Draft version November 13, 2025
Typeset using LaTeX default style in AASTeX631

A robust replacement for principal components analysis, designed to account for missing data and outliers.

Thomas Hilder[1] and David W. Hogg[2]

[1] School of Physics and Astronomy, Monash University VIC 3800, Australia
[2] TODO

1. INTRODUCTION

Matrix factorization methods are a workhorse of astronomy. This is especially true in stellar spectroscopy where it is hard to build detailed theoretical models, but we have a lot of data. Models of this class model a rectangular $N \times M$ data matrix $\mathbf{Y}$ with a matrix $\mathbf{L}$ of lower rank $K \leq \min{(N, M)}$ that minimizes the residual $|\mathbf{Y} - \mathbf{L}|$. The oldest and most widely used of these is principal component analysis,

Prior work:—

- At the beginning of time, there was principal components analysis (PCA) (**?**), the last universal common ancestor (LUCA) of all self-supervised (or unsupervised) machine-learning methods. PCA replaces (or models) the $N \times M$ rectangular data Y with a low-rank matrix L that minimizes the sum of squares of the residual $Y - L$, subject to $\mathrm{rank}(L) = K < \min(N, M)$. Like almost all of its machine-learning-method descendants, PCA requires complete, rectangular data; it treats every data point identically. It is extremely sensitive to outliers; a single bad pixel in one data record can spoil many or all of the delivered eigenvectors. After all, it is a model to minimize unmodeled *variance* (squared error); whereas the empirical variance in a block of data can easily be dominated by one or a few pixels.

- Unrelated to the successes of PCA, humankind evolved, looked at the stars, embraced weighted least squares (chi-squared fitting) (**?**), got upset about the influences of outliers, and immediately started sigma clipping (**?**). This has been the dominant method for making weighted-least-squares fits insensitive to rare outliers or data anomalies; that is, to make them more *robust.* [HOGG: Algorithm??] This method is subject to a kind of "mode collapse" in which large amounts of data get clipped out and the model just doesn't represent those data at all.

- If ancient astronomers had looked at the statistics literature, they would have seen iteratively reweighted least squares (IRLS) (**?**). This method is really a family of methods; but there is a form you can write down [HOGG DO THAT HERE] in which IRLS has all the good properties of weighted least squares with sigma-clipping (and a similar kind of anomaly threshold parameter), but is way less prone to mode collapse. The IRLS method is a workhorse in many domains; it has been used in astronomy now and then (**?**).

- A mathematically rigorous but robust empirical model for (representation for?) data is the robust principal components analysis (Robust-PCA) method of Candés et al (1). This method attempts to describe the rectangular block of data Y as a sum of a low-rank block L and a sparse block S. Interestingly, the method attempts to make this exact, such that $Y = L + S$ exactly. When the Robust-PCA algorithm is iterated for finite time, it comes finitely close [HOGG: CORRECT?]. Fundamentally, the Robust-PCA is an alternation of a singular value decomposition (with a threshold on the singular values) to make L and an outlier identification (with a threshold on the residual) to make S.

- Along a separate thread, Tsalmantza & Hogg introduced heteroskedastic matrix factorization (HMF) (2) as a data-driven dimensionality reduction for astronomical spectra (or other kinds of noisy data). The HMF model of rectangular data Y is the rectangular matrix L of rank $K < \min(N, M)$ that minimizes the chi-squared residual (weighted sum of squares). HMF is a replacement for PCA that does not require complete data (because missing data can be assigned vanishing weights), and it has the satisfying property that every data point is weighted with its associated inverse uncertainty variance, which represents the amount of information it brings. However, HMF is still very sensitive to outliers or anomalies in the data, if they are not weighted appropriately.

The algorithm that follows—Robust-HMF—is very much a mash-up of HMF (2) and Robust-PCA (1). It adds to Robust PCA data weighting and the ability to handle missing and low-information pixels. It sacrifices the nice property of Robust-PCA that it explicitly decomposes the data (or the model for the data) into sparse and low-rank components. It also [HOGG THINKS] can't be expressed precisely as the optimization of a single scalar objective function, which makes it harder to analyze mathematically. [TOM: I think we can express things either as a heavy-tailed MLE, or as a latent-variance hierarchical model with a Gaussian likelihood. See below and Appendix. The latter is undercooked, I only started thinking about it this morning, but I think it's ultimately the nicest. Either way, we have nice statistics.] However, empirically, it works very well in standard astronomical contexts, as we will see.

Other prior art that Tom has found:

1. Wright+2009: Write problem with Langrange multiplier framework swap from L0 to L1 loss on sparse/outlier matrix to get a tractable optimization problem. If they used a different loss I think they would have what we do?
2. Wipf+2009: Actually kind of similar to ours but super expensive and I got a bit lost in the sauce quite fast. Start of section 2 has text about a very similar probabilistic view to ours.
3. Candes+2011: Robust PCA suggestion but not IRLS-based. Called PCP.
4. Polyak+2017: IRLS-based PCA using Huber loss
5. Centofanti+2025: cellPCA. IRLS-based PCA that simultaneously learns per-object and per-"cell" (per y_{ij}) weights by combining two robust losses. The have a nested rescaling when evaluating the loss where each time is hyperbolic tangent as a loss. They handle the missing data case but not heteroskedasticity. I think our loss is much nicer because our latent weights have statistical meaning. Actually they also make a big deal out of their provable convergence. I think we do everything they do but better, apart from potentially their segregation between object and cell weights? We sort of have that though.
6. Rodriguez+2013: fast-PCP. Basically Candes but fast via Lanczos.
7. Cai+2021: LRPCA. Idk this one is confusing but they are doing some gradient-descent-ish thing.
8. Guyon+2012: Candes with an L1 regularisation spatially (they are doing foreground/background identification for moving objects in security cameras and stuff). They do use an IRLS scheme to solve it.

No one else has all of the following together as far as I can tell:

1. Heteroskedastic measurement uncertainties
2. Missing data
3. Student-t/Cauchy loss
4. Closed form ALS/IRLS update rules with guarantees
5. A Bayesian hierarchical interpretation

2. METHOD

2.1. *Model setup and assumptions*

Let $N \in \mathbb{Z}^+$ be the number of spectra, and $M \in \mathbb{Z}^+$ be the number of pixels in each spectrum. y_{ij} is then the value of pixel j for spectrum i. We assume also that the investigator has Gaussian measurement uncertainties σ_{ij} corresponding to each pixel value, an assumption we will return to shortly, and also that these uncertainties have the same units as y_{ij}. In practice we will often refer to "weights" instead, which are simply inverse variances $w_{ij} = \sigma_{ij}^{-2}$.

Conceptually, the forward model for the data is

$$y_{ij} = \sum_{k=1}^{K} a_{ik} g_{kj} + \text{outliers} + \text{noise}, \tag{1}$$

which is a low dimensional linear embedding of rank $K \in \mathbb{Z}^+$. a_{ik} are the entries of an $N \times K$ matrix $\mathbf{A}$ where each of the N rows contains K *coefficients*, and g_{kj} are the entries of a $K \times M$ matrix $\mathbf{G}$ where each of the K rows contain *basis vectors* of length M. The outliers appear explicitly in the above, but ultimately cannot be separated from the noise.

