The Astrophysical Journal, 833:26 (14pp), 2016 December 10 doi:10.3847/0004-637X/833/1/26

MAPPING THE SIMILARITIES OF SPECTRA: GLOBAL AND LOCALLY-BIASED APPROACHES TO SDSS GALAXIES

David Lawlor[1,2], Tamás Budavári[3,4], and Michael W. Mahoney[5,6]
[1] Statistical and Applied Mathematical Sciences Institute, USA
[2] Dept. of Mathematics, Duke University, USA
[3] Dept. of Applied Mathematics and Statistics, The Johns Hopkins University, USA
[4] Dept. of Computer Science, The Johns Hopkins University, USA
[5] International Computer Science Institute, USA
[6] Dept. of Statistics, University of California, Berkeley, USA
Received 2016 May 20; revised 2016 August 19; accepted 2016 September 8; published 2016 December 5

ABSTRACT

We present a novel approach to studying the diversity of galaxies. It is based on a novel spectral graph technique, that of *locally-biased semi-supervised eigenvectors*. Our method introduces new coordinates that summarize an entire spectrum, similar to but going well beyond the widely used Principal Component Analysis (PCA). Unlike PCA, however, this technique does not assume that the Euclidean distance between galaxy spectra is a good global measure of similarity. Instead, we relax that condition to only the most similar spectra, and we show that doing so yields more reliable results for many astronomical questions of interest. The global variant of our approach can identify very finely numerous astronomical phenomena of interest. The locally-biased variants of our basic approach enable us to explore subtle trends around a set of chosen objects. The power of the method is demonstrated in the Sloan Digital Sky Survey Main Galaxy Sample, by illustrating that the derived spectral coordinates carry an unprecedented amount of information.

Key words: catalogs – galaxies: general – methods: data analysis – methods: statistical – surveys

1. INTRODUCTION

The physical properties of the universe and the internal mechanisms of galaxies are ultimately intertwined in astronomical observations. Characterizing the diversity of galaxies is vital, not only for understanding their evolution, but for unraveling the nature of dark energy in the context of our cosmological models. While today's large-scale spectroscopic surveys provide a plethora of data, the usual tools designed to capture specific aspects of the spectra might not take full advantage of our experiments. The large homogeneous collections of spectra currently available offer new opportunities for statistical studies, and our goal here is to develop novel approaches, which can empirically find trends in the data that later can be understood in the context of galaxy evolution.

Current analysis approaches generally fall into one of two broad categories. In the first category, the observed spectra are fitted by theoretical or semi-analytic models, e.g., Bruzual & Charlot (2003), to infer their parameters. These estimates in turn provide a model-dependent coordinate system with absolute scales such as age and metallicity. These physical measurements are then used in subsequent population studies, etc. Challenges for these methods typically include systematic biases due to imperfect models as well as correlated parameters. In the second category, one adopts a more empirical approach, where galaxies are analyzed in relation to other galaxies based on the original measurements, i.e., the observed spectra. A major challenge for these more empirical methods is the conceptual problem of how best to compare empirical spectra, e.g., which features of a spectrum are most important for identifying similarities between two objects. The approach we describe, in this paper, falls into the second category, and it aims to address the fundamental issue of measuring similarity between galaxy spectra as well as how to use this information in conjunction with principled machine learning algorithms to obtain astronomical insight.

1.1. Embedding Spectra

A canonical example of the empirical approach is Principal Component Analysis (PCA), which is widely used to find the globally dominant linear trends in the data. PCA was first applied to galaxies by Connolly et al. (1995b), who found that a significant fraction of the variance in the spectra can be captured by only three components. In other words, the analyzed spectra could be well approximated by a linear combination of three *eigenspectra*. The coefficients serve as summaries of the high-dimensional spectra, and in this coordinate system galaxies could be meaningfully compared to one another. This is called a *low-dimensional embedding,* because every high-dimensional spectrum is mapped to just a few coefficients, i.e., to a low-dimensional vector. PCA has been used in many areas of astronomy, including photometric redshift estimation (Connolly et al. 1999; Budavári et al. 2000), sky subtraction (Wild & Hewett 2005), and the classification of galaxies and quasars (Francis et al. 1992; Connolly & Szalay 1999; Yip et al. 2004a, 2004b). The Sloan Digital Sky Survey (SDSS) (York et al. 2000) has adopted this method in its data reduction pipeline, and it automatically derives the first five eigencoefficients (called `eCoeff_0`—`eCoeff_4`). Figure 1 shows the mixing angles θ and ϕ of the three leading eigencoefficients for the Main Galaxy Sample (MGS) in SDSS Data Release 7 (Abazajian et al. 2009). These coordinates are defined as in Yip et al. (2004a) by

$$\phi = \tan^{-1}\left(\frac{\texttt{eCoeff\_1}}{\texttt{eCoeff\_0}}\right) \quad (1)$$

$$\theta = \cos^{-1}(\texttt{eCoeff\_2}). \quad (2)$$

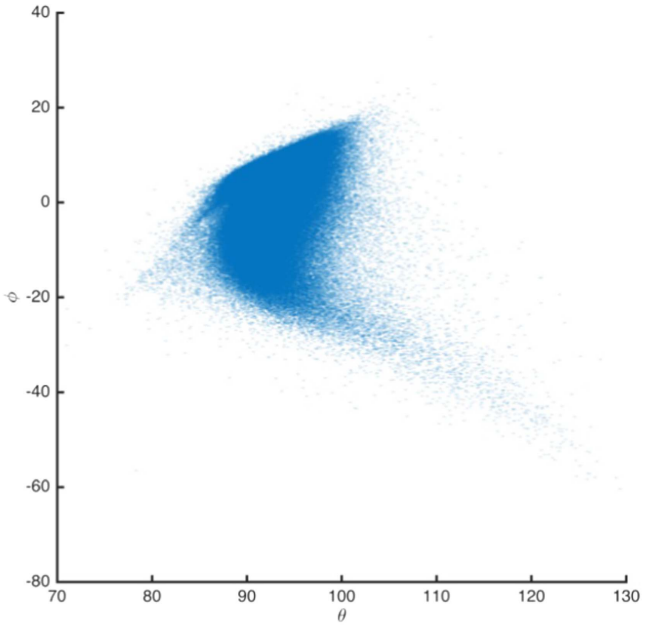

Figure 1. PCA provides a powerful compression for the Main Galaxy Sample of the SDSS Data Release 7. Every galaxy is summarized by a point in a low-dimensional space spanned by the eigenvectors. Here, we show the the mixing angles θ and ϕ of the first three eigencoefficients.

In the embedding illustrated in Figure 1, every point is a galaxy, and nearby points (i.e., galaxies or points which are near each other in the two-dimensional representation) have similar observations. This is achieved by construction of the empirical coordinate system. In particular, "red and dead" galaxies appear at the top of the plot, while star-forming blue ones are on the lower right.

The simplified view of the data provided by PCA does, however, have shortcomings. First of all, a significant fraction of the galaxies is actually removed from or scattered out of the plot. Consequently, we can usually only really see the core of the distribution in a figure such as this one. The lack of structure in this visualization is surprising, especially considering the large amount of high-quality data and the wide range of known galaxy types. Also, the interpretation of principal components is somewhat difficult, as they are linear combinations of input data vectors. Extensions and variants of PCA have been proposed to overcome these challenges, including non-negative matrix factorization (Lee & Seung 1999), the use of robust statistics (Budavári et al. 2009), and CX/CUR matrix decompositions (Mahoney & Drineas 2009; Yip et al. 2014). While these methods have alleviated some of the issues associated with PCA, the fundamental limitation of the linear model remained.

Perhaps the biggest conceptual change in the area was introduced by VanderPlas & Connolly (2009), who applied the Locally Linear Embedding (LLE) method of Roweis & Saul (2000). This more sophisticated empirical approach attempts to identify and exploit local structure in the data, and thus it breaks away from the straightforward global linear model underlying PCA. While there are other related nonlinear approaches (Tenenbaum et al. 2000; Belkin & Niyogi 2003), LLE in particular attempts to provide an angle-preserving mapping, which assigns coordinates to galaxies such that each galaxy is approximately a linear combination of its nearest neighbors.

The power and practical usefulness of LLE (as well as other related nonlinear methods including Tenenbaum et al. 2000; Belkin & Niyogi 2003; Coifman & Lafon 2006), however, is known to be severely diminished in many practical situations. The reasons for this are many, perhaps most notably that these methods are quite sensitive to realistic noise in the data and to the "details" of constructing the nearest neighbor graph. (This is in spite of a large body of theory stating that in idealized situations these details do not matter.) In addition to exploiting the strong algorithmic and statistical theory underlying our main method (Mahoney et al. 2012; Hansen & Mahoney 2014), dealing appropriately with these and other related practical graph construction issues will be central to our approach, and thus we postpone further discussion of this until Sections 2 and 3.

1.2. BPT Diagrams

Not all embeddings of astronomical data come from statistical procedures designed to assign new features to galaxies. In fact, any set of measurements extracted from the spectra could be considered as summaries for further analyses. A well known example is based completely on line measurements. The high resolution in wavelength often allows the identification and measurement of various spectral lines, and it is common to plot spectra in terms of carefully chosen line ratios. That is, while not usually described as an embedding method, the typical use of line measurements often involves embedding or mapping the data to a low-dimensional space.[7] In particular, the Baldwin, Phillips & Terlevich (BPT) diagrams (Baldwin et al. 1981) plot different line ratios on a logarithmic scale, enabling, e.g., the classification of galaxies (Brinchmann et al. 2004; Kewley et al. 2006). Figure 2 shows several BPT diagrams of the SDSS MGS. In the left panel, the characteristic V-shape of the embedding on the ratios N_{II}/H_{α} versus O_{III}/H_{β} is clearly visible, despite significant scatter, which is partly due to noisy measurements of the individual lines. Little to no structure is evident in the middle and right panels, whose x-axes plot different line strengths, but share the same y coordinates. The insight conveyed by the BPT plots can be considered complementary to that of the PCA results, which is primarily based on the continuum shape. As with PCA, these plots have significant scatter and, while we see some global trends, one would hope to see more subtle trends from a survey of hundreds of thousands of objects with close to 4000 wavelength elements.

