maximum likelihood estimate (MLE) of z can be calculated very efficiently via the normal equation. For small c, the likelihood will domniate the prior and the MLE will be close to the MAP estimate. Figure 3 shows the error-relative difference between an arbitrary held-out spectrum and its denoised form, $\widehat{F^2} = \widehat{F^2}(\hat{z}(F^2))$ for $c = 2, 10, 50$. Note that the number of very narrow deviations decreases with c. I hypothesize that these are the locations of strong lines of elements whose abundances aren't captured by z. Figure 4 shows two spectra from the held-out set with aberant non-stellar features that aren't present in their denoised counterparts. It works! I looked at 50 spectra in the held-out set and didn't see any obvious

[3]I suspect that a form of this idea is already in the statistics literature. If you know what name it goes by, I'd love to know.

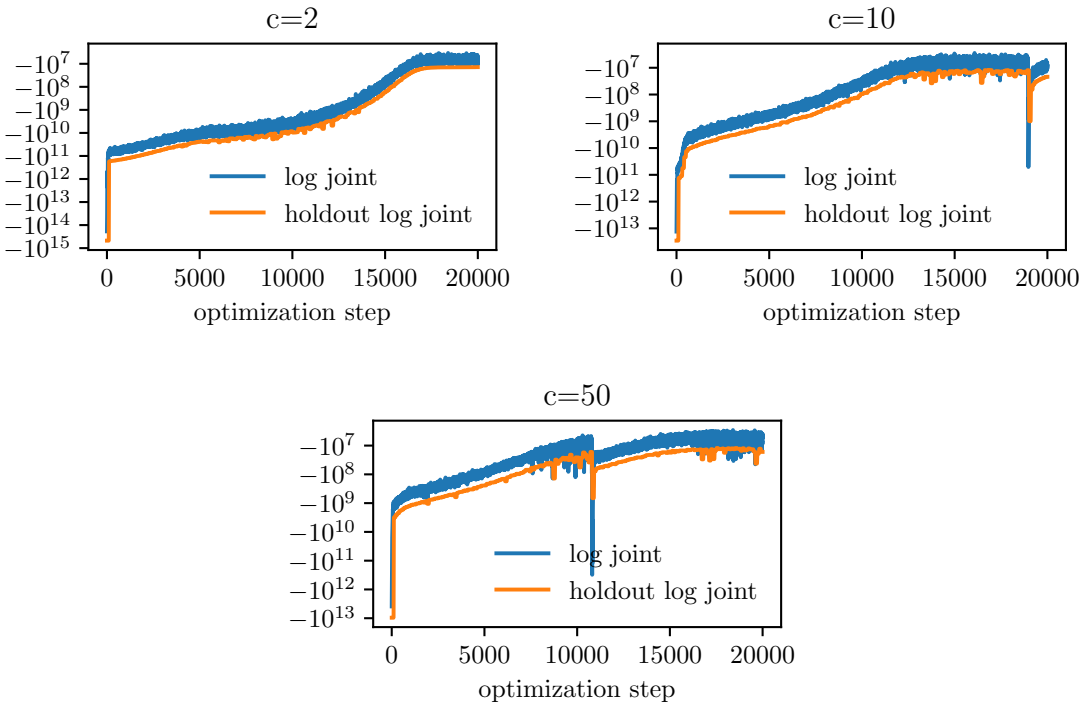

Figure 1: Log joint probability of latent parameters with training and held-out data as a functino of epoch. Continuing the optimization for 80,000 more epochs did not result in a higher joint probability. Apparently, sudden itermittent decreases are somewhat common when optimizing with `Adam` and its variants. Despite these, convergence was still reached more quickly than when using `AdaGrad` or standard SGD. They are caused by an "unlucky" minibatch. When using the whole dataset for every batch (i.e. using non-stochastic gradient ascent) they are not present.