We then make the following assumptions:

1. **Unreliable measurement uncertainties**: The data y_{ij} have known, and approximately Gaussian measurement uncertainties σ_{ij}, although these may not all be *representative* in that some y_{ij} may be outliers or have underestimated σ_{ij}.
2. **Heteroskedasticity**: The measurement uncertainties σ_{ij} may vary for different i, j.
3. **Uniform data grid**: For fixed pixel index j, pixel values across all spectra i correspond to the same wavelength. [TOM: although one could just have shift operators?]
4. **Missing data**: Spectra may be missing values at particular j are handled with vanishing weights by setting the $w_{ij} = 0$ at for each missing y_{ij}, equivalent to infinitely large measurement uncertainties $\sigma_{ij} \to \infty$.
5. **Basis orthogonality**: The inferred basis vectors, and so the rows of G, will be strictly orthogonal and ordered by explained variance, similarly to principal component analysis. The rank of the model is restricted to $1 \leq K \leq N$.

2.2. *Probabilistic view*

Such a model as above, and also for PCA/HMF, are usually inferred by minimizing the χ^2 metric or equivalently a Gaussian likelihood

$$\hat{\mathbf{A}}, \hat{\mathbf{G}} = \underset{\mathbf{A},\mathbf{G}}{\operatorname{argmin}} \left[\sum_{ij} w_{ij} \left(y_{ij} - \sum_{k} a_{ik} g_{kj} \right)^2 \right], \tag{2}$$

$$\text{or equivalently} \quad y_{ij} \sim \text{Normal}\left(\sum_{k} a_{ik} g_{kj}, \sigma_{ij}^2 \right), \tag{3}$$

where the notation on the second line denotes that each y_{ij} is *drawn from* a Normal distribution with mean $a_{ik} g_{kj}$ and variance σ_{ij}^2. This induces a quadratic penalty in the residuals $r_{ij} = y_{ij} - \sum_k a_{ik} g_{kj}$, which causes outliers to have a large influence on the fit.

We instead replace the likelihood with a heavy-tailed distribution, which results in a sub-quadratic penalty in the residuals. Here, we choose Student's t-distribution, but there is a large literature on heavy-tailed distributions for robust inference [TOM cite some of that here], and our method is generalizable to any of those. Thus, we replace the above with[1]

$$\hat{\mathbf{A}}, \hat{\mathbf{G}} = \underset{\mathbf{A},\mathbf{G}}{\operatorname{argmin}} \left[\sum_{ij} \log \left(1 + \frac{w_{ij} r_{ij}^2}{\nu s^2} \right) \right], \tag{4}$$

$$\text{or equivalently} \quad y_{ij} \sim \text{StudentT}_{\nu}\left(\sum_{k} a_{ik} g_{kj}, s^2 \sigma_{ij}^2 \right) \tag{5}$$

where the number of degrees of freedom $\nu \in \mathbb{Z}^+$ and the scale s are hyperparameters. In the limit $\nu \to \infty$ with $s = 1$ this converges to the Normal likelihood we had before. We will discuss how to choose ν and s in a later section [TOM return to this and add a link or fix].

[1] Note that strictly there is a $\nu s^2/2$ prefactor for the full negative log-likelihood. Also for Hogg, your $Q = \nu s^2$.

This setup can be equivalently viewed as a hierarchical model with latent, unknown variances τ_{ij}^2

$$y_{ij} \sim \text{Normal}\left(\sum_k a_{ik} g_{kj}, \tau_{ij}^2\right), \tag{6}$$

$$\tau_{ij} \sim \text{InverseGamma}\left(\frac{\nu}{2}, \frac{\nu}{2} s^2 \sigma_{ij}^2\right) \tag{7}$$

where the Inverse Gamma distribution is a strictly positive distribution, here with *shape* $\frac{\nu}{2}$ and *scale* $\frac{\nu}{2} s^2 \sigma_{ij}^2$. Both the shape and scale control the location of the mode, while the scale mostly controls the support for larger values. s corresponds to a global scale for the confidence in our beliefs about our provided uncertainties, and in practice it will control how pessimistically we treat large residuals during fitting. Very small values $s \ll 0$ correspond to very high confidence in the measurement uncertainties and no outliers, $s = 1$ has mostly correct measurement uncertainties, and larger values $s > 1$ have very low confidence.

The equivalence between the student-t likelihood and the hierarchical view is due to the fact that the Inverse Gamma distribution is the conjugate prior distribution of an unknown Gaussian variance; marginalizing t_{ij} from Eq. (6) with Eq. (7) yields Eq. (5).

For the fitting methods we will present here s and ν are not uniquely identifiable. This is because our estimates are either maximum likelihood in the Student-t view, or type-II maximum likelihood (empirical Bayes) in the hierarchical view. The fit instead depends only on the product $Q = \nu s^2$, which we show in Appendix A.

2.3. *Fitting*

While in principle it's possible to optimize Eq. (4) for all a_{ik} and g_{kj}, we can instead invoke a majorization-maximization scheme that instead implicitly minimizes the objective by iterating between re-weighting and least-squares solves. This type of algorithm is known as iteratively-reweighted least squares (IRLS), and is commonly used in robust regression settings [CITE]. We prove the equivalence between the maximum likelihood and IRLS approaches in Appendix A.

2.3.1. *Initialization*

In either case, we must provide a sensible initialization, since there will be in general very many local optima. The most straightforward approach is the singular-value decomposition

$$[\mathbf{Y}]_{ij} = y_{ij}, \tag{8}$$

$$\mathbf{U}\mathbf{S}\mathbf{V}^\top = \text{svd}\,(\mathbf{Y}), \tag{9}$$

$$a_{ik} \leftarrow [\mathbf{S}]_{kk}^{1/2}\,[\mathbf{U}]_{ik}, \tag{10}$$

$$g_{kj} \leftarrow [\mathbf{S}]_{kk}^{1/2}\,[\mathbf{V}^\top]_{kj}, \tag{11}$$

where $\mathbf{U}$ and $\mathbf{V}^\top$ are unitary matrices, and $\mathbf{S}$ is a matrix with diagonal entries equal to the singular values of $\mathbf{Y}$ in non-decreasing order, and zeros elsewhere.

TODO: note about what to do if $\mathbf{Y}$ is massive?

2.3.2. *Iterative solver*

The algorithm consists of 4 steps iterated in turn until convergence. First, let $Q = \nu s^2$. [TOM: oh damn I just realized the MLE for Student t with ν and s^2 is equivalent to the Cauchy MLE with $s^2 \leftarrow \nu s^2$] These are as follows.