In this paper, we present a novel approach to studying galaxies, which combines elements of several aforementioned techniques, but that moves away from the limiting assumptions in their underlying mathematical models. Our method, which is an extension of *semi-supervised eigenvectors* (Hansen & Mahoney 2014) from *locally-biased machine learning* (Mahoney et al. 2012), uses ideas from spectral graph theory and local spectral methods to study the properties of the SDSS MGS data.

1.3. Ideas from Graph Theory

The aforementioned LLE (Roweis & Saul 2000) along with ISOMAP (Tenenbaum et al. 2000), Laplacian eigenmaps

[7] The state-of-the-art in this area actually uses PCA as a preprocessing step to subtract the continuum, in order to better measure the line strength relative to this baseline (Tremonti et al. 2004).

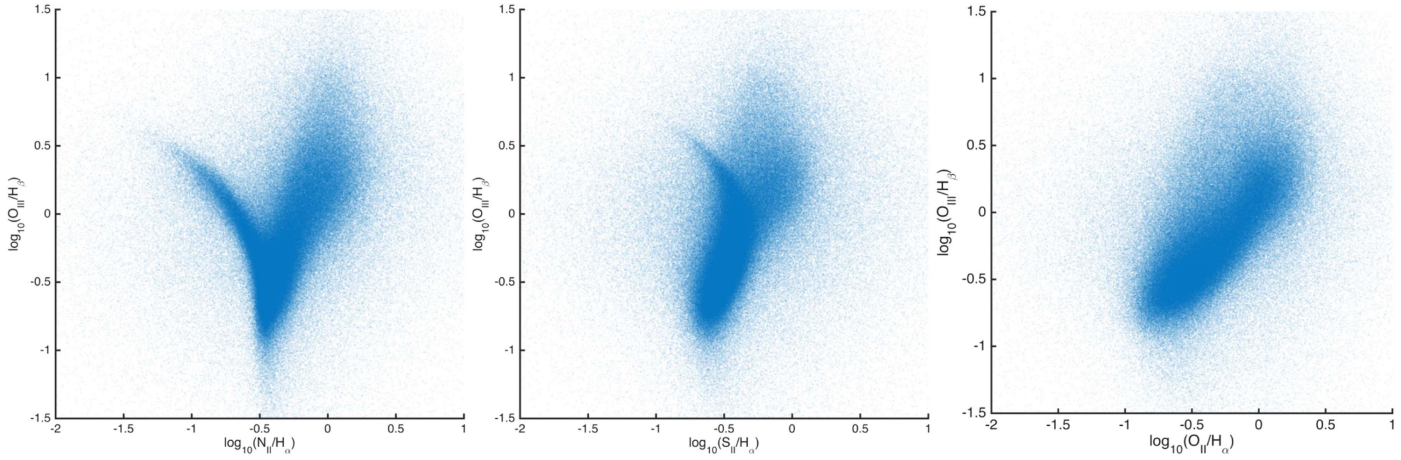

Figure 2. BPT diagrams of line ratios can also be thought of as a particular data-driven embedding, which provide insight not available from, but complementary to those provided by PCA.

(Belkin & Niyogi 2003) and the related diffusion maps (Coifman & Lafon 2006), recently generated a great deal of interest in statistical learning. These methods are also the precursors to our approach and share a number of common elements. They all first find for each data point the closest neighbors, and they focus on assigning new coordinates such that some neighborhood metric is preserved. For example, LLE preserves local angles and ISOMAP approximates the geodesic distance. In practice, they all solve an eigenproblem or a related global graph problem.

Here, we provide a very brief introduction to the most relevant ideas from spectral graph theory. For an elementary introduction, see Gallier (2013); and for more details on the particular method used in this paper, see Mahoney et al. (2012), Hansen & Mahoney (2014). Let us consider a matrix whose columns and rows correspond to our galaxies and the matrix elements encode their similarity to each other. For example, the k nearest neighbors could have the value 1 assigned to their corresponding matrix entry, and all other pairs would then have the value 0, indicating that they are not neighbors. This *adjacency matrix* A is large, but typically very sparse. Alternatively, we can also think of this network of connections as a graph, where the galaxies are nodes and the edges connect only those nearby. At this point, we can now define the so-called *Laplacian matrix* (also known as the graph Laplacian) as

$$L = D - A, \tag{3}$$

where D is the diagonal degree matrix, which simply counts the number of connections (or neighbors) each object has. It can be shown that the bilinear expression of L for any vector $\nu \in \mathbb{R}^n$ has the form

$$\nu^T L\, \nu = \sum_{\text{edges } (i,j)} (\nu_i - \nu_j)^2. \tag{4}$$

Hence, if our goal is to assign low-dimensional vectors to each graph node (or galaxy) i such that the new vectors are close if the pair is adjacent in the original data space, then we simply have to minimize the above formula and keep a small number of such eigenvectors. The trivial solution is that all $\nu_i = \text{const.}$, and this mode can be removed by using constrained optimization. Much like PCA is the eigensolution of the covariance matrix, the nonlinear Laplacian eigenmap is the eigensolution of the graph Laplacian.

If the adjacency matrix encodes different similarity measures, instead of just having $0/1$ entries, then one obtains weighted edges, and these propagate appropriately all the way into the eigenmap. Laplacian eigenmaps and diffusion maps use weights, which provide a gradual drop-off of the adjacency, which is motivated by the physical diffusion process. Hereafter, we will make use of these key ideas in the development of our method for galaxies.

In Section 2, we briefly review the approach, and in Section 3, we illustrate how to optimize the method for specific goals and data sets. In Section 4, we present the first results in the context of previously successful methods, and in Section 5, we illustrate how to *focus* the embedding on any objects of interest. Section 6 concludes the study and discusses possible future research directions.

2. GLOBAL AND LOCAL EMBEDDINGS

The nonlinear embedding introduced previously is considered *global*, in the sense that all galaxies are equally important in the construction of the graph and thus, the embedding map. The method of *locally-biased semi-supervised eigenvectors* (Mahoney et al. 2012; Hansen & Mahoney 2014), which we use, follows a similar approach, except that it introduces extra constraints to bias the embedding map creation toward objects of interest. These constraints, which are typically indicator vectors of *seed sets* of nodes, and the neighborhoods of these seed sets, will be best resolved in the locally-biased embedding, while the data far away will tend toward the origin and thus be less well resolved in the locally-biased embedding.

2.1. Similarity Graph

Given a collection of n data points $\{x_i\}_{i=1}^n$ in the d-dimensional space $\mathbb{R}^d$, in our case the full set of MGS spectra from SDSS DR7, we start the analysis by constructing the neighborhoods. These relationships form a graph which can be represented as a (sparse) matrix. To construct a weighted graph on the data, where the vertices are the data points x_i and the edges represent local connectivity information, we add an edge (i, j) to the graph if x_j is one of x_i's k nearest neighbors or if x_i is one of x_j's k nearest neighbors. Note that this ensures the adjacency matrix of the graph is symmetric. We then weight

each edge with a measure of local similarity given by

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right), \quad (5)$$

where $\|x_i - x_j\|$ is the Euclidean distance of the two high-dimensional data vectors and σ is a parameter, which controls the amount of "locality" present in the weight function. Alternatively, one can think of adding an edge between every pair of points, with weight given by Equation (5), in which case the value is near 0 for distant pairs of points. In practice, sparse matrices are preferred for computational reasons. The σ parameter can either be constant across all data points or it can be allowed to vary. A particularly useful choice often used in machine learning is $\sigma^2 = \sigma_i \sigma_j$, where σ_i is the distance of point x_i to its $k/2$ nearest neighbor. This accounts for the differences in spectra. Alternatively, one could choose to pick neighbors based on a distance threshold ϵ. We will study the effects of these tuning parameters later. We find that within reasonable ranges, the basic characteristics of the mappings remain similar. However, we also find that, unless due care is taken with astronomical issues, it is very easy to choose parameters, which overly homogenize the data and that lead to embeddings, which are much less useful astronomically.

2.2. Locally-biased Semi-supervised Eigenvectors

With the previously introduced *data graph* in hand, the new maps can be derived, which will solve the eigenproblem of the Laplacian matrix. Formally, the embedding assigns a low-dimensional vector to every high-dimensional data vector x, which in our case are the galaxy spectra. In our notation, the following ν components correspond to the new coordinates:

$$x \mapsto (\nu_2, \nu_3, ..., \nu_m). \quad (6)$$

The first element ν_1 is omitted as it corresponds to the trivial constant solution as discussed in the Introduction. The number of elements m is a free parameter similar to PCA, where the components are truncated based on variance arguments. Informally, the meaning of the new ν coordinates can be related to the "distance" on the data graph. In the embedding space, the standard Euclidean distance between two points is proportional to the average length of a random walk starting at one point and reaching the second. In this sense, the embedding given by these eigenvectors preserves "connectivity" information about the original data. For further details on these methods, we refer the reader to the theory[8] of Belkin & Niyogi (2003 and Coifman & Lafon (2006).

For the locally-biased embeddings, the key observation is that the optimization using the graph Laplacian can be further constrained to focus on a *seed set* of data points (Mahoney et al. 2012; Hansen & Mahoney 2014). (The choice of seed is based on the astronomical question of interest.) If s is the indicator vector of the seed set, then the locally-biased optimization becomes

$$\begin{aligned} \min_{\nu \in \mathbb{R}^m} \; & \nu^T L \, \nu \\ \text{s.t. } & \nu^T D \, \nu = 1 \\ & \nu^T D \, s \geqslant \sqrt{\kappa}, \end{aligned} \quad (7)$$

where the first constraint guards against the 0 solution and $\kappa \in [0, 1]$ is the correlation parameter, which controls the bias toward the seeds s. We can assume (without loss of generality) that s is properly normalized and orthogonalized so that $s^T D \, s = 1$ and $s^T D \, 1 = 0$. We also note that, for $\kappa = 0$, the locally-biased results coincide with the usual global objective, while for $\kappa > 0$, this produces solutions which are biased toward the seed vector s.

