Multiscale Feature Attribution for Outliers

Jeff Shen ⓘ
Department of Astrophysical Sciences
Princeton University
shenjeff@princeton.edu

Peter Melchior ⓘ
Department of Astrophysical Sciences
Center for Statistics & Machine Learning
Princeton University
peter.melchior@princeton.edu

Abstract

Machine learning techniques can automatically identify outliers in massive datasets, much faster and more reproducible than human inspection ever could. But finding such outliers immediately leads to the question: which features render this input anomalous? We propose a new feature attribution method, Inverse Multiscale Occlusion, that is specifically designed for outliers, for which we have little knowledge of the type of features we want to identify and expect that the model performance is questionable because anomalous test data likely exceed the limits of the training data. We demonstrate our method on outliers detected in galaxy spectra from the Dark Energy Survey Instrument and find its results to be much more interpretable than alternative attribution approaches.

1 Introduction

Machine learning (ML) is becoming increasingly necessary in astronomy to deal with the massive datasets that are already available or will become available in the near future. But the adoption of novel tools does change the scientific motivation, namely to *understand* the phenomena in the data. It is therefore important to investigate not just if a ML method works, or how well it works, but also why it produces the reported results. For complex methods, such as deep neural networks, explainability is an open challenge, which suggest a rephrasing of the question: what aspects of the data led to the reported results? Several feature attribution approaches have been developed to answer this question and yield insights not just into the data, but also into the inner workings of the ML method.

We investigate feature attribution methods for the task of outlier detection. ML methods are very capable of finding anomalous data in large data sets, but with every detection comes the question: what made that input an outlier? We present a new feature attribution method and compare it to existing alternatives on outliers identified by Liang et al. (2023a) in spectra of the Early Data Release (EDR; DESI Collaboration et al., 2023) of the Dark Energy Survey Instrument (DESI; DESI Collaboration et al., 2016) Bright Galaxy Survey (BGS; Hahn et al., 2023).

The outlier detection method is based on the autoencoder model by Melchior et al. (2022). It encodes spectra into a low-dimensional latent space and then uses a normalizing flow in the latent space to determine objects that have low probability relative to the rest of the data (see Liang et al., 2023b, for details). We now want to identify—without human intervention—the features in a spectrum that most contribute to the ML model (here the combination of autoencoder and normalizing flow) identifying that spectrum as an outlier.

2 Feature Attribution Methods

Numerous methods have been proposed for feature attribution in machine learning. We discuss two main classes, beginning with gradient-based attribution methods. The most direct one, **Instantaneous**
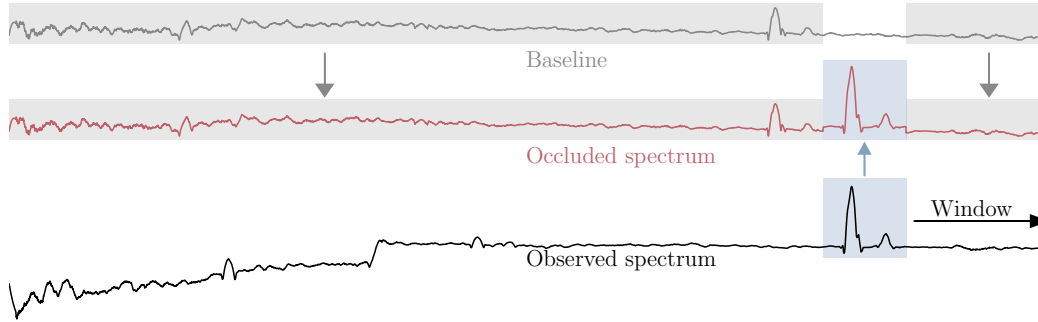
Figure 1: In inverse occlusion, a segment of the baseline is replaced with a segment of the test spectrum to create an occluded spectrum, which is then used to calculate the network output. This approach is more robust for outliers because the bulk of the occluded spectrum is not anomalous.