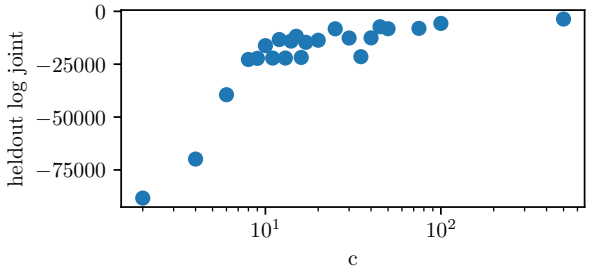

Figure 2: held-out log joint probability for MAP estimates of latent variables for various values of c.

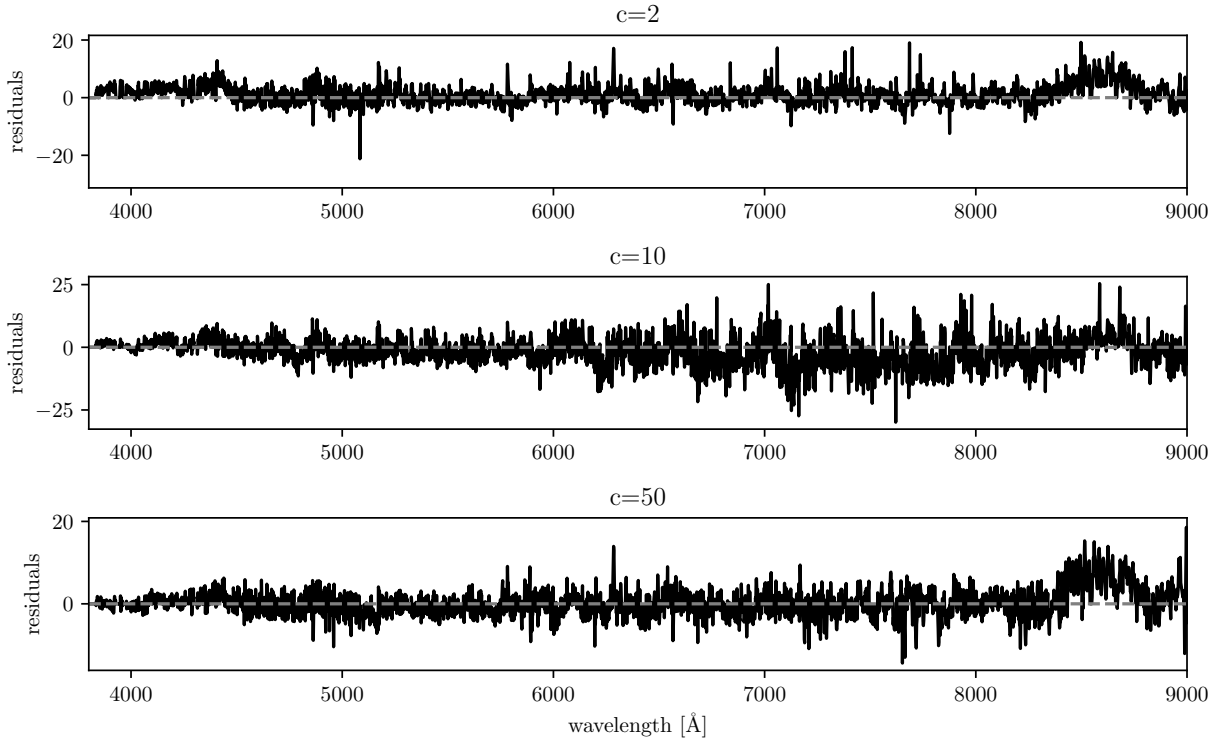

Figure 3: Error-relative residuals, $(F^2 - \widehat{F^2})/E^2$, for an arbitrary LAMOST spectrum for $c = 2, 10, 50$.

non-stellar features that weren't removed by denoising. It remains to be seen if spectra denoised with this model are still suitable for a given downstream analysis, but at the very least the model provides way to identify aberant spectra.

5 Model variants

The model presented above is simple. While working on this project, I trialled more sophisticated variants, but none were fruitful. Particularly, I sought to infer stellar labels ℓ_i (effective temperature, surface gravity, abundances) from each LAMOST spectrum though a model that factorizes like this:

$$p(F^2_{1:n}, L, \theta_z, \theta_F, \theta_\ell | F_1) = p(\theta_z)p(\theta_F)p(\theta_\ell) \prod_{\text{stars } i} p(F_i^2|\theta_z, \theta_F, F^1)p(\ell_i|\theta_z, \theta_\ell). \quad (3)$$

A linear model analogous to the one discussed above can predict labels, but with poor precision, even for large c.