In this paper, we are primarily interested in understanding the features and properties of different mappings, and thus we adopt an "exploratory" approach. For other downstream learning tasks, e.g., classification or regression, various model-selection methods can be used to select k, σ, m, κ, etc. (Friedman et al. 2001). Extending the methodology of locally-biased semi-supervised eigenvectors to model selection and other related statistical questions, e.g., those considered in Richards et al. (2009), VanderPlas & Connolly (2009), is straightforward, but we do not consider it in this paper.

2.3. Implementation

The aforementioned optimization is related to a number of other methods, which are computationally more advantageous, e.g., random walks and linear equation solvers. We adopt an approach based on diffusion that is proven to solve a regularized version of the above problem. For details of these connections, we refer the interested reader to papers by Mahoney et al. (2012) and Hansen & Mahoney (2014).

Our computations of global diffusion embeddings were performed in MATLAB using a modified version of the `DiffusionGeometry` package of Bremer and Maggioni[9] (Coifman et al. 2005), and our computations of local embeddings were performed in MATLAB using the `sseigs_demo` package of Hansen and Mahoney.[10]

3. APPLICATION TO SDSS GALAXIES

We apply the new data-driven exploratory analysis to the well-studied collection of spectra in Data Release 7 (Abazajian et al. 2009) of the SDSS (York et al. 2000). Our goal is to explore the new data-driven parametrization of galaxies to test whether known trends appear in these low-dimensional maps.

3.1. Similarity of MGS Spectra

The MGS (Strauss et al. 2002) has become a testbed for a wide range of astronomical studies. There are several reasons for this, including the well understood selection function, large volume of high-quality data, and the prior systematic analyses, which serve as a reference for new techniques. In our study, we use the entire restframe wavelength range between 3450 and 8350 Å. Starting from the spectrophotometrically calibrated

[8] This theory is most relevant when the k or σ parameter is chosen to be relatively large, meaning that each data point has a relatively large number of neighbors. In that case, the data are typically homogenized and the corresponding matrix is relatively well approximated by a low-rank matrix. As our results below will show, this is often *not* the region of greatest astronomical usefulness, since in that regime small-scale or local structure in the data graph is largely lost. Indeed, this was the original motivation for the development of locally-biased semi-supervised eigenvectors (Mahoney et al. 2012; Hansen & Mahoney 2014).

[9] http://www.math.duke.edu/~mauro/code.html

[10] https://sites.google.com/site/tokejansenhansen/

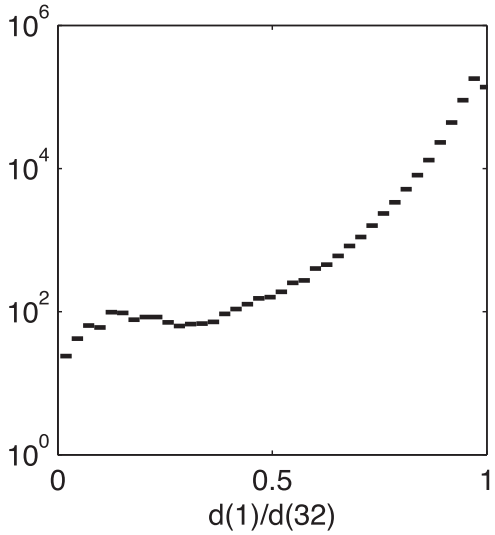

Figure 3. Histogram of ratios of distances to the 1st and 32nd nearest neighbors carries information about the distribution of galaxies in the high-dimensional space of observations. The vast majority of the 517,000 galaxies lie in very dense regions of the parameter space (corresponding to ratios near unity), while a non-negligible fraction lie in very sparsely populated regions (the bumps near zero).

normalized restframe spectra, our only preprocessing consists of dealing with gaps in the wavelength coverage. The missing parts of the spectra are filled-in based on the best-fit linear combination using eigenspectra from a prior PCA analysis following Yip et al. (2004a). While such preprocessing certainly simplifies our current implementations, we note that the approach is capable of dealing with incomplete and noisy data, e.g., similar to the gappy PCA of Connolly et al. (1995a).

Analyzing the distances of galaxy pairs reveals a heterogeneity. Figure 3 shows a histogram (in logarithmic scale) of the distance ratios of 1st and 32nd nearest neighbors over the entire data set. It is clear from the figure that the vast majority of galaxies lie in very dense regions of space (for which this ratio is close to one), while a non-negligible fraction lie in more sparsely populated regions (for which the ratio is less than one half.)

To build the graph of the MGS spectra, we *autotune* the bandwidths and fix the number of nearest neighbors. For computational reasons, we decided to start our exploration using the so-called *Markov normalization* (Belkin & Niyogi 2003; Coifman & Lafon 2006), which does not take the density of points into account.

3.2. First Results for SDSS

Figure 4 illustrates a global embedding. In the top-left panel we plot the second and fourth eigenvectors. We recall that the first eigenvector is the trivial mode, which corresponds to the 0 eigenvalue and has no information content, hence the second is the leading component. In this plot, every point is a galaxy and the color corresponds to the value of the second component (that is also shown on the horizontal axis). Unlike PCA, where a large fraction of the galaxies is typically scattered out from any reasonable plot, here we actually see all galaxies with only a small number of outliers that turned out to be pipeline errors or are completely mysterious looking (and are currently being investigated).

We see no clusters or classes of galaxies, which would separate from the others. Instead we observe that all spectra fall into a contiguous pattern. While this is in accord with our current understanding, the new method is the most powerful data-driven approach, which could detect such peculiarities. We see a lower envelope to the data points along which the spectra go from star-forming to red ellipticals. The boxes labeled A1–A5, R1–R5, and E1–E5 denote regions of embedding coordinates over which we calculated the mean spectra to illustrate the changes in the real spectra with increased signal-to-noise ratio. These delineations were admittedly drawn in an ad hoc manner, and they are meant only to indicate the changes along the embedding dimensions. The corresponding composite spectra are shown in Figure 4(b) through 4(d). From these average spectra, we can clearly see the correlation of the second eigenvector with the shape of the continuum: positive values of ν_2 tend to correspond to red, while negative values correspond to blue shapes. There also appears to be a correlation with the strengths of emissions lines, with negative ν_2 values corresponding to larger fluxes. A natural continuation of this trend is in the E1–E5 boxes shown in the Figure 4(c), where the lines become overwhelmingly strong in the young galaxies. At first glance, we see a dramatic change in not only the lines, but their ratios. For example, the [O III] (4959 Å, 5007 Å) lines grow significantly in comparison to the H α (6563 Å) line. In Figure 4(b), we present the mean and standard deviation of spectra in boxes A1–A5 of Figure 4(a). These spectra trace a trend, which is different from the main direction, but we see increasing line strengths and bluer spectra. This population of galaxies will be immediately obvious in Section 4.4, where we show the results of the SDSS classification (Brinchmann et al. 2004): these are active galactic nuclei (AGN).

3.3. Effects on the Embeddings

The embedding shown before is the result of a careful study of the parameters, which affect the mapping to varying degrees. We start with the choice of k, the number of nearest neighbor edges on the constructed graphs. While, in general, the effects of the choice can be highly problem-dependent and hence, it ultimately should be determined by a downstream astronomical selection criterion, e.g., optimizing the mean-squared error, or a precision-recall metric if one is performing classification. We have noted several trends of interest. For large k values, the maps emphasize relations across the entire data set and one tends to identify well these large-scale or global structures in the data, while washing out or homogenizing small-scale or local structure, which is often of interest. The global nature of these embeddings is often referred to as large-scale structure in the machine learning community. For small k, the emphasis is on the connections of the closest objects or at shorter scales.

In Figure 5, we show the embeddings of the full data set on the third and fourth eigenvectors for k ranging from 2^1 to 2^{11} by factors of four.[11] The points are color coded by the value of the second eigenvector, as before. These plots hide density

[11] The "correct" value of k may well depend on the density of points at various places on the graph, but it is also a non-trivial statistical model-selection problem (Friedman et al. 2001). For example, it depends on whether one is interested in identifying properties of very large clusters in the data or properties of very small clusters in the data. Our point here is simply that different values of k can be used to identify very different properties in the data.