Gradients, considers only the gradient of the output $f(\mathbf{x}) \in \mathbb{R}^L$, in our case the likelihood $\log p(\mathbf{x})$ from a normalizing flow. An early example of this approach is the saliency maps of Simonyan et al. (2014). While extremely fast and simple to implement, instantaneous gradients may saturate when a change in input does not lead to a change in output (Shrikumar et al., 2019). Sundararajan et al. (2017) proposed **Integrated Gradients (IG)** to solve this saturation issue. The idea is to integrate the gradients of the output with respect to the input along a straight line (i.e., interpolations) from a baseline input to the observed input. Doing so requires numerical integration over the interpolations between the baseline and the observed input, which make this is a slower method. The choice of the baseline is not trivial (see Sturmfels et al., 2020, for an extensive discussion). For our particular application, the common baseline of a zero vector will turn out to be inappropriate (see Section 3). **Expected gradients (EG)** mitigate the choice of baseline by averaging over many different baselines drawn randomly from the underlying dataset (Erion et al., 2020).

The second class of methods perturb the inputs. **Feature ablation (FA)** replaces a single input feature with its corresponding value from some baseline, forming a modified spectrum $\hat{\mathbf{x}}$, and calculates the difference between the two versions, $\text{FA} = f(\mathbf{x}) - f(\hat{\mathbf{x}})$, repeating the same procedure for all features. If changing the feature value has a large impact on the output, then that feature is important. However, input features are often correlated. In our case, we know that the majority of variation in spectra can be modelled by very few variables, a property that is exploited by the autoencoder compression that the outlier detection model of Liang et al. (2023a) is based on. Furthermore, many interesting features are wider than a single pixel (e.g., broad lines from quasars). The ablation of a single feature may therefore miss many important features. **Occlusion** is an attribution method similar to ablation; it has been used in computer vision tasks (e.g., Zeiler & Fergus, 2013) and involves ablating a contiguous block of size W with a baseline rather than a single feature.

3 Feature Attribution for Outliers

We propose **Inverse Multiscale Occlusion (IMO)** for feature attribution in detected outliers, a modification of standard occlusion with several desirable features. Here we describe the motivations for the development of this algorithm and provide the pseudocode in Algorithm 1. Code to reproduce the results in this paper is available at `https://github.com/al-jshen/imo`.

Outliers have by construction low probability. Modifying an already anomalous input further by occlusion with a baseline spectrum will likely lead to erratic behavior because the modified feature vector $\hat{\mathbf{x}}$ is almost guaranteed to be outside of the training data. We therefore perform "inverse occlusion", which replaces a segment of length W, centered at position i, of the baseline spectrum $\mathbf{x}_b$ with the corresponding segment of the input spectrum to form the modified spectrum $\hat{\mathbf{x}}_{\overline{W_i}}$ (see Figure 1), and use the probability of the baseline, rather than the input spectrum, as the reference. For a given position, window size, and baseline b, the attribution metric is thus $\text{IMO}_{W_i,b} = f(\mathbf{x}_b) - f(\hat{\mathbf{x}}_{\overline{W_i}})$. By sliding this window across the entire input, we get a vector $\text{IMO}_{W,b} \in \mathbb{R}^L$. Larger values of IMO_W

indicate a higher probability for the original baseline not containing a particular segment from the input spectrum, i.e., this segment is unusual.

Anomalous features may be broad (spread over multiple pixels) or narrow. In the case of galaxy spectra, contamination by another source is a broad feature because the entire spectrum will look different, but outflows causing asymmetric emission lines is a much more localized feature. We want our method to identify the relevant features in both cases without having to specify a priori what we are looking for. Taking inspiration from wavelet filtering, we perform occlusion with variable window sizes, i.e., we get IMO_W for each choice of window size W. Smaller window sizes respond to local features, while larger windows sizes are sensitive to broad or even global features. We compute attributions for all window sizes, so that users can decide what features are relevant in each case.

This leaves the question as to how to choose the baseline, which serves as a counterfactual to determine how the model would have responded to an input if it did not have that feature. We adopt the use of multiple baselines, randomly drawn from the underlying dataset, from the EG method, but we make a domain-specific adjustment for galaxy spectra. We need to ensure that the baselines are consistent with the input spectrum even if they come from galaxies at different redshifts, i.e., where spectral features are stretched by a different amount from the input spectrum. To remove this source of discrepancy, we redshift the baseline spectra to the same redshift as the input spectrum. In other words, we are finding anomalous features for objects at a particular redshift because we are comparing to objects at the same redshift.