The **w-step** updates the data weights to downweight outliers, given the current best guess of $\mathbf{A}$ and $\mathbf{G}$:

$$w_{ij}^{\text{total}} \leftarrow w_{ij}^{\text{data}} w_{ij}^{\text{robust}}, \tag{12}$$

$$w_{ij}^{\text{robust}} = \frac{Q^2}{Q^2 + w_{ij}^{\text{data}} r_{ij}^2}, \tag{13}$$

$$r_{ij} = y_{ij} - \sum_K a_{ik} g_{kj}, \tag{14}$$

where we note that $w_{ij}^{\text{robust}} \in (0, 1]$ provides a per-data-point measure of outlier-y-ness, and that $w_{ij}^{\text{total}} = \tau_{ij}^{-2}$. This rule also respects data weights of zero, and gives interpretability to Q in that it is a dimensionless soft outlier threshold that sets the degree to which large weighted residuals cause downweighting.

The **a-step** finds the best-fit values for the coefficients a_{ik} given the current estimate of the basis vectors g_{kj}:

$$a_{ik} \leftarrow [\boldsymbol{\alpha}_i]_k \quad \text{for } i \text{ in } 1, ..., N, \tag{15}$$

$$\boldsymbol{\alpha}_i = \text{solve}\,(\mathbf{X}_i, \mathbf{b}_i), \tag{16}$$

$$[\mathbf{X}_i]_{kk'} = \sum_{j=1}^{M} g_{kj} w_{ij} g_{jk'}, \tag{17}$$

$$[\mathbf{b}_i]_k = \sum_{j=1}^{M} g_{kj} w_{ij} y_{ij}, \tag{18}$$

where the operator solve$(\mathbf{X}, \mathbf{b})$ returns $\mathbf{X}^{-1}\,\mathbf{b}$. This is just the weighted least-squares (WLS) solution for the rows of $\mathbf{Y}$ given fixed $\mathbf{G}$.

The **g-step** finds the best-fit basis vectors g_{kj} given the current estimate of the coefficients a_{ik}:

$$g_{kj} \leftarrow [\boldsymbol{\gamma}_j]_k \quad \text{for } j \text{ in } 1, ..., M, \tag{19}$$

$$\boldsymbol{\gamma}_j = \text{solve}\,(\mathbf{X}_j, \mathbf{b}_j), \tag{20}$$

$$[\mathbf{X}_j]_{kk'} = \sum_{i=1}^{N} a_{ik} w_{ij} a_{ik'}, \tag{21}$$

$$[\mathbf{b}_j]_k = \sum_{i=1}^{N} a_{ik} w_{ij} y_{ij}, \tag{22}$$

which is just the WLS solution for the columns of $\mathbf{Y}$ given fixed $\mathbf{A}$.

The **rotation** suppresses the huge set of degeneracies in the model by enforcing a standard orientation in either data or feature space. Here, we will require that the basis vectors be orthonormal:

$$\mathbf{A} \leftarrow \mathbf{A}\mathbf{V}\,\text{diag}\left(\boldsymbol{\lambda}^{-1}\right)\mathbf{V}^{\top}, \tag{23}$$

$$\mathbf{G} \leftarrow \mathbf{V}\,\text{diag}\left(\boldsymbol{\lambda}\right)\mathbf{V}^{\top}\mathbf{G}, \tag{24}$$

$$\boldsymbol{\lambda}, \mathbf{V} = \text{eig}\left(\mathbf{G}\mathbf{G}^{\top}\right), \tag{25}$$

where the operator eig$(\mathbf{G}\mathbf{G}^{\top})$ returns a K-vector and a $K \times K$ matrix, containing the eigenvalues and eigenvectors of $\mathbf{G}\mathbf{G}^{\top}$ respectively. $\boldsymbol{\lambda}^{-1}$ is shorthand for the element-wise reciprocal of the eigenvalues vector $\boldsymbol{\lambda}$. Note that here the eigendecomposition is guaranteed to be performed on a real and symmetric matrix, and so allows for slightly faster numerical routines than the general case.

Convergence is assessed every few cycles by a dimensionless estimate of the size of the g-step adjustment. The output of the procedure is the full matrices $\mathbf{A}$ and $\mathbf{G}$. The robust weights w_{ij}^{robust} for any data point are also calculable by Eq (13).

2.3.3. *Validation and hyperparameter choice*

At test time, a new data object y_* with M pixel values y_{*j} is introduced, with associated weights w_{*j}, including probably some missing data with vanishing weights. The a-step and w-step are iterated on this object to convergence, keeping all the components g_{kj} fixed. Convergence is judged by a dimensionless estimate of the size of the a-step adjustment. The output of test time is K converged coefficients a_{*k}, or equivalently the low-rank representation $\sum_k a_{*k}\, g_{kj}$.

We use this held-out data validation approach to select appropriate robust scale Q and rank K. This involves partitioning the data randomly between a training and a test set. The following metric should then be calculated using the test set data y_{*j}, the fitted coefficients a_{*k}, the basis vectors g_{kj}, and the total weights w_{*j}^{total} calculated with the w-step rule using the test set measurement uncertainties w_{*j}^{data},

$$\text{score}\,(Q,K) = \log\left(1 - \text{std}\left[\sqrt{w_{*j}^{\text{total}}}\left(y_{*j} - \sum_{K} a_{*k} g_{kj}\right)\right]\right). \tag{26}$$

This might seem arcane at first, but it essentially just tests the degree to which after fitting, the data follow the distribution given by the likelihood Eq (6), using the residuals weighted by the inferred weights. That is, we expect the data to be normally distributed with standard deviation equal to the inferred weights $\tau_{*j} = (w_{*j}^{\text{total}})^{-1/2}$, and we test the degree to which the fit results in a distribution with that width.

I think Hogg is also going to advocate for splitting into two groups and testing the consistency between the predictions of actual observables for each model? I don't think you can test consistency for inferred coefficients because it's not a bug that the bases don't have to come out identical even if "correct", right?

2.3.4. *Row- and column-level outlier quantification*

[Tom to rename section something more explanatory]

While the inferred weights on either the training or test sets w_{ij}^{total} give a per-pixel-per-spectrum level view of outlier-y-ness, we can also assemble simple metrics to assess the level to which we should consider individual observations i, or cross-object pixel-grid features j as outliers. We can do this as

$$w_i^{\text{spectrum}} = \frac{1}{M}\sum_{j=1}^{M} w_{ij}, \tag{27}$$

$$w_j^{\text{pixel}} = \frac{1}{N}\sum_{i=1}^{N} w_{ij}, \tag{28}$$

where the former has more obvious utility, and the latter may be used to look for consistently problematic pixels across objects from the instrument or data reduction.

These weights should be calculated for all spectra or all pixels, and the resultant distribution over weights informs the confidence with which one can conclude a particular spectrum is outlying. In Section 4.1 we show that for a carefully chosen K, and data containing true spectra-level outliers, the distribution of weights will consist of two separate modes for each the outliers and the not outliers (? how to say).

3. IMPLEMENTATION

`Robusta-HMF` in `JAX` blah blah. It's open source and easy to use, you should use it.

4. DATA EXPERIMENTS

4.1. *Toy*

We generated 4000 toy spectra with 1200 pixels on the same wavelength grid, from a linear sum of a few polynomials and a low- and fixed-frequency sinusoid to represent continuum, as well as a set of absorption lines present in all the spectra. This means the true underlying model is linear and has a rank of around 5 depending on how orthogonal the components are. We added Gaussian noise with a standard deviation that is proportional to a random factor across spectra, a systematic factor across wavelength, and with a random additive factor per pixel. We assume that we know these noise scales exactly as our measurement uncertainties. We also randomly added bad pixels with null values and zero weights, representing *known* bad pixels [TOM: didn't do this yet].