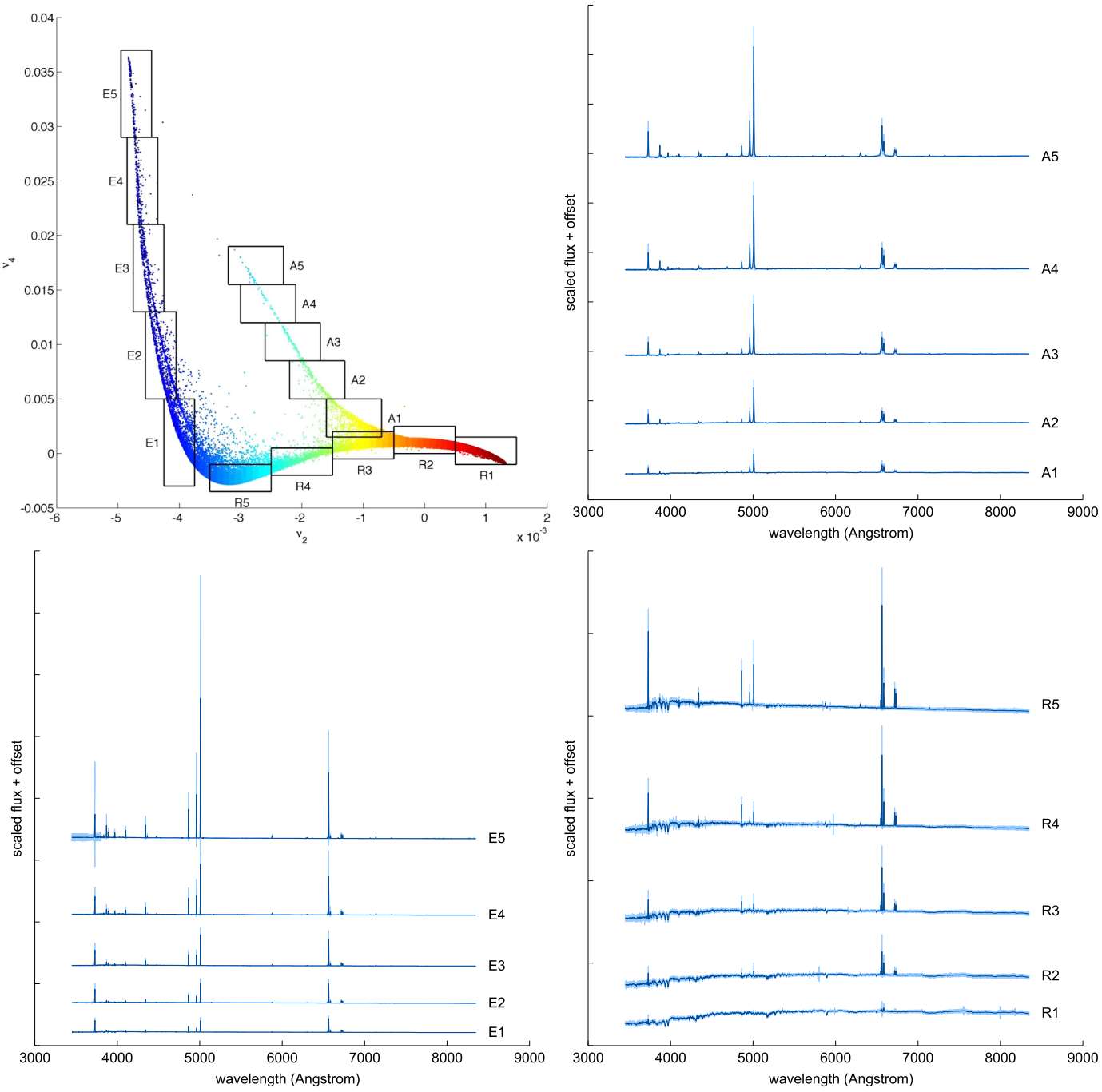

Figure 4. Leading components of our nonlinear mappings provide great detail of galaxy diversity. Every point in the top-left panel corresponds to a galaxy colored by the first non-trivial component. The R-, E-, A- labels correspond to trends that the boxes intend to illustrate: R for red to blue, E toward extremely blue, and A for active galactic nuclei. The other panels show the composite spectra for objects in the labeled black boxes. The light blue envelopes around the spectra illustrate the one standard deviation.

information, but they illustrate that as k decreases, the structures tend to become a skeleton, e.g., for $k = 2$, where very fine-scale structure can be seen. In particular, the red part of the embedding, which corresponds to redder spectra, varies considerably as k is adjusted, and small-scale local structure such as the cyan "heel" pointing downward and to the right for $k = 2$ is lost for larger values of k. For the largest k, the mappings start to lose some of the subtleties. We seem to find a good balance for $k = 32$, which is what we adopt as the fiducial setting for the rest of the paper, but we emphasize that other smaller values of k may be more appropriate for finer-scale analyses.

To make these informal observations somewhat more quantitative, we study the eigenvalues as well as the eigenvectors used in our embeddings. In Figure 6(a), we plot the decay of the top 101 eigenvalues for values of k ranging from 2^1–2^{11} by powers of two. From the figure, we can see that as k is increased and thus, as more edges are added to the graph, the rate at which the eigenvalues decay increases, i.e., the matrix is more well-embeddable in a low-dimensional space.

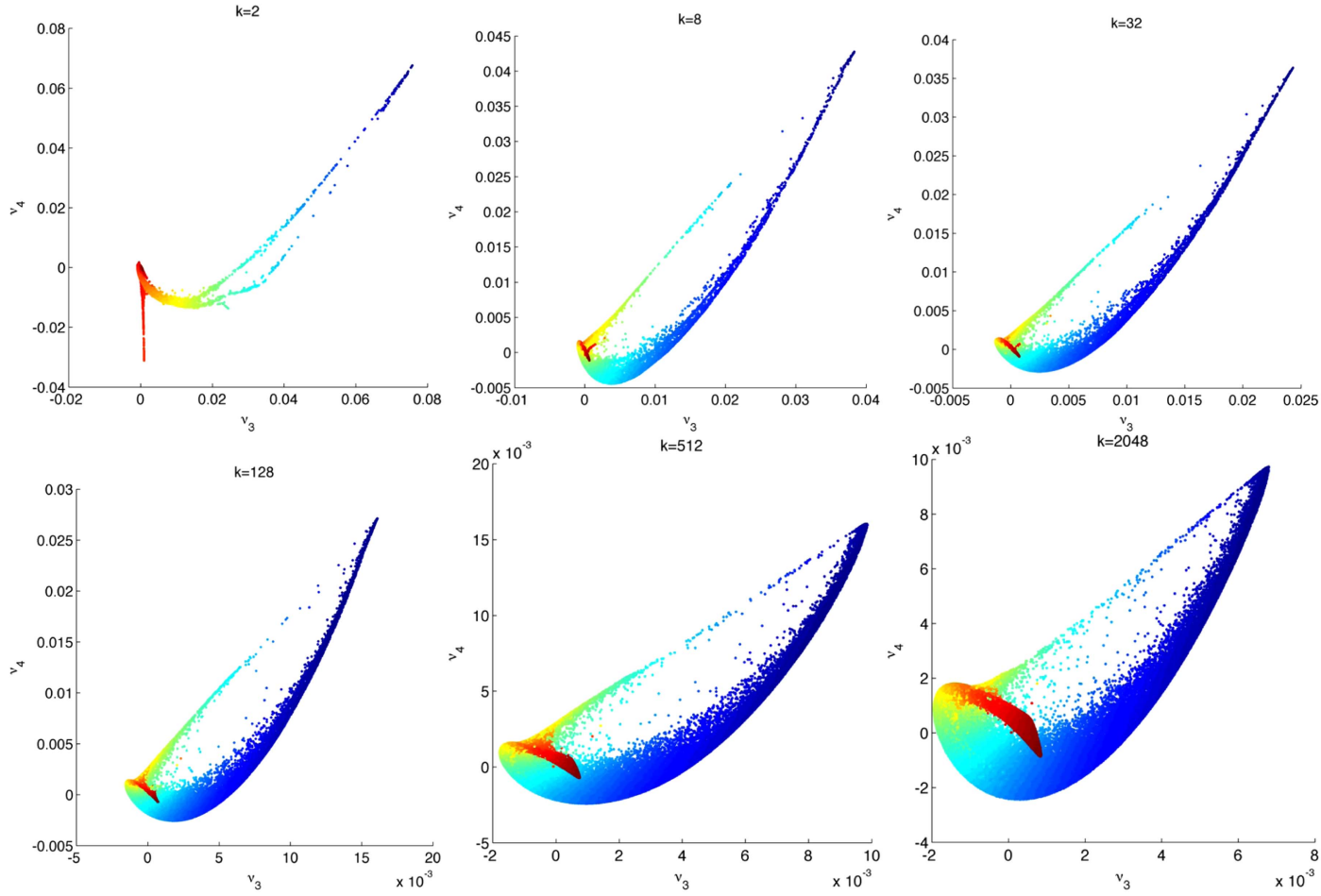

Figure 5. Series of plots show the variations in the embedding coordinates as the number of neighbors is adjusted to $k = 2^n$, $n \in \{1, 3, 5, 7, 9, 11\}$. Points are color coded by the value of the most important eigen-coefficient and the axes correspond to the next two components.

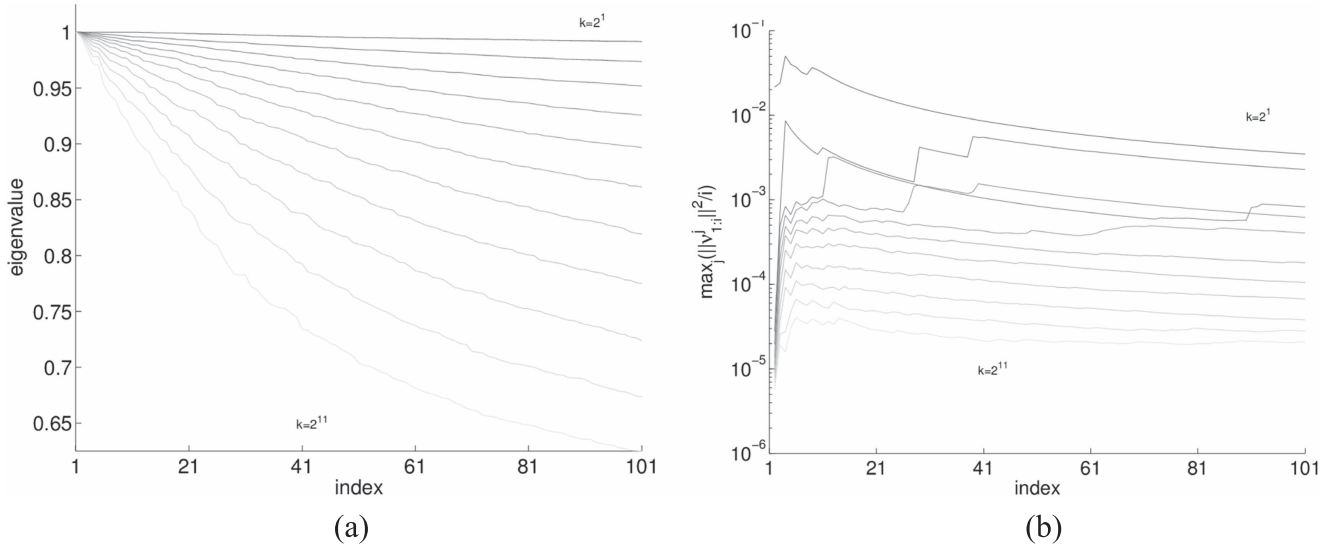

Figure 6. Empirical spectral properties, i.e., eigenvalues and eigenvectors, of embedding matrices as a function of parameter k. (a) Top 101 eigenvalues of the lazy Markov operator with autotuned bandwidths, for $k = 2^1$ (red) to $k = 2^{11}$ (yellow). This illustrates faster eigenvalue decay as k increases, meaning that for small k there is more heterogeneous structure not well approximated by a low-rank space. (b) Max-to-median ratio of eigenvector norms, as a function of the embedding dimension for the lazy Markov operator with autotuned bandwidths, for $k = 2^1$ (red) to $k = 2^{11}$ (yellow). This illustrates more smooth eigenvectors as k increases, meaning that for small k there is more heterogeneous local structure.