The use of an ensemble of baselines $\mathcal{B}$ also enables us to form a single feature attribution result as the minimum-variance combination across all window sizes:

$$\text{IMO}_W = \mathbb{E}_{b \sim \mathcal{B}}\left[\text{IMO}_{W,b}\right], \ \ \sigma_W^2 = \text{var}_{b \sim \mathcal{B}}\left[\text{IMO}_{W,b}\right]$$
$$\text{IMO} = \sum_W \text{IMO}_W \odot \sigma_W^{-2} \Big/ \sum_W \sigma_W^{-2},$$

with all operations computed element-wise in $\mathbb{R}^L$. This attribution method is emphasizes, at every location i, the window size for which the most significant discrepancies are found in the ensemble.

4 Application to DESI Outliers

We apply the feature attribution methods we discussed above to the DESI spectrum of outlier 4, as listed by Liang et al. (2023a). The code, data, and models to create similar figures for all top 200 outliers listed in Liang et al. (2023a) are available at `https://github.com/al-jshen/imo`. As shown in the top panel of Figure 2, the spectrum shows a calibration problem in the B arm of the spectrograph—which is roughly responsible for the left third of the spectrum—causing a jump and negative flux values on the left side of the spectrum. The same spectrum also shows double-peaked emission lines, which is interesting as it indicates that there may be matter moving at different relative velocities, leading to both a leftward and rightward shift in the observed wavelength of an emission line[1]. This is a localized anomalous feature, particularly visible in the emission of Hα around $6563\,\text{Å}$ (insets in Figure 2). With large- and small-scale anomalous features, outlier 4 is a good test case for attribution methods.

Gradient-based methods show no significant attributions in the left part of the spectrum, largely missing the obvious calibration problem. FA shows a significant bump at the transition between the anomalous and the correct parts of the spectrograph (the transition from the left third to the middle third of the spectrum), but becomes flat at wavelengths less than $4500\,\text{Å}$, missing the progressive deviation towards the left side. IMO detects this large-scale anomaly. Looking at right panel of Figure 2, we can see that the sharp jump is picked up by moderate window sizes of $W = 16$ and $W = 64$, while the progressive drop towards the left side shows up when using larger window sizes. With a single window size, e.g., for standard occlusion with $W = 64$, the calibration mismatch does not show up as a prominent feature.

The emission line shapes, especially for Hα and N$\textsc{ii}$, are detected as anomalous in almost all methods. The gradient-based methods show negative attributions on either side of the Hα line, indicating that suppressing the data there would make the spectrum appear less anomalous. By averaging over

---

[1]Note that this is distinct from having two emission lines whose wavelengths are intrinsically close together.
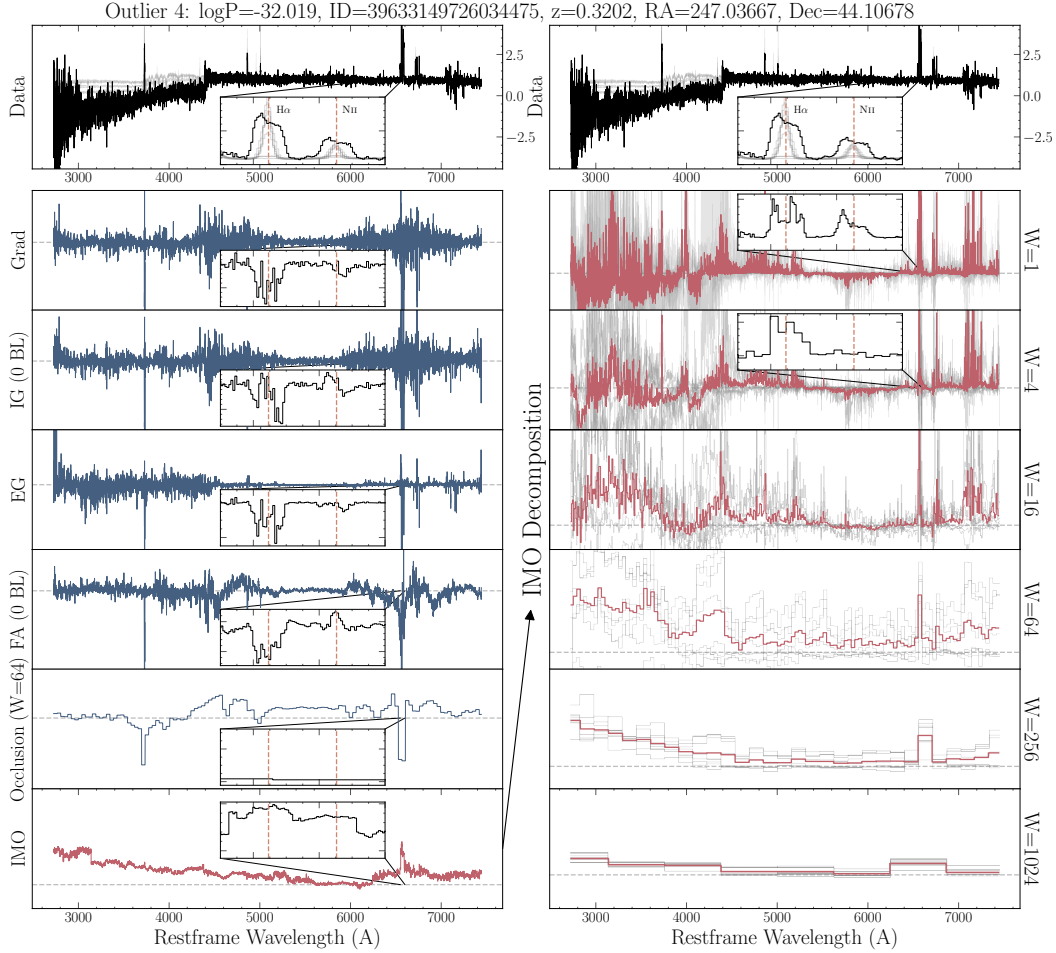
Figure 2: Comparison of attribution methods for spectroscopic outlier 4 from (Liang et al., 2023a), which exhibits a large-scale calibration problem and double-peaked emission lines. The observed spectrum is shown in the top panel, with several baselines in light gray. Existing feature attributions are shown in blue in the left column; the last panel shows the our combined Inverse Multiscale Occlusion (IMO) in red. The right column shows the IMO attributions separately for each window size, with individual baseline results in gray. The insets zoom in on a region around the Hα line.