In addition to this, we injected multiple types of outliers to test the models ability to distinguish each. First, we injected 20 [for now] outlier spectra which consist of only a high frequency sinusoid with a randomly drawn frequency and amplitude in some range. We injected outlier columns, we at a fixed pixel index we injected random large (or very negative) values in a random 30% of all the spectra, intended to represent some systematic reduction, calibration, or instrument issue. We also inject individual bad pixels at completely random locations, consisting of a fixed 0.1% of all pixels in the data. Unlike the bad pixels mentioned in the preceding paragraph, we did *not* set the corresponding measurement weights to zero as these are to represent *unknown* bad pixels or pixel-level outliers. Finally, in 10 [for now] spectra we injected 3 additional absorption lines at random wavelength locations with random amplitudes, to test the model's ability to distinguish outliers that cover a few adjacent pixels in individual spectra.

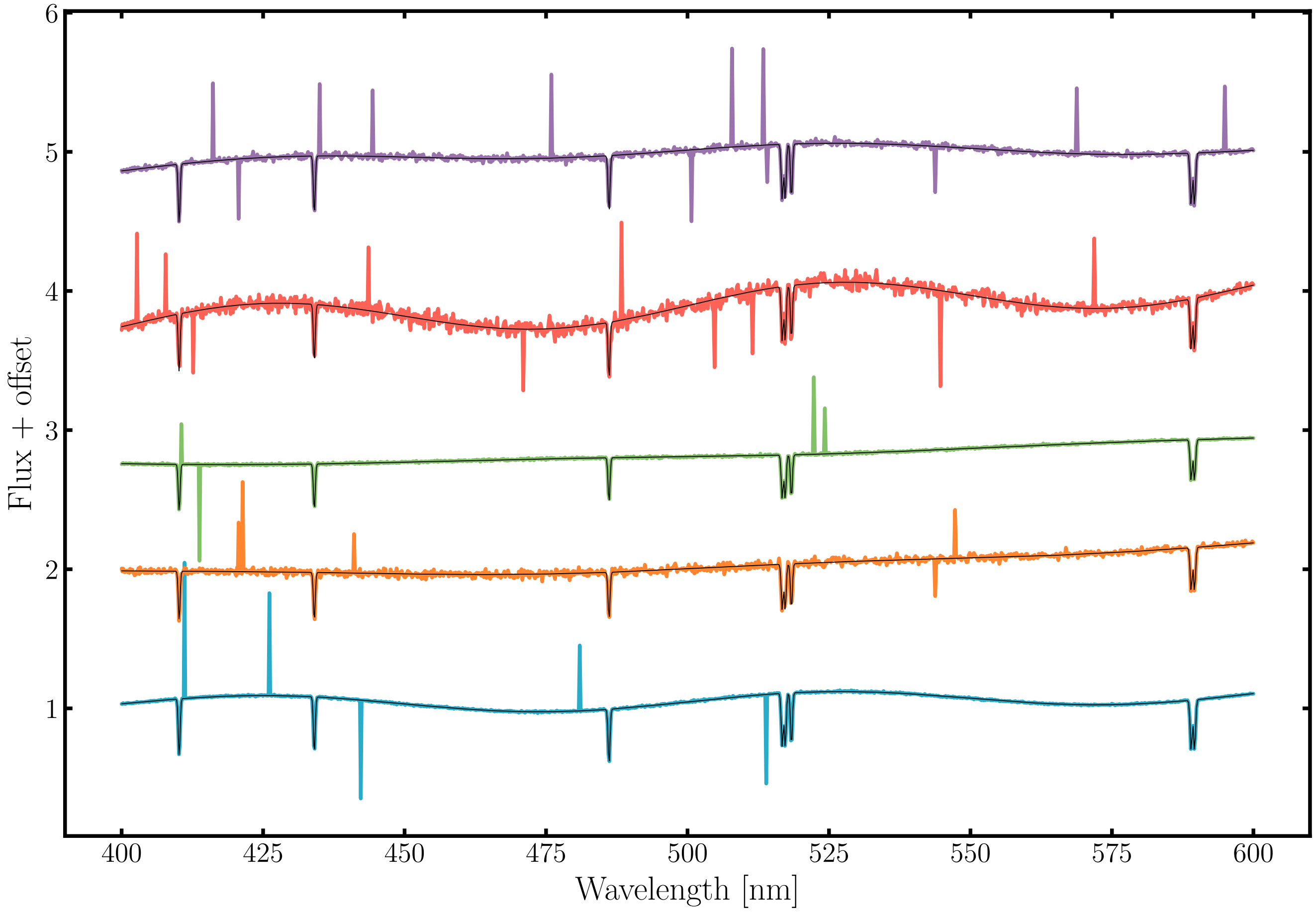

Figure 1. Synthetic noisy and corrupted spectra shown with their best-fits from $K = 5$, $Q = 3$. The fit ignores all of the outlier pixels.

[TOM: I guess I should add all the details of what what distributions the random draws were from in an appendix.]

We fit the model over a grid of 4 potential values of $Q \in \{0.5, 1, 2, 3\}$ and 5 different ranks $K \in \{3, 4, 5, 6, 7\}$ for a total of 20 possibilities. In each, we fit to the same random subset of 3500 spectra and held-out the other 500 for validation purposes.

Figure 1 shows 5 random spectra from the training set, along with the predictions from the best-fit model with $K = 5$ and $Q = 3$. [Tom: Maybe I should show fits to the test set instead?] The model clearly picks up on the structure that is common to all of the spectra like the lines and the continuum structure, but ignores individual bad pixels.

Figure 2 shows 5 random outlier spectra from the training set, along with the model predictions. We see that the predictions do not look like the data here, which is the intent. The model does not spend basis vectors trying to explain these data that do not having any shared low-rank structure amoungest themselves or the rest of the spectra. [Tom: I picked sinusoids of differing periods here because I knew that would prevent a "family" of outliers with shared linear structure that get a basis vector. Is this too optimistic?] The spectra will have very large weighted residuals with respect to the input data weights $w_{ij}^{\rm data}$, but still be approximately normally distributed as according to the *total* weights $w_{ij}^{\rm total}$. [Tom: maybe would be a nice plot if we don't have too many.]

Figure 3 shows the inferred basis functions from the model. The don't really match onto the basis used to make the synthetic data, but they surely span approximately the same subspace. In this case these aren't really interpretable, but extension to include labels, or causal structure, could help that.