Next, in Figure 6(b), we plot the ratio of the largest eigenvector norm to the median eigenvector norm, as a function of the index of the eigenvectors. Specifically, if $\nu^j_{1:i}$ is the embedding of spectrum j on eigenvectors 1 to i, we plot the quantity

$$\frac{1}{i}\max_{1\leqslant j\leqslant n}(\|\nu^j_{1:i}\|_2^2), \tag{8}$$

as a function of i. In general, the eigenvectors become more uniform with increasing k, and local heterogeneities—as captured by localized eigenvectors—become less prominent with the inclusion of additional edges.

We note that for these values of k, the Markov matrix is not particularly well approximated by a low-rank matrix. Nevertheless, the leading eigenvectors do correlate well with physical intuition, and they do provide meaningful low-dimensional representations of the data. We should also note that these effects of increasingly localized eigenvectors and slower eigenvalue decay, as the data graphs are made sparser, have been observed previously in connection with the Nyström method in large-scale machine learning applications, as in Gittens & Mahoney (2016). (While a detailed discussion of this is beyond the scope of this paper, see Gittens & Mahoney 2016 and in particular the discussion of the statistics on the sparsified radial basis function kernels.) Our results are consistent with this prior work (Gittens & Mahoney 2016), which demonstrated that sparser graphs (within a parameterized family such as with radial basis function kernels or k-NN graphs) are much less "nice" in the sense that their eigenvalues decay more slowly and that their eigenvectors are more localized.

4. GLOBAL STRUCTURE VIA GLOBAL EMBEDDINGS

Our approach yields results that resemble our intuitions formed by decades of galaxy research. The goal for future research is to understand the empirical parameters in the context of synthetic models and perform inference using the embedding coordinates, but here we discuss some observations of the data-drive methodology. While these are not new astronomy results, they illustrate the usefulness of such embeddings, which could originally have been used to make the discoveries.

4.1. Red Galaxies, Trends and Outliers

Figure 7 illustrates how the new method can be used to study the diversity of red galaxies. Panel (a) shows the second (ν_2, i.e., the first non-trivial) and fifth (ν_5) component. The latter discriminates red galaxies in the very dense region of spectrum space better than other low-order eigenvectors. As in Figure 4, color corresponds to the value of ν_2, and we have drawn bounding boxes in an ad hoc manner.

In Figure 7(b), we present the mean and standard deviation of spectra in boxes RR1–RR5 from panel (a). It is clear that all spectra in this area of the embedding share a red continuum shape, with small or absent emissions lines. We also see the increasing strength of prominent Hα line with increasing values of ν_5. Panel (c) is the same for boxes RG1–RG5, which correspond to a region of high density. The transition from red to blue continuum shape is perhaps the most evident in this small region of the embedding space.

Finally, for completeness, in Figure 7(d) a set of outliers is shown, labeled O1–O5 in panel (a). Spectra O4 and O5 have been scaled by factors of $1/10$ and $1/100$, respectively, for legibility. Outlier detection is a critical and important area to remove artifacts from the data and to identify potentially new types of objects or phenomena. All of these outlier spectra appear to be artifacts of the pipeline or the gap-correction preprocessing step. We note that each of these erroneous spectra appear separated from the remainder of the data, thus indicating the robustness and usefulness of the method for identifying outliers.

4.2. Relationship to `eCoeff`

Next, we compare the new embeddings with those obtained via PCA, which is a dimension-reduction method that optimally preserves linear structure in high-dimensional data sets. These embeddings are computed in the SDSS pipeline and stored as `ecoeff_i` for `i` ranging from 0–4. These coefficients are routinely used, for example, to classify galaxy types (Yip et al. 2004a).

In Figure 8, we plot the galaxies in the embedding on the mixing angles θ and ϕ. The opacity is proportional to the density and the coloring is determined by the eigenvector in the left panel, and by eigenvector 5 on the right. From the former, it is clear that ν_2 and ϕ are highly correlated. Given our results in the previous subsection, this is not surprising, since both measures mediate between red and blue continuum shapes. We confirm that the principal trends from red to blue captured by previous methods is the leading term of the new embedding. We omitted plots showing that ν_3 and ν_4 discriminate among blue spectra. The right panel displays how ν_5 picks up in a more complicated way variation among red spectra.

4.3. Relationship with BPT Diagrams

Next, we compare the new embeddings to BPT diagrams, which are based on the flux in four wavelength bins, corresponding to the N_{II}, H_α, O_{III}, and H_β emissions lines. Figure 9 contains the usual BPT diagrams of line ratios

$$\log_{10}\left(\frac{N_{II}}{H\,\alpha}\right) \text{ versus } \log_{10}\left(\frac{O_{III}}{H\,\beta}\right) \text{ versus } \log_{10}\left(\frac{O_{II}}{H\,\beta}\right).$$

The opacity is again proportional to the density and the coloring is determined by the first non-trivial eigenvector ν_2. These color versions of Figure 2 show that the apparent bifurcation in the BPT plots is really a continuous and gradual change. The most dominant component shown in color resolves the degeneracies in these plots and shows that the scatter is primarily due to the red galaxies. These plots hint at the possibility of a better classification algorithm, which uses the coordinates provided by our method, potentially in combination with the traditional line measurements. The advantage of using the new approach is that the ν coordinates summarize the entire spectrum and are not limited to the small wavelength regions on which the line measurements are based. This makes the quantities less noisy and the subsequent inference more robust.

The comparison of the apparent structure in the embedded coordinates of galaxies in Figures 1 and 4, along with the illustrations of similarities to PCA and BPT in this section, suggests that the new mapping will be a useful tool to study galaxy diversity. It seems to capture information about both the continuum shapes and the spectral lines in a single analysis. In the next section, we look at the results in the context of galaxy classification.

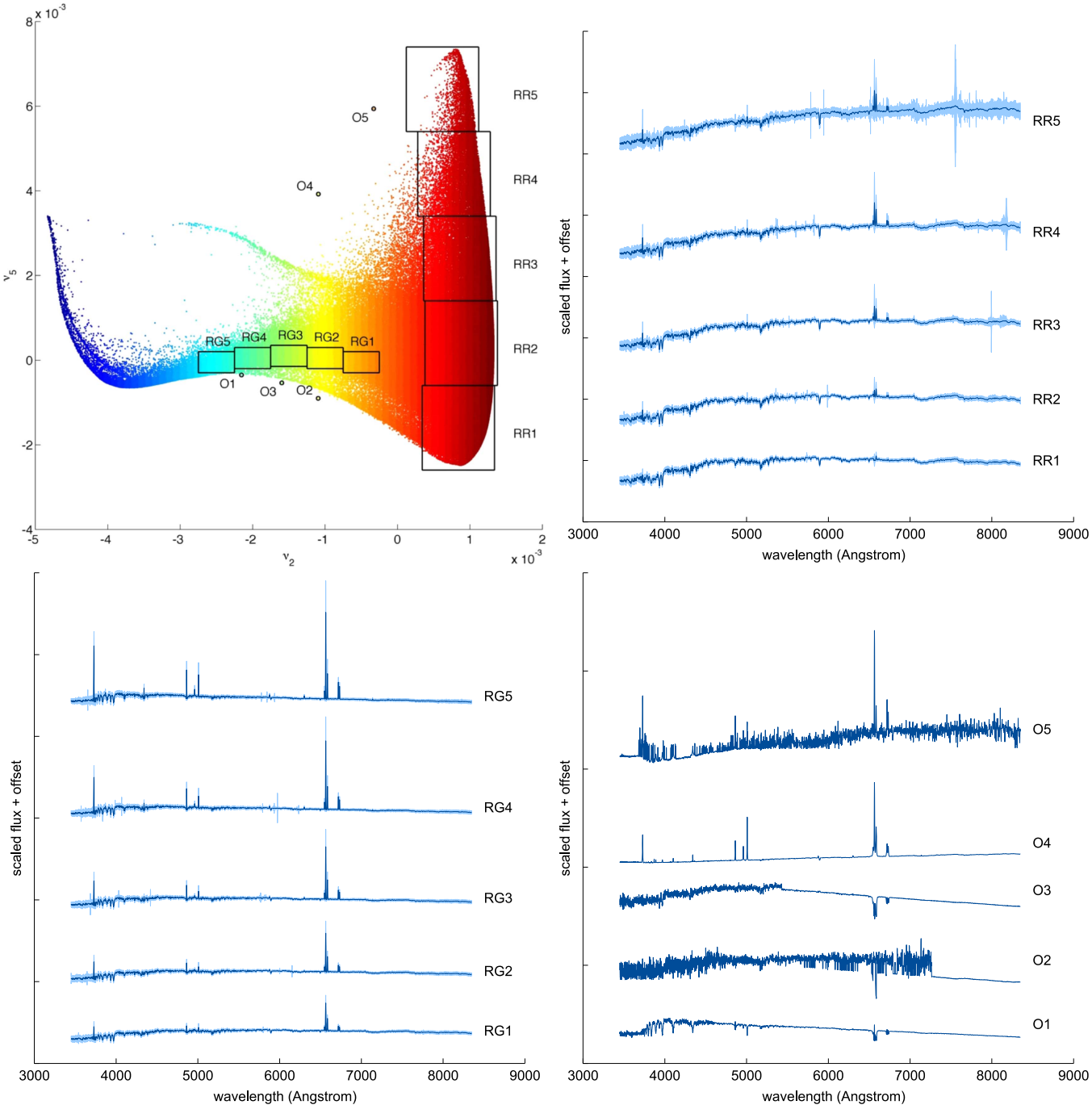

Figure 7. Higher-order components help distinguish different types of galaxies. In the projection shown in the top-left panel, the red galaxies are resolved better by the 5th component. The other panels show composites of the appropriate boxes and illustrate the red–blue transition (RG1-5), the diversity of red galaxies (RR1-5) and some of the outliers (O1-5).