multiple baselines, EG produces the least noisy attributions. FA also detects the double-peaked line shapes, with negative values indicating a region where the test spectrum has lower probability than the baseline. On the other hand, standard occlusion with $W = 64$ recognizes the line region itself as somewhat anomalous, but fails to resolve any details because the window size is too large compared to the size of the features. Finally, for IMO the Hα line region produces the highest outlier attribution. In the combined statistic, in the bottom left panel of Figure 2, the line appears somewhat broadened, a consequence of having significant attributions over a range of window sizes. Decomposing the outlier attribution into different window sizes reveals a very clear double-peak anomaly for $W = 1$ and $W = 4$, with peaks on either side of the nominal line center, demonstrating the utility of reporting all IMO$_W$ separately. Finally, IMO with $W = 1$ localizes the major anomalies as distinct peaks much more evidently as FA, despite both being based on single-pixel occlusion. This result is a clear demonstration of the benefit of the inverted occlusion method. In IMO, the baseline is (typically) not anomalous, and the occluded vector is thus likely in the region of good model performance, so that anomalous features produce more meaningful attributions.

4

5   Conclusion

Our new attribution method, Inverse Multiscale Occlusion, is able to identify anomalous features—both large-scale and small-scale—for known outliers. By applying IMO to a spectroscopic outlier in DESI BGS, we demonstrate that it peaks at a large calibration offset and at double-peaked emission lines, reproducing the findings from the visual inspection performed by Liang et al. (2023a). We show that IMO outperforms gradient-based methods, which often capture only highly localized anomalies and are harder to interpret, and is also more flexible and better suited to outliers than other perturbative methods. IMO produces a meaningful outlier attribution summary as well as scale-dependent attributions, which enables the interpretation of anomalies with little human supervision.

6   Supplementary Material

We list the pseudo-code of our feature attribution method IMO below:

Algorithm 1 Inverse multiscale occlusion (IMO)