Figure 4 demonstrates a mock of a real scientific use-case, which is identifying rare emission/absorption lines. The plot shows the data-weighted residuals, looking mostly flat apart from 5 sharp peaks. Three of these peaks correspond to locations where outlier bad pixels where injected, indicated with the green shading. The other two peaks correspond

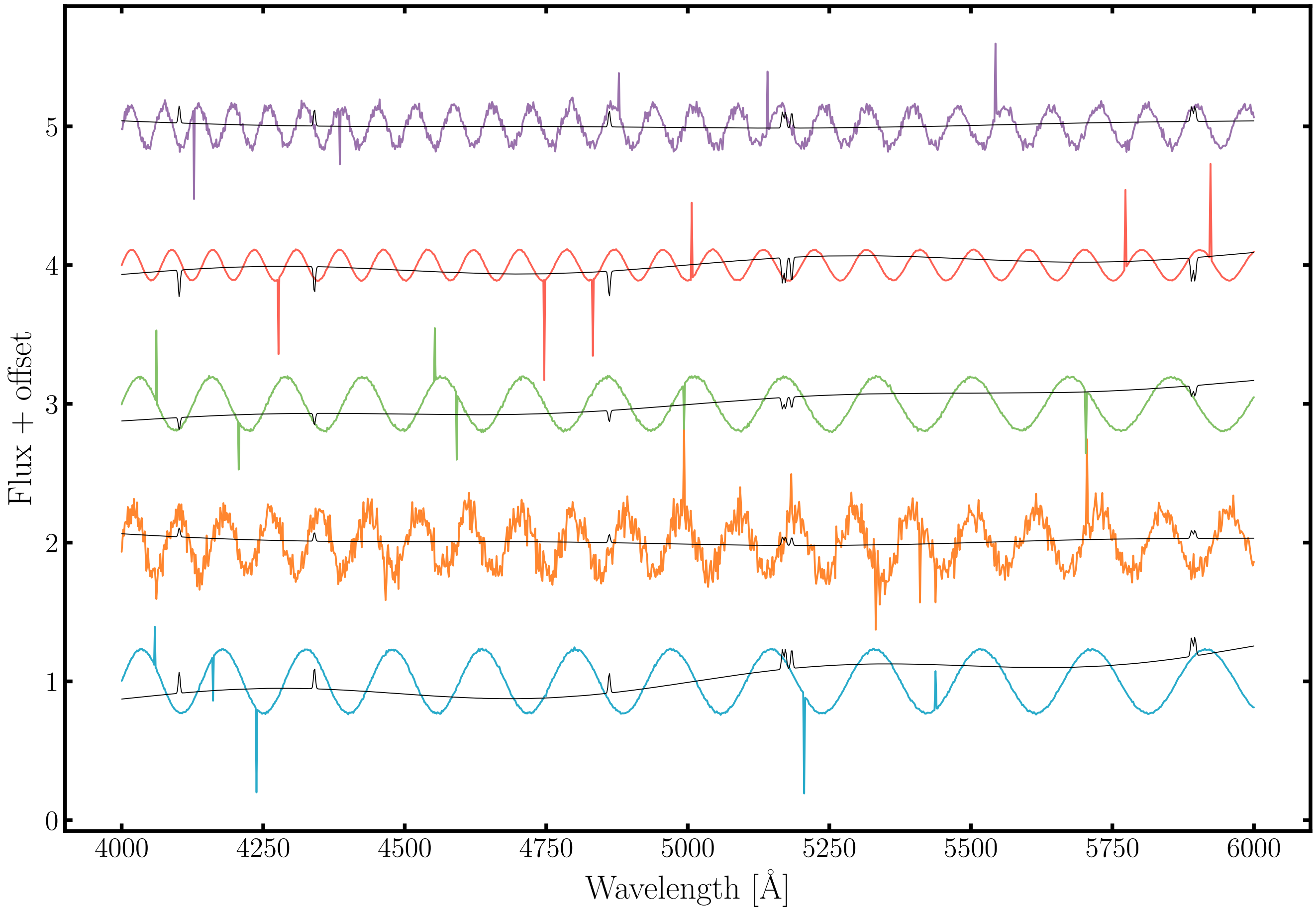

Figure 2. The outlier spectra and their reconstructions from the model. The model clearly does not fit these well at all, since they do not have any shared low-rank structure with either one-another or the other non-outlier spectra.

to multiple pixels where absorption lines where present only in this spectrum, indicated by the orange bands. This makes them very easy to identify by eye after subtracting the model.

Figure 5 shows the distribution of the robust weights w_{ij}^{robust} which give some notion of how much a likely a data point y_{ij} is to be an outlier. The histograms are coloured by their true known origin; blue for untouched pixels that are just the true model plus accurately characterized noise, orange for pixels in the outlier high-frequency-sinusoid spectra, green for unflagged bad pixels, red for bad pixels in the same location across spectra (so a bad "column" in $\mathbf{Y}$), and in purple the outlier absorption lines like those in Figure 4. We see that the model clearly distinguishes these populations, with the clean pixels clustered toward $w_{ij}^{\text{robust}} \approx 1$, and the outliers mostly clustered around $w_{ij}^{\text{robust}} \approx 0$, although there is a reasonable tail of outliers that read toward larger weight values.

Figure 6 shows the validation scores calculated with Eq (26)

4.2. *Hot Stars*

Next, we will use the method to investigate hot stars in the SDSS-V BOSS spectra. We limited our selection to consider only spectra with a signal-to-noise ratio above 29, and I think only spectra from Milky Way Mapper? We selected only spectra that share a common wavelength grid. Additionally, we cut spectrally to include only data between 370 nm and 1030 nm.

We floored the flux measurement uncertainties at 1%, assuming that uncertainties less than this are not unphysical. Hogg then did some other magic floor and ceiling-ing that Tom does not immediately understand.

We then selected [TODO] spectra pseudorandomly [TODO describe] from the much larger set that satisfies our criterion. These spectra were then randomly split into two groups of equal size called A and B.