4.4. Comparison to the SDSS Classification

We augment the embeddings with the addition of the SDSS class labels as derived by Brinchmann et al. (2004). Our sample of about 517,000 spectra are split into the following six categories: star forming, low signal-to-noise star forming, composite, AGN, low signal-to-noise LINER; and unclassified. This state-of-the-art classification scheme goes beyond the BPT diagrams and uses a total of seven lines to distinguish the separate classes. See Figure 10 for BPT line-ratio diagrams with galaxies color coded by class label. The linear class boundaries through high-density regions of the data, as well as a quick comparison with Figure 9(a), highlight the arbitrariness of the class boundaries.

In Figure 11, we present the embeddings on eigenvectors 2 and 5 of the same graph Laplacian of the galaxies. This is the same plot as Figure 7(a), but now with points color coded according to their type: blue for star forming, cyan for low signal-to-noise star forming, green for composite, magenta for AGN, and red for LINER. We have omitted unclassified spectra from these figures in order to make the embeddings more legible. One can discern a transition from star forming to composite to LINER with increasing ν_2. This agrees with the earlier remarks regarding the correlation of ν_2 with continuum

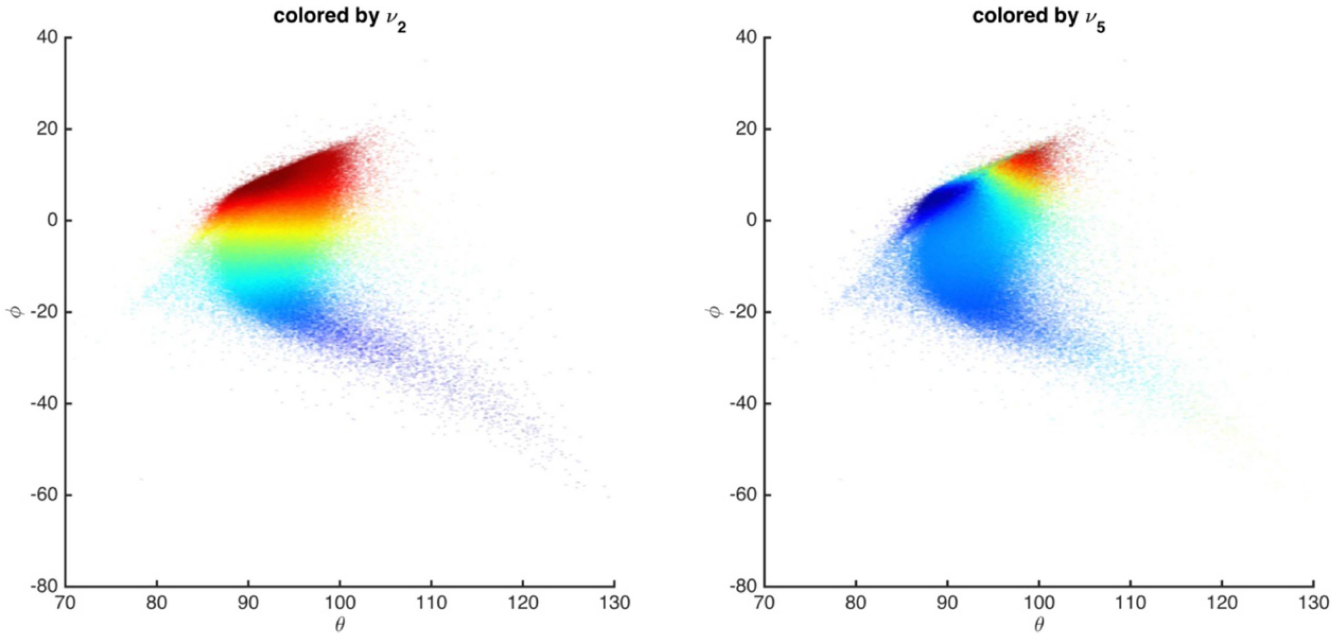

Figure 8. Leading components of the PCA and our nonlinear embedding show similar trends, as can be seen on the left, where the galaxies are plotting on the mixing angles, but colored by the first leading component (ν_2) of the new analysis. The next two components differentiate between blue galaxies, not shown in this figure, and ν_5 resolves the red, as also shown in the previous figure.

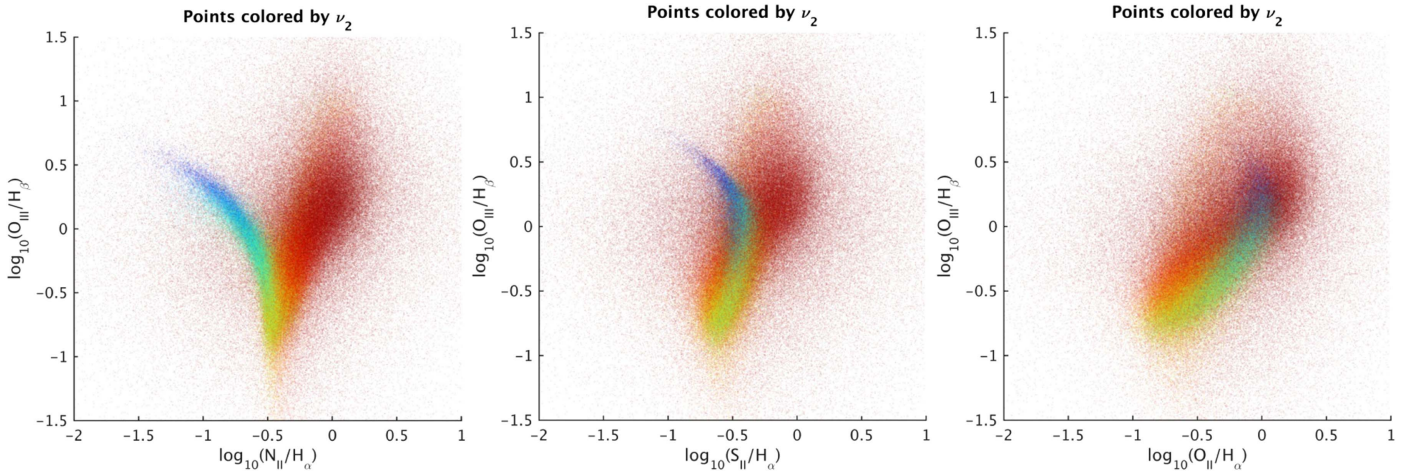

Figure 9. BPT line-ratio diagrams, color coded by the leading component of the new embedding, reveal previously hidden trends.

shape. In addition, as noted previously, the AGN form a separate "spur" between the star-forming galaxies and LINERs. We note the concentration of composite galaxies, which exhibit a transition from blue to red continuum shapes. We also note the concentration of LINER and low signal-to-noise star-forming galaxies along the lower right rim of the data set, which are difficult to distinguish from one another. We see a separation between the bulk of the spectra and those labeled as AGN. Other projections show a clear boundary between LINERs and AGN, which does not appear to be data driven, but astronomically motivated (Brinchmann et al. 2004).

4.5. Bimodality of the Blue Ridge

Upon closer inspection of Figure 4(a), the blue sequence of galaxies actually appears to split: there are two parallel ridges going through the the boxes E1–E3. As with the unexpected resolution of the AGN, this feature of the map was also not foreseen and is most tantalizing at first. To highlight and understand better this apparent bimodality, consider Figure 12, where we present a map of higher-order eigenvectors. The separate trendline is clearly visible. The key insight, however, comes from the color scheme, which represents the redshift of each galaxy. We see that one of the strands contains the lowest redshift spectra in the entire MGS with $z < 0.02$. Upon examination of the individual sources and their postage stamps from the SkyServer Image Cutout service, we determined that this feature is in fact an artifact. The SDSS `photo` pipeline is known to break apart large galaxies, and here we witness its power to pick out individual star-forming H II regions, which the target selection identified as galaxies. Their morphology is in complete agreement and in some of the cases the enhanced star formation appears to be induced by merging galaxies.

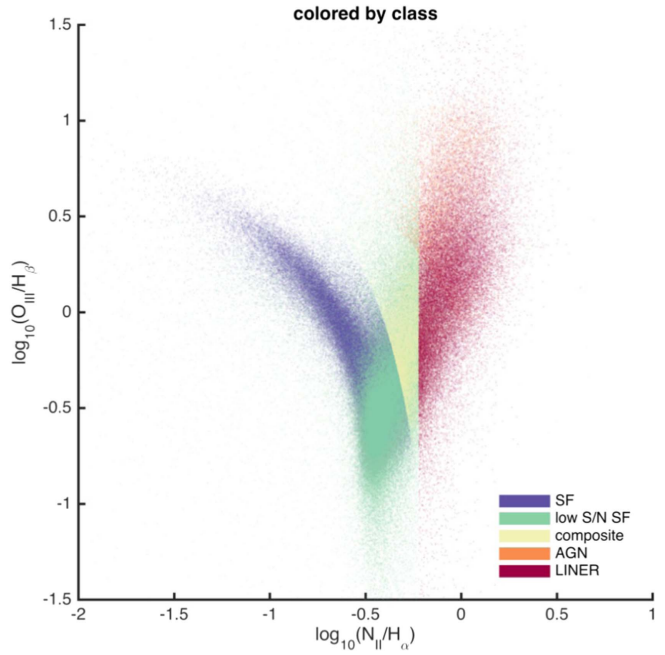

Figure 10. BPT line-ratio diagram colored by the SDSS classification scheme is comparable to the previous figure.