1: **Inputs:**
    Input spectrum $\mathbf{x} \in \mathbb{R}^N$
    Input redshift $z \in \mathbb{R}$
    Encoder model $f : \mathbb{R}^N \to \mathbb{R}^S$
    Decoder model $g : (\mathbb{R}^S, \mathbb{R}) \to \mathbb{R}^N$
    Normalizing flow log-probability function $h : \mathbb{R}^S \to \mathbb{R}$
    Spectrum dataset $\mathcal{D}$
    Number of baseline samples M
    Window sizes $W = W_1, \ldots, W_K$
2: **function** INVERSEMULTISCALEOCCLUSION($\mathbf{x}, z, \mathcal{D}, M, W$)
3:    **Initialize:**
         $\mathbf{A} \leftarrow \texttt{zeros}(K, N), \mathbf{p} \leftarrow \texttt{zeros}(N), \mathbf{R} \leftarrow \texttt{zeros}(M, N)$
4:    **for** $j = 1, \ldots, M$ **do**   ▷ *Loop over all baselines*
5:       $\mathbf{B} \sim \mathcal{D}$ ▷ *Sample a baseline spectrum from the dataset*
6:       $\mathbf{R_j} \leftarrow g(f(\mathbf{B}), z)$ ▷ *Reconstruct the baseline spectrum at the input redshift*
7:       $p_j \leftarrow h(f(\mathbf{R_j}))$ ▷ *Encode and compute the log-probability of*
                               the redshifted baseline spectrum
8:    **end for**
9:    **for** $k = 1, \ldots, K$ **do**   ▷ *Loop over all window sizes*
10:      $\mathbf{T} \leftarrow \texttt{zeros}(M, N)$ ▷ *Initialize temporary attribution array*
11:      **for** $j = 1, \ldots, M$ **do**   ▷ *Loop over all the baseline samples*
12:         $i \leftarrow 0$ ▷ *Initialize the counter index*
13:         **for** $c = 1, \ldots, \texttt{ceil}(N/W_k)$ **do**   ▷ *Loop over all chunks*
14:           $u \leftarrow \min(i + W_k, N)$ ▷ *Calculate the upper bound of the current chunk*
15:           $\mathbf{y} \leftarrow \mathbf{R_j}$ ▷ *Copy the baseline spectrum*
16:           $\mathbf{y}_{i:u} \leftarrow \mathbf{x}_{i:u}$ ▷ *Replace the current chunk with the corre-*
                               sponding chunk from the input spectrum
17:           $l \leftarrow h(f(\mathbf{y}))$ ▷ *Compute the log-probability of the occluded spectrum*
18:           $T_{j,i:u} \leftarrow (p_j - l)/(u - i)$ ▷ *Calculate the attribution*
                               for the current chunk
19:           $i \leftarrow i + W_k$ ▷ *Increment the counter index*
20:         **end for**
21:      **end for**
22:      $\mathbf{A_k} \leftarrow \frac{1}{M} \sum_{j=1}^{M} \mathbf{T_j}$ ▷ *Compute average attribution for each window size*
23:    **end for**
24:    **return A**

References

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016. `http://arxiv.org/abs/1611.00036`

DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2023, doi: 10.5281/zenodo.7964161

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S., & Lee, S.-I. 2020, Improving Performance of Deep Learning Models with Axiomatic Attribution Priors and Expected Gradients, arXiv. http://ascl.net/1906.10670

Hahn, C., Wilson, M. J., Ruiz-Macias, O., et al. 2023, The Astronomical Journal, 165, 253, doi: 10.3847/1538-3881/accff8

Liang, Y., Melchior, P., Hahn, C., et al. 2023a, Outlier Detection in the DESI Bright Galaxy Survey, arXiv. http://ascl.net/2307.07664

Liang, Y., Melchior, P., Lu, S., Goulding, A., & Ward, C. 2023b, Autoencoding Galaxy Spectra II: Redshift Invariance and Outlier Detection, arXiv. http://ascl.net/2302.02496

Melchior, P., Liang, Y., Hahn, C., & Goulding, A. 2022, Autoencoding Galaxy Spectra I: Architecture, arXiv. http://ascl.net/2211.07890

Shrikumar, A., Greenside, P., & Kundaje, A. 2019, Learning Important Features Through Propagating Activation Differences, arXiv. http://ascl.net/1704.02685

Simonyan, K., Vedaldi, A., & Zisserman, A. 2014, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, arXiv. http://ascl.net/1312.6034

Sturmfels, P., Lundberg, S., & Lee, S.-I. 2020, Distill, 5, doi: 10.23915/distill.00022

Sundararajan, M., Taly, A., & Yan, Q. 2017, Axiomatic Attribution for Deep Networks, arXiv. http://ascl.net/1703.01365

Zeiler, M. D., & Fergus, R. 2013, Visualizing and Understanding Convolutional Networks, arXiv. http://ascl.net/1311.2901