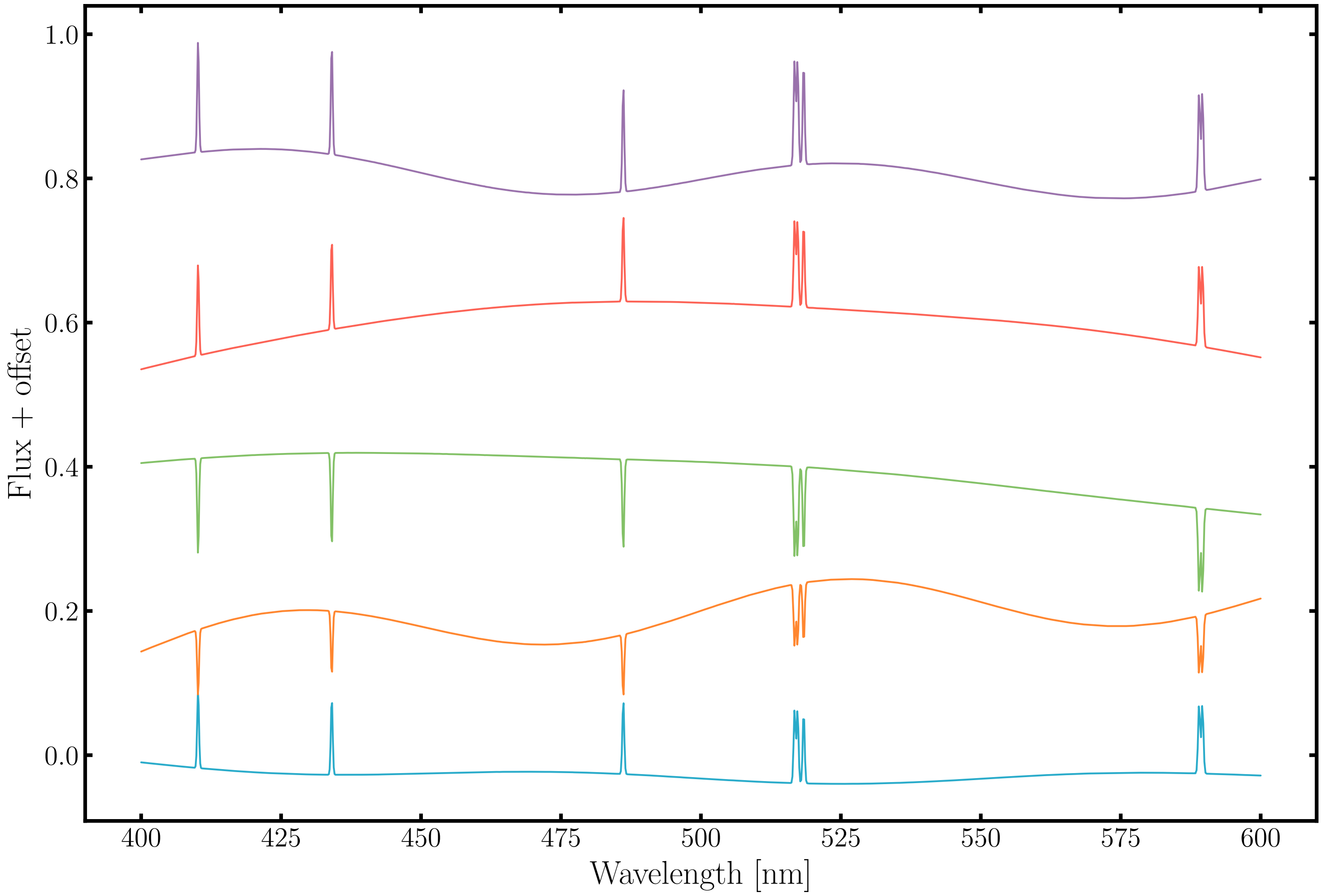

Figure 3. The fitted basis. Clearly did not separate out nicely into a constant + straight line + quadratic + lines + sinusoid, but those things might not be perfectly orthogonal anyway. [TOM: this plot looks better for real data?]

4.3. *Gaia RVS*

5. DISCUSSION

The useful-ness of the robustness of the model is two-fold, one can either learn what to ignore, or what to investigate. We showed in our toy example that this approach does not suffer the issue of outliers spoiling the eigenvectors. We do this without any ad-hoc sigma-clipping, instead taking a purely data-driven approach with a probabilistic interpretation, in that we are marginalizing over our ignorance about the true uncertainties by using the measurement uncertainties as a prior. We also provide relaxed assumptions about the noise distribution, and bad pixels do not ruin rectangularity.

Validation metric cannot be prediction performance on held-out data because there simply is not enough information. We could always score better by increasing K. It's also complicated by the fact that we are deliberately fitting a subset of the data poorly because we believe they are outliers. And worse, we don't know what is truly an outlier nor how to ignore it for the sake of quantifying prediction quality.

We can circumvent this problem by doing something Bayesian-ish, because our optimized total variances τ_{ij}, τ_{*j} have actually performed implicit marginalization. [Tom to think about this more carefully but] this effectively gives us some posterior-ish information in terms of a predictive distribution type thing. The metric we propose in this paper leverages this by simply comparing the distribution of z-scores obtained with our inferred weights from that which the likelihood predicts.

APPENDIX

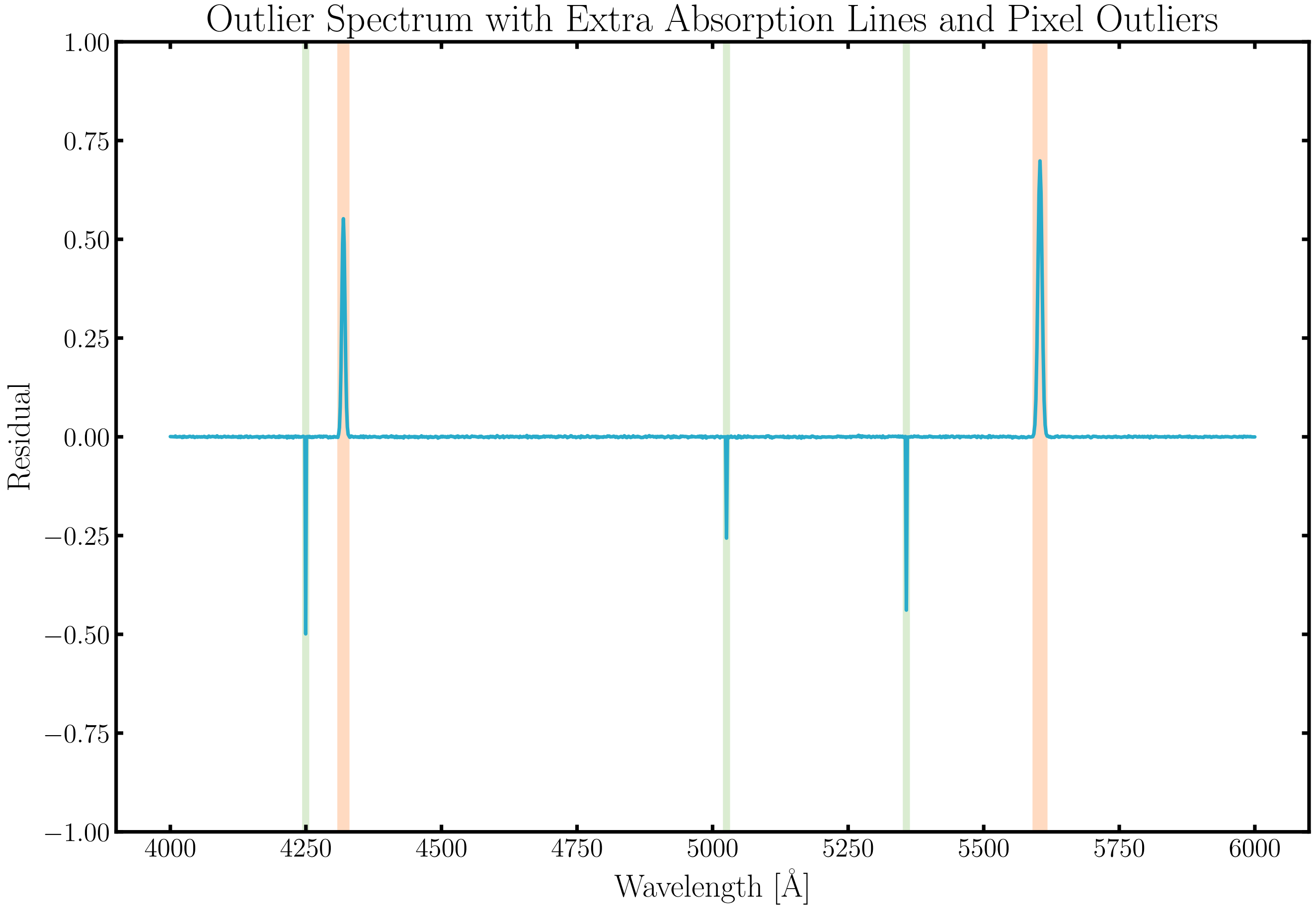

Figure 4. Model-subtracted (actually model - data whoops) spectrum with injected outlier absorption lines indicated by the orange bands. The particular lines are present only in this spectrum and so treated as outliers by the model. This makes them easy to identify in residuals. The green bands indicate where bad pixels were injected, showing that the model also ignores those.