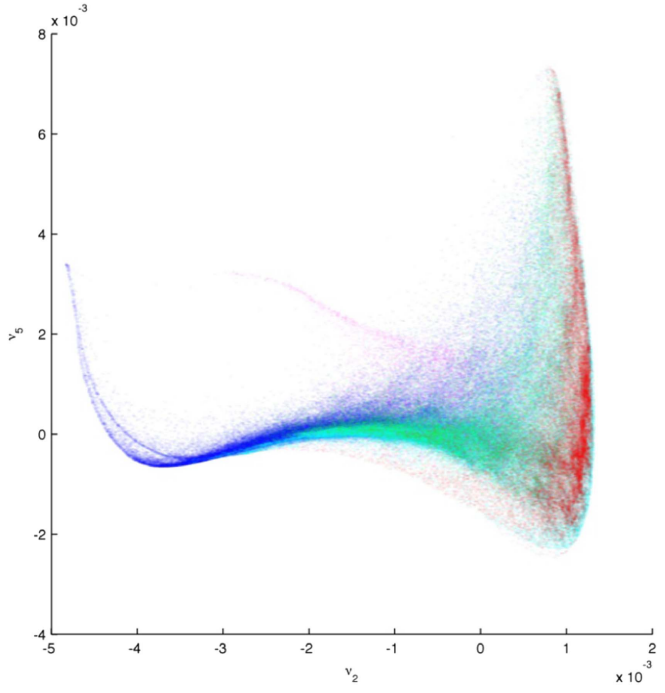

Figure 11. Using the same ν_2–ν_5 projection of our new embedding, we show all classified galaxies in the DR7 MGS in Brinchmann et al. (2004). In comparison to the previous two figures, this map conveys global trends, which are in accord with our intuitions, e.g., the AGN shown in magenta nicely separate from the star-forming galaxies in blue, and the composite types in cyan are in the transition region. We also see unexpected trends, such as the bimodality of the blue branch, see text.

5. LOCAL STRUCTURE VIA LOCAL EMBEDDINGS: FOCUSING AROUND SELECTED OBJECTS

While embedding maps provided by global variants of locally-biased semi-supervised eigenvectors provide a great deal of information about the structure in the data and the diversity of the galaxies, locally-biased variants are expected to

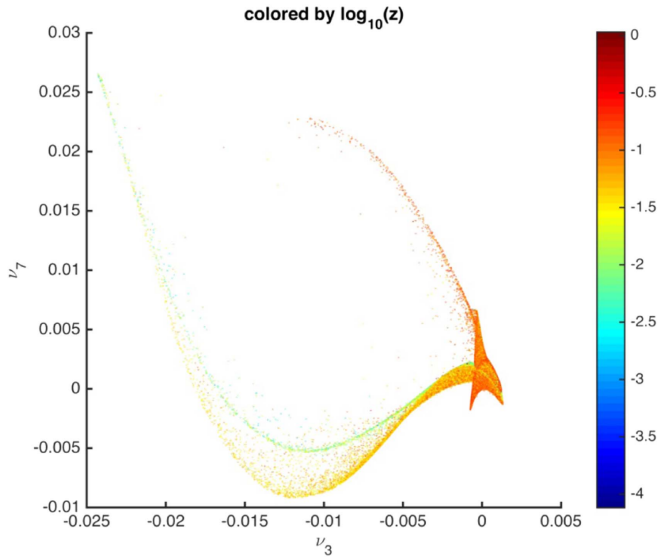

Figure 12. Higher-order components further differentiate the bimodality in the blue galaxies. The solution to this puzzle is revealed by coloring the galaxies by their redshifts. The green strand contains H II regions targeted as galaxies by the photometric pipeline.

go well beyond this in terms of resolving power for any points of particular interest. In this section, we investigate locally-biased embeddings and the choice of seed vectors s introduced earlier in Section 2. (Recall that the seed vector s in Problem (7) is an indicator vector of the seed set of nodes, i.e., it is a vector indicating which galaxy spectra the method should be biased toward.) In Figure 13, we compare global and locally-biased mappings. In the left panel, we plot the galaxies on global eigenvectors 2 and 3, and color the points by the second global eigenvector, as before. Also indicated with black outlines are two subsets of galaxies chosen somewhat arbitrarily on the AGN branch and among red galaxies. These seed sets were defined by manually choosing one galaxy of interest and taking its 100 nearest neighbors in the global embedding space shown in the left panel. (That is, the seed vectors for the embedding shown in the second and third panel were indicator vectors for one or the other of the black sets of nodes in the first panel.) Using biased mapping with κ set to $1/4$ for each eigenvector, we create two embeddings for the two sets of seeds and plot the results in the middle and left panels using the same color scheme. The plots of the two leading non-trivial eigencomponents appear very different and are dominated by different populations of galaxies even though all DR7 MGS galaxies are plotted in all three panels. In the local setting, galaxies similar to the seeds are drawn away from the bulk of the embedding, offering a "zoomed-in" view of the local region of interest, and galaxies very dissimilar to those in the seed vector are given lower importance and clumped together. With the new focus, these galaxy maps reveal very different subtle trends that were previously hidden in the global view, and galaxies unlike the seeds are marginalized in a corner of the figure. The full story of these trends is still unfolding and will be studied in future work.

5.1. Locally-biased Learning

To illustrate how the local analysis can be used to perform improved classification, we constructed a set of five locally-biased embeddings, one for each class per the definition of

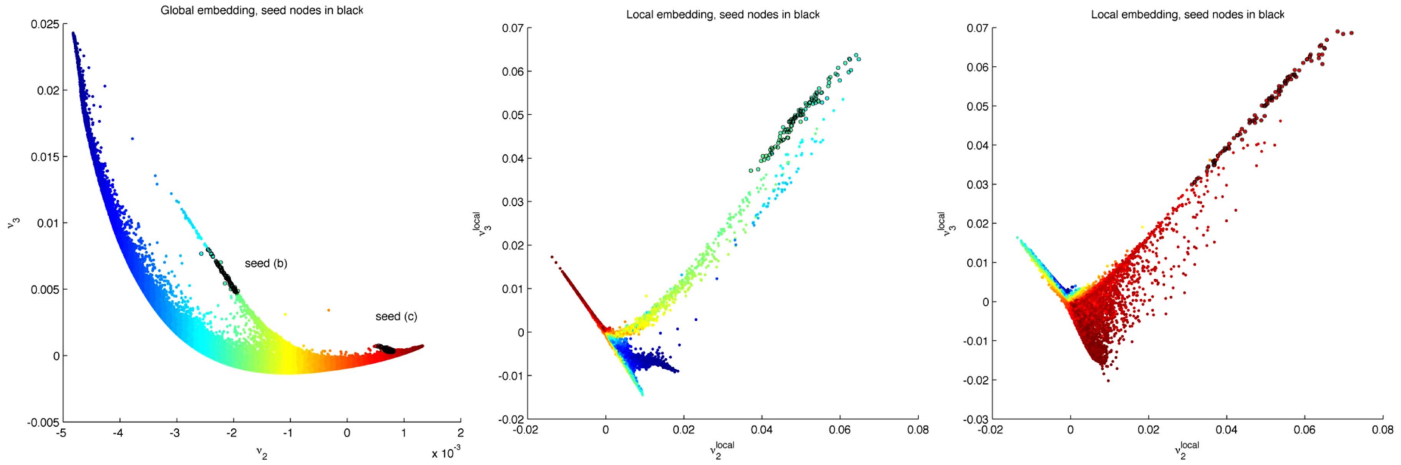

Figure 13. Two sets of black points in the initial global embedding (left) are selected as seed for further examination. The middle and right panels illustrate the locally-biased embeddings focusing on these two sets of points. The new maps reveal local features previously hidden in the global embeddings.

