Exploiting Low-Dimensional Structure in Astronomical Spectra

Joseph W. Richards, Peter E. Freeman, Ann B. Lee, Chad M. Schafer

jwrichar@stat.cmu.edu

Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

ABSTRACT

Dimension-reduction techniques can greatly improve statistical inference in astronomy. A standard approach is to use Principal Components Analysis (PCA). In this letter we apply a recently-developed technique, diffusion maps, to astronomical spectra, and develop a robust, eigenmode-based framework for regression and data parameterization. We show how our framework provides a computationally efficient means by which to predict redshifts of galaxies, and thus could inform more expensive redshift estimators such as template cross-correlation. It also provides a natural means by which to identify outliers (e.g., misclassified spectra). We analyze 3846 SDSS spectra and show how our framework yields an approximately 99% percent reduction in dimensionality. Finally, we show that the prediction error of the diffusion map-based regression approach is markedly smaller than that of a similar approach based on PCA, clearly demonstrating the superiority of diffusion maps over PCA and traditional linear data reduction techniques.

Subject headings: galaxies: distances and redshifts — galaxies: fundamental parameters — galaxies: statistics — methods: statistical — methods: data analysis

1. Introduction

Galaxy spectra are classic examples of high-dimensional data, with thousands of measured fluxes providing information about the physical conditions of the observed object. To make computationally efficient inferences about these conditions, we need to first reduce the dimensionality of the data space while preserving relevant physical information. We then need to find simple relationships between the reduced data and physical parameters of interest. Principal Components Analysis (PCA, or the Karhunen-Loève transform) is a standard method for the first step; its application to galaxy spectra is described in, e.g., Boroson & Green (1992), Connolly et al. (1995), Madgwick et al. (2003), Yip et al. (2004a), Yip et al. (2004b), Li et al. (2005), Jian-Nan et al. (2006), Vanden Berk et al. (2006), Rogers et al. (2007), and Re Fiorentin et al. (2007). In most cases, the authors do not proceed to the second step but only ascribe physical significance to the first few eigenfunctions from PCA (such as the "Eigenvector 1" of Boroson & Green). Notable

exceptions are Li et al., Jian-Nan et al., and Re Fiorentin et al. However, these authors combine eigenfunctions in an ad hoc manner with no formal methods or statistical criteria for regression and risk (i.e., error) estimation.

In this letter we present a unified framework for regression and data parameterization of astronomical spectra. The main idea is to describe the important structure of a data set in terms of its *fundamental eigenmodes*. The corresponding eigenfunctions are used both as coordinates for the data and as orthogonal basis functions for regression. We also introduce the *diffusion map* framework (see, e.g., Coifman & Lafon 2006, Lafon & Lee 2006) to astronomy, comparing and contrasting the method with PCA for regression analysis of SDSS galaxy spectra. PCA is a global method that finds linear low-dimensional projections of the data; the method attempts to preserve Euclidean distances between all data points and is often not robust to outliers. The diffusion map approach, on the other hand, is non-linear and instead retains distances that reflect the (local) connectivity of the data. This method is robust to outliers and is often able to unravel the intrinsic geometry and the natural (non-linear) coordinates of the data.

In §2 we introduce diffusion maps. In §3 we apply both PCA and diffusion maps to the problem of *adaptive regression* using eigenmodes. In §4 we demonstrate the effectiveness of our proposed PCA- and diffusion-map-based regression techniques for predicting the redshifts given by SDSS. Our PCA- and diffusion-map-based approaches provide a fast and statistically rigorous means of identifying outliers in redshift data. The returned embeddings also provide an informative visualization of the results. In §5 we summarize our results.

2. Diffusion Maps

We first use diffusion maps for data parameterization, i.e., to find a natural coordinate system for the data. When reducing the dimensionality of the data, one needs to decide what features to preserve and what aspects of the data one is willing to lose. The diffusion map framework attempts to retain the cumulative local interactions between its data points, or their "connectivity" in the context of a diffusion process. We demonstrate how this can be a better method to learn the intrinsic geometry of a data set than by using, e.g., PCA, which simply projects all data points onto a lower-dimensional hyperplane.

Our goal is to define a distance metric $D(\mathbf{x}, \mathbf{y})$ that reflects the connectivity of two points $\mathbf{x}$ and $\mathbf{y}$. (Note that in our case a "point" in p-dimensional space represents a complete spectrum of p wavelength bins.) The general idea is that we call two data points "close" if there are many short paths between $\mathbf{x}$ and $\mathbf{y}$ in a jump diffusion process. Our starting point is defining $w(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{s(\mathbf{x},\mathbf{y})^2}{\epsilon}\right)$, where $s(\mathbf{x}, \mathbf{y})$ is a locally relevant similarity measure, e.g., the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ (denoted here $\|\mathbf{x} - \mathbf{y}\|$) when $\mathbf{x}$ and $\mathbf{y}$ are vectors. The tuning parameter ϵ is chosen small enough that $w(\mathbf{x}, \mathbf{y}) \approx 0$ unless $\mathbf{x}$ and $\mathbf{y}$ are similar, but large enough such that the data set is connected.