I also explored replacing the linear transformations with neural networks. Unfortuntely, this approach didn't improve on the linear model in any way. I'm not surprised that the more flexible model no better at denoising spectra, but I don't understand why it completely failed to predict stellar labels (even for shallow and narrow network architecture without a huge number of weights). I suspect that a bug in my code is to blame, although it's possible that my training set is too small.

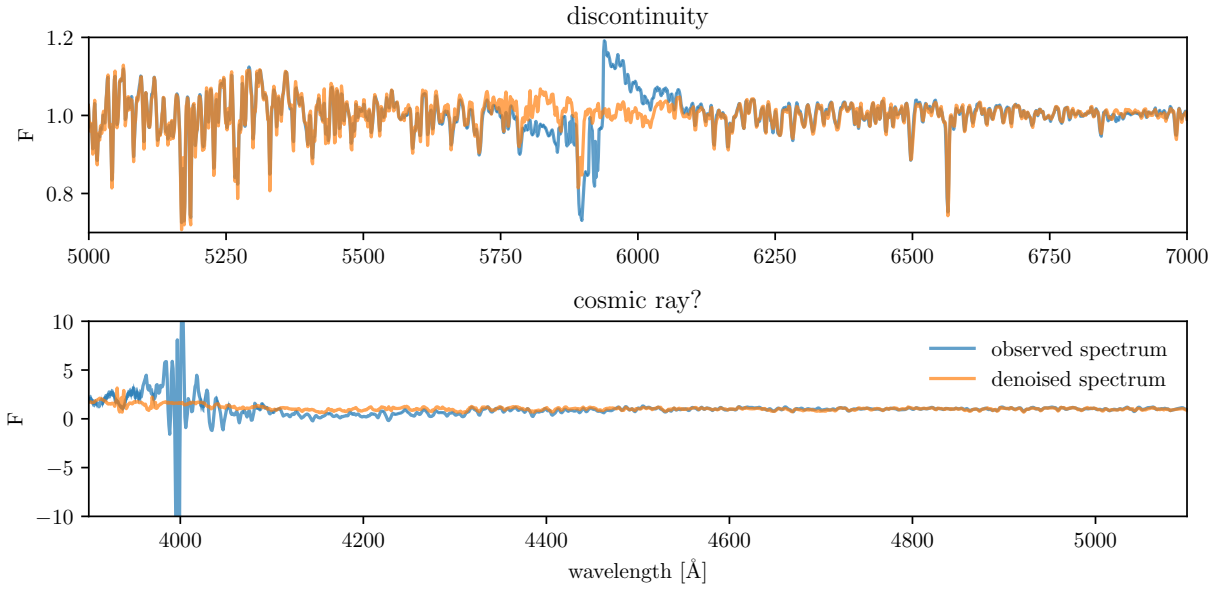

Figure 4: Two spectra with non-stellar features removed by denoising. **top:** a nonphysical discontinuity, perhaps at the chip gap? **bottom:** extremely high- and low- flux spectral pixels, possibly caused by a cosmic ray.

6 Future work

I haven't made as much progress on this project as I'd hoped. Here are some potential future directions.

- More robust inference than MAP estimation. Even if a point estimate of latent parameters is all I want, an approximation of the posterior mean obtained via variational inference may improve generalization to new data.
- While more computationally expensive, modelling the measurement error of F^1 might result in a better inference.
- It would be interesting to see what happens if z is instantiated as a random variable.
- A similar model could be applied to whole star clusters, which have stars of equal-to-within-the-errors abundnaces and age, but which span the gamut of mass and evolutionary stage[4]. A model capturing only the things that are the same for all stars in a cluster might be more effective for identifying the members of disolved clusters than chemical tagging.
- There is other data to which these idea can be applied without modification. The GALAH survey is of particular interest.

The code used for this project is available at `https://github.com/ajwheeler/stereo`.

[4]Star of different masses age at different rates.