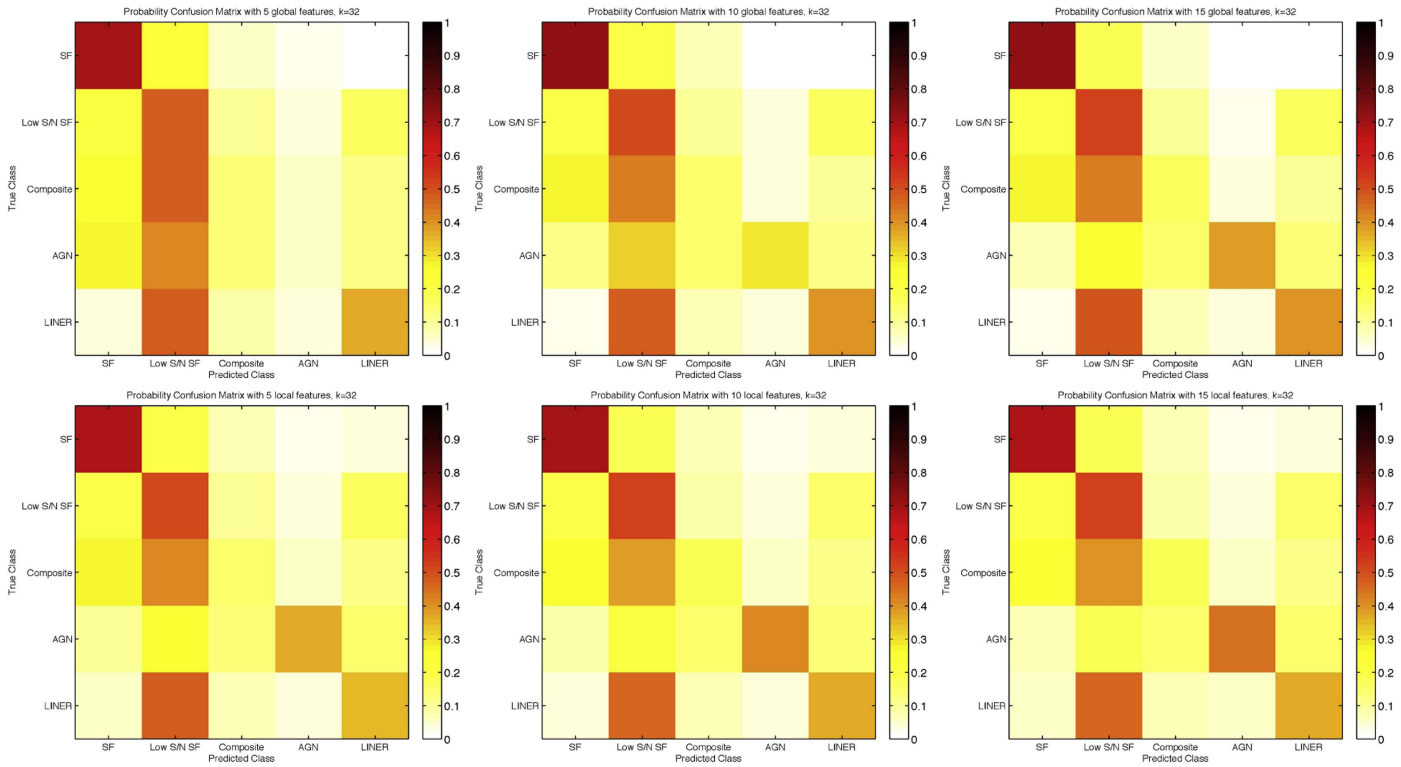

Figure 14. A suite of galaxy classifiers was developed on the embedding coordinates. The top row shows the confusion matrix of different types as a function of leading components in the global setting. We tested the top 5, 10, and 15 most important terms shown from left to right. The bottom row illustrates the same for the locally-biased case. We see faster improvement in the latter, especially in the case of the AGN.

Brinchmann et al. (2004). Our goal is not to recover the same class memberships for all galaxies, but rather to look at the difference in the light of the diversity maps of enhancing the neighborhood of sample galaxies in known classes.

In particular, we choose random spectra of each class and set the corresponding elements of s to $+1$. To increase the contrast, we further choose the same number of random spectra from the other classes and set the corresponding elements of s to -1 (Friedman et al. 2001). We expect that the local embedding with such a seed vector will better separate class c from the remaining spectra. For this study, we varied n in a wide range and chose to illustrate the application using $n = 640$. We calculate the top nine local eigenvectors for each class, with uniform correlation parameter $\kappa = 1/9$. We then trained a five-class logistic regression model using the global and local embeddings as features. We grouped the features together into sets of five, taking a fixed number of local eigenvectors per class. Thus, we tested one model using the top local eigenvector from each class, another using the top two local eigenvectors from each class, and so on. We compared these with models built using the same number of global eigenvectors. For each number of features, we cross-validated the model using ten folds with a 10%/90% train/test split.

The logistic regression model returns a vector of probabilities that a spectrum in the test set belongs to a given class. Using these probabilities, we constructed *confusion matrices*,

whose i, j entry represents the average probability (over the test set) that a spectrum of type i is assigned to type j. The results are shown in Figure 14 for models constructed with 5, 10, and 15 eigenvectors. The top row corresponds to the increasing number of global eigenvectors and the bottom to local eigenvectors. Comparing the top and bottom left panels of Figure 14 , we note that with five global features, the model has difficulty correctly classifying AGN. However, using five local features leads to a significant increase in the correct classification of such spectra. Thus, for a fixed budget of features, the local model outperforms the global version, as expected. Increasing the number of features, global or local, leads to improved performance on this rare galaxy type, and the local model consistently outperforms its global counterpart.

We also note that for all models shown, there is a large ambiguity between low signal-to-noise star-forming galaxies and LINERs. Indeed, for all cases, the probability that a galaxy labeled LINER is classified as low signal-to-noise star forming, is greater than the probability that it is correctly labeled. This is not surprising in light of Figure 11(c), where significant mixing of these two types, colored cyan and red, respectively, is visually evident. We find this result to be very robust to the choice of graph construction, i.e., the k nearest neighbors.

5.2. Labeling Unclassified Spectra

There is a large number of galaxies without reliable classification, and one of our goals was to see if the embeddings provide a way to classify these. We found that this is very challenging and fundamentally limited by the data. The reason is that unclassified spectra (even when they are not obvious outliers caused by experimental artifacts) typically have properties which are *very* different than any classified spectra. That is, they are unclassified for a good reason, i.e., since they are very different than spectra in one of the main classes, and in some sense they form their own (diverse) "other" class, whether viewed from a global or a locally-biased perspective.

6. CONCLUSION

We have presented the method of *locally-biased semi-supervised eigenvectors* (Mahoney et al. 2012; Hansen & Mahoney 2014), which is a novel data-driven technique, to study the diversity of galaxy spectra in the SDSS DR7 MGS. By constructing low-dimensional maps, which respect the local spectral similarity, we are able to visualize a range of known astronomical phenomena, e.g., the continuous transition of blue galaxies to red ones or the varying strength of AGN. Unlike previously used methods such as PCA, our method can focus on local properties and subtle trends in the data, in addition to the global context. These aspects of the maps of all galaxies can be varied by just a couple of parameters.

In particular, we (1) studied the new method in the context of PCA and demonstrated unprecedented detail in our data-driven maps. The new parameters clearly track changes in the continuum shape and the strengths of spectral lines. We confirmed that there are no disjoint groups of galaxies to indicate natural classes, but instead there are smooth continuous transitions between different types of objects. We found that prominent features of these maps correspond to known astrophysical phenomena such as star formation and AGN. We also found (2) a tantalizing bimodality in the bluest observations, which turned out to be the result of the photometric pipeline detecting H II regions in the closest galaxies. We studied (3) the relation of the new results to BPT diagrams as well as the state-of-the-art classification methods and proposed the use of continuous parameters instead of rigid boundaries in astrophysical studies. The derived empirical parameters summarize entire spectra and hence carry more information than measurements extracted in narrow wavelength ranges and could offer a more robust alternative. We developed (4) new classifiers and attempted to categorize the "unclassified" spectra with modest success, due to the inherent limitations of the data. Furthermore, we (5) demonstrated that locally-biased maps emphasize subtle trends, which can be used as a general data exploration tool to focus on galaxies of interest. In the new maps, oddities and artifacts are automatically separated and can be discovered by visual inspection or the usual machine learning tools, which makes our method a great candidate for outlier detection.

Based on these results, the new method of locally-biased semi-supervised eigenvectors may be viewed as a new type of *computational microscope* for astronomical data. It can not only reproduce known properties of the data or identify outliers and artifacts, as done in this paper, but can be used as a novel research tool to enhance subtle trends around selected objects and potentially to facilitate new discoveries. We plan future studies to improve galaxy classification and derive continuous spectral models, which in turn can be used for subsequent analyses, e.g., for photometric redshift estimators, and to better understand galaxy evolution when combined with stellar population synthesis models.

This work was partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. TB would like to acknowledge partial support from the National Science Foundation via ACI-1261715 and the Astrophysical Research Consortium via SSP430. MWM would also like to acknowledge the Defense Advanced Research Projects Agency and the Department of Energy for providing partial support for this work.

REFERENCES

Abazajian, K., Adelman-McCarthy, J., Agüeros, M., et al. 2009, ApJS, 182, 543
Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Belkin, M., & Niyogi, P. 2003, Neural Comp., 15, 1373
Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. 2000, AJ, 120, 1588
Budavári, T., Wild, V., Szalay, A., Dobos, L., & Yip, C. 2009, MNRAS, 394, 1496
Coifman, R., & Lafon, S. 2006, App. Comp. Harmon. Anal., 21, 5
Coifman, R., Lafon, S., Lee, A., et al. 2005, PNAS, 102, 7426
Connolly, A. J., Budavári, T., Szalay, A. S., Csabai, I., & Brunner, R. J. 1999, in ASP Conf. Ser. 191, Photometric Redshifts and the Detection of High Redshift Galaxies, ed. R. Weymann et al. (San Francisco, CA: ASP), 13
Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995a, AJ, 110, 2655
Connolly, A. J., & Szalay, A. S. 1999, AJ, 117, 2052

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995b, AJ, 110, 1071
Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, ApJ, 398, 476
Friedman, J., Hastie, T., & Tibshirani, R. 2001, The Elements of Statistical Learning, Vol. 1 (2nd ed.; New York: Springer)
Gallier, J. 2013, Notes on Elementary Spectral Graph Theory Applications to Graph Clustering Using Normalized Cuts, Tech. Rep. (arXiv:1311.2492)
Gittens, A., & Mahoney, M. W. 2016, J. Mach. Learn. Res., 17, 1
Hansen, T. J., & Mahoney, M. W. 2014, J. Mach. Learn. Res., 15, 3691
Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961
Lee, D. D., & Seung, H. S. 1999, Natur, 401, 788
Mahoney, M. W., & Drineas, P. 2009, PNAS, 106, 697
Mahoney, M. W., Orecchia, L., & Vishnoi, N. 2012, J. Mach. Learn. Res., 13, 2339
Richards, J., Freeman, P., Lee, A., & Schafer, C. 2009, ApJ, 691, 32
Roweis, S., & Saul, L. 2000, Sci, 290, 2323
Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
Tenenbaum, J. B., De Silva, V., & Langford, J. C. 2000, Sci, 290, 2319
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
VanderPlas, J., & Connolly, A. 2009, AJ, 138, 1365
Wild, V., & Hewett, P. C. 2005, MNRAS, 358, 1083
Yip, C.-W., Connolly, A., Szalay, A., et al. 2004a, AJ, 128, 585
Yip, C. W., Connolly, A. J., Vanden Berk, D. E., et al. 2004b, AJ, 128, 2603
Yip, C.-W., Mahoney, M. W., Szalay, A., et al. 2014, AJ, 147, 110
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579