A. PROOF OF OBJECTIVE

[TOM: update now that I am using Student-t not Cauchy.]

A.1. *Auxiliary Form*

Claim: the loss can be expressed for any r in the following way

$$\rho(r) = \min_{0<w\leq 1} \left[\tfrac{1}{2}wr^2 + \phi(w)\right], \tag{A1}$$

where

$$\phi(w) = \frac{Q^2}{2}\,(w - 1 - \log w). \tag{A2}$$

Let's prove that. Define

$$J\,(w; r) = \tfrac{1}{2}wr^2 + \tfrac{Q^2}{2}\,(w - 1 - \log w) \tag{A3}$$

Differentiating with respect to w:

$$\frac{\partial J}{\partial w} = \tfrac{1}{2}r^2 + \frac{Q^2}{2}\left(1 - \frac{1}{w}\right), \tag{A4}$$

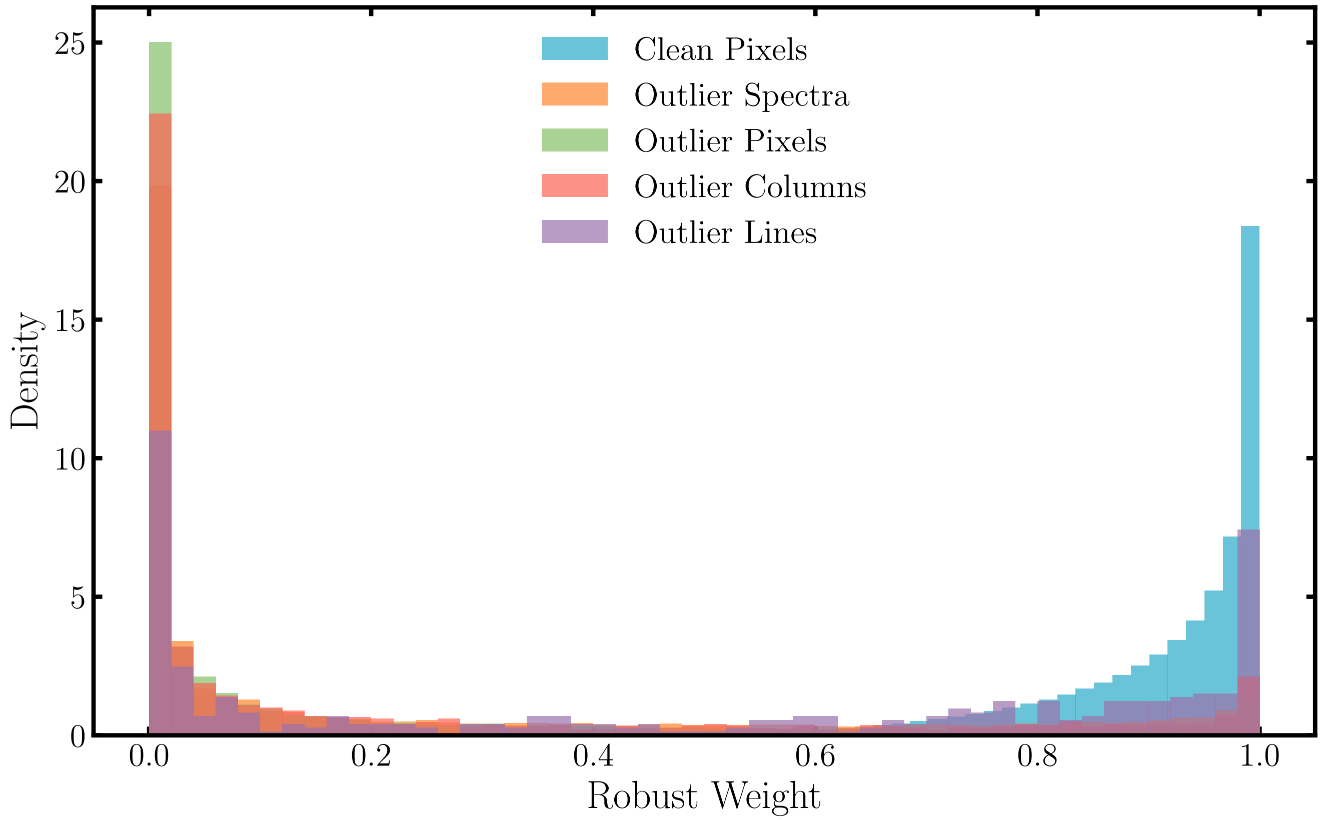

Figure 5. Histogram of all robust weights $w_{\rm ij}$, grouped by whether the associated data point was an injected outlier or "clean", and if an outlier what kind. [TOM: probably should update it such that they are side-by-side rather than overlaid.]

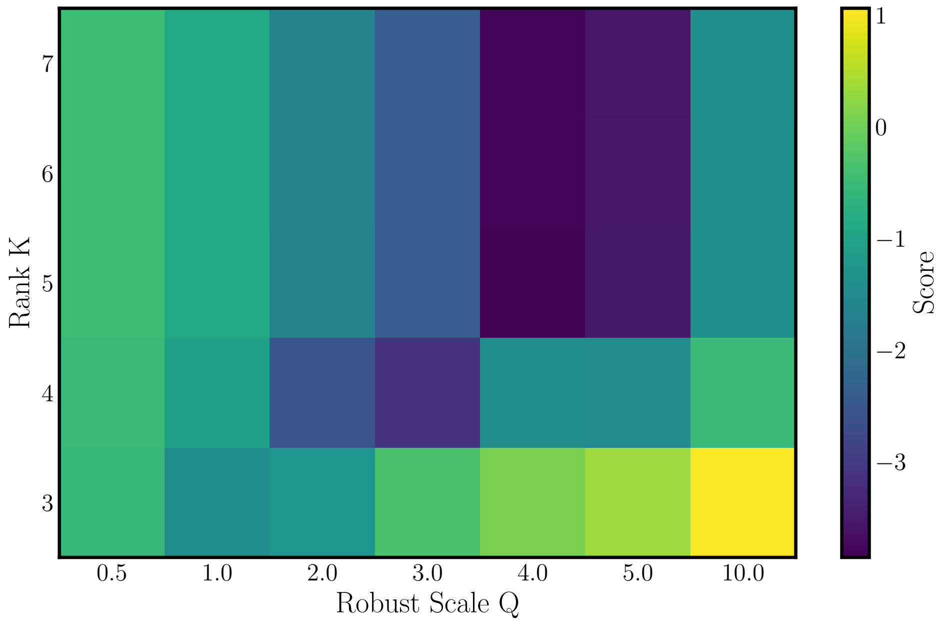

Figure 6. Validation scores based on performance on held-out data. Performance is assessed by metric described in text. Smaller numbers are better, and K should be chosen conservatively in that if the score does not improve much one should not chose a larger K.

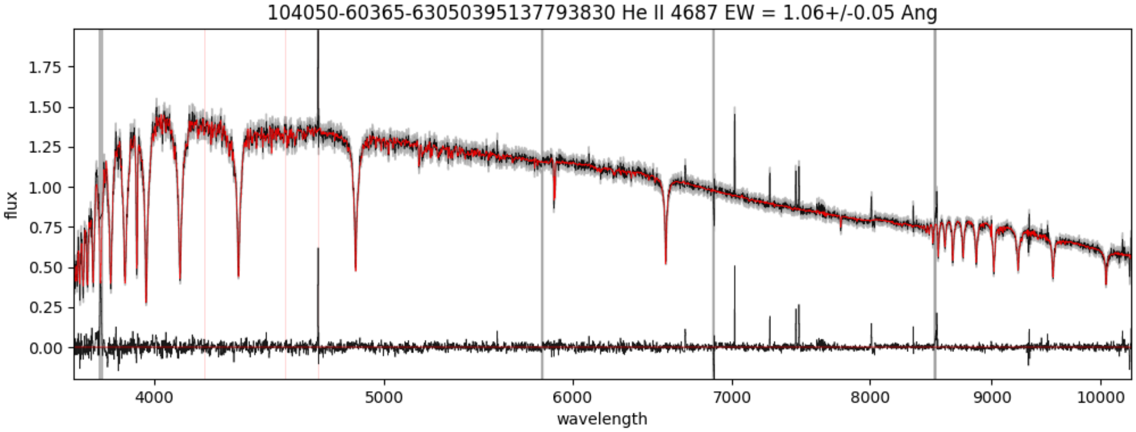

Figure 7. Star with very strong He II in the residual.

and

$$\frac{\partial^2 J}{\partial w^2} = \frac{Q^2}{2}\frac{1}{w^2} > 0, \qquad \forall\, Q, w > 0. \tag{A5}$$

thus we know that the critical point minimises J. Setting $\partial_w J = 0$ yields

$$\hat{w}(r) = \operatorname{argmin}_w J(w; r) \tag{A6}$$

$$= \frac{1}{1 + (r/Q)^2}, \tag{A7}$$

and note that $\hat{w} \in (0, 1]$ because $(r/Q)^2 \geq 0$. We now prove the claim. First set $t = (r/Q)^2 \geq 0$, then

$$\hat{w} = \frac{1}{1+t}, \qquad r^2 = Q^2 t, \tag{A8}$$

such that

$$J(\hat{w}; r) = \tfrac{1}{2}\hat{w}r^2 + \phi(\hat{w}), \tag{A9}$$

substituting and simplifying one piece at a time:

$$\Rightarrow \tfrac{1}{2}\hat{w}r^2 = \frac{Q^2}{2}\frac{t}{1+t} \tag{A10}$$

$$\Rightarrow \phi(\hat{w}) = \frac{Q^2}{2}\left(\hat{w} - 1 - \log \hat{w}\right) \tag{A11}$$

$$= \frac{Q^2}{2}\left(-\frac{t}{1+t} + \log(1+t)\right) \tag{A12}$$

$$\Rightarrow J(\hat{w}; r) = \frac{Q^2}{2}\log\left(1 + \left(\tfrac{r}{Q}\right)^2\right). \tag{A13}$$

Thus

$$\rho(r) = J(\hat{w}; r) \tag{A14}$$

$$= \min_{0<w\leq 1}\left[\tfrac{1}{2}wr^2 + \phi(w)\right], \tag{A15}$$

as claimed.

A.2. *Three-Step Algorithm*

Define new objective:

$$J(A, G, W) = \frac{1}{2} \sum_{ij} \left[w_{ij} r_{ij}^2 + \phi(w_{ij}) \right], \tag{A16}$$

with $r_{ij} = (Y_{ij} - a_i^\top g_j)/\sigma_{ij}$. By construction,

$$L(A, G) = \min_W J(A, G, W), \tag{A17}$$

and if

$$\hat{W} = \mathrm{argmin}_w \, J(A, G, W), \tag{A18}$$

then

$$\left[\hat{W}\right]_{ij} = \hat{w}(r_{ij}) \tag{A19}$$

$$= \frac{1}{1 + (r_{ij}/Q)^2}. \tag{A20}$$

This immediately yield's Hogg's procedure

$$\begin{aligned} \text{w-step:} \quad & w_{ij} \leftarrow \hat{w}(r_{ij}), \\ \text{a-step:} \quad & \text{solve WLS for } A \text{ with new weights}, \\ \text{g-step:} \quad & \text{solve WLS for } G \text{ with new weights}. \end{aligned}$$

where the a-step optimises the quadratic

$$Q(A \,|\, G, W) = \frac{1}{2} \sum_{ij} w_{ij} r_{ij}^2, \tag{A21}$$

and the g-step optimises $Q(G \,|\, A, W)$. It should be pretty apparent now that the procedure gives the MLE with a Cauchy likelihood.

A.3. *Extra convincing (showing that the procedure optimises L)*

Consider one outer cycle starting at $(A^{(t)}, G^{(t)})$. Choose $W^{(t)} = \hat{w}(r(A^{(t)}, G^{(t)}))$. Then

$$L(A^{(t)}, G^{(t)}) = J(A^{(t)}, G^{(t)}, W^{(t)}). \tag{A22}$$

With frozen $W^{(t)}$, the a- and g-steps minimize $Q(\cdot \,|\, W^{(t)})$. Since our total objective is $J = Q + \sum \phi(W^{(t)})$, this implies

$$J(A^{(t+1)}, G^{(t+1)}, W^{(t)}) \leq J(A^{(t)}, G^{(t)}, W^{(t)}). \tag{A23}$$

We're guaranteed to be helped by the w-step again now, so setting

$$W^{(t+1)} = \hat{w}\left(r(A^{(t+1)}, G^{(t+1)})\right), \tag{A24}$$

and using our result from the previous section gives

$$J(A^{(t+1)}, G^{(t+1)}, W^{(t+1)}) \leq J(A^{(t+1)}, G^{(t+1)}, W^{(t)}). \tag{A25}$$

Thus chaining the inequalities and $L(A, G) = \min_W J(A, G, W)$ gives

$$L(A^{(t+1)}, G^{(t+1)}) \leq L(A^{(t)}, G^{(t)}). \tag{A26}$$

This is enough to guarantee that robust HMF with Hogg's w-step converges to the Cauchy MLE.

REFERENCES

[1]Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[2]P Tsalmantza and David W Hogg. A data-driven model for spectra: Finding double redshifts in the Sloan Digital Sky Survey. *The Astrophysical Journal*, 753(2):122, 2012.