Using the weight matrix $\mathbf{W}$ with elements $w(\mathbf{x}, \mathbf{y})$, we then construct a Markov random walk on our data with a transition matrix $\mathbf{P}$ whose elements are $p_1(\mathbf{x}, \mathbf{y}) = \frac{w(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{z}} w(\mathbf{x},\mathbf{z})}$. We interpret $p_1(\mathbf{x}, \mathbf{y})$ as the probability of moving from $\mathbf{x}$ to $\mathbf{y}$ in one time step. Given a positive integer t, the matrix power $\mathbf{P}^t$, with elements $p_t(\mathbf{x}, \mathbf{y})$, therefore represents the probability of moving from $\mathbf{x}$ to $\mathbf{y}$ in t steps. Increasing t moves the random walk forward in time, propagating the local influence of a data point (as defined by the kernel w) with its neighbors.

For a fixed time or scale t, the points $\mathbf{x}$ and $\mathbf{y}$ are close if the conditional distributions after t steps in the random walk, given by the vectors $p_t(\mathbf{x}, \cdot)$ and $p_t(\mathbf{y}, \cdot)$, are similar. This leads to a natural definition of the diffusion distance at a scale t as

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{z}} \frac{(p_t(\mathbf{x}, \mathbf{z}) - p_t(\mathbf{y}, \mathbf{z}))^2}{\phi_0(\mathbf{z})} \tag{1}$$

where $\phi_0(\cdot)$ is the stationary distribution of the random walk. The distance will be small if $\mathbf{x}$ and $\mathbf{y}$ are connected by many short paths with large weights. This construction of a distance measure is robust to noise and outliers because it simultaneously accounts for *all* paths between the data points. An example of a situation where a diffusion distance is more appropriate than the Euclidean distance in the original space is shown in Figure 1.

In applying this technique for dimensionality reduction, the data set attribute we wish to preserve is the diffusion distance between all points. A biorthogonal spectral decomposition of the matrix $\mathbf{P}^t$ gives $p_t(\mathbf{x}, \mathbf{y}) = \sum_{j \geq 0} \lambda_j^t \psi_j(\mathbf{x}) \phi_j(\mathbf{y})$, where ϕ_j, ψ_j, and λ_j, respectively, represent left eigenvectors, right eigenvectors and eigenvalues of $\mathbf{P}$. By retaining the m eigenmodes corresponding to the m largest nontrivial eigenvalues and by introducing the diffusion map

$$\Psi_t : \mathbf{x} \mapsto [\lambda_1^t \psi_1(\mathbf{x}), \lambda_2^t \psi_2(\mathbf{x}), \cdots, \lambda_m^t \psi_m(\mathbf{x})] \tag{2}$$

from $\mathbb{R}^p$ to $\mathbb{R}^m$, we have that (see Coifman & Lafon (2006))

$$D_t^2(\mathbf{x}, \mathbf{y}) \simeq \sum_{j=1}^{m} \lambda_j^{2t} (\psi_j(\mathbf{x}) - \psi_j(\mathbf{y}))^2 = ||\Psi_t(\mathbf{x}) - \Psi_t(\mathbf{y})||^2, \tag{3}$$

i.e., Euclidean distance in the m-dimensional embedding approximates diffusion distance. In contrast, PC maps approximate the original Euclidean distances $\|\mathbf{x} - \mathbf{y}\|$. For the example in Figure 1, a diffusion map onto one dimension ($m = 1$) approximately recovers the arc length parameter of the spiral. A one-dimensional PC map, on the other hand, simply projects all the data onto a straight line through the origin.

3. Adaptive Regression Using Orthogonal Eigenfunctions

Assume that our data is defined on a set $\mathcal{X} \subset \mathbb{R}^p$, where p is very large but the intrinsic dimension of $\mathcal{X}$ is small. The set $\mathcal{X}$ could, for example, be a non-linear submanifold embedded in $\mathbb{R}^p$;

see Figure 1 for an example where $\mathcal{X}$ is a one-dimensional spiral. We may view the eigenfunctions from PCA or diffusion maps (a) as *coordinates* of the data points, as shown in the previous section, or (b) as forming a *Hilbert basis* for any function supported on the subset $\mathcal{X}$. Rather than applying an arbitrarily chosen prediction scheme in the computed diffusion or PC space (as in, e.g., Li et al., Jian-Nan et al., and Re Fiorentin et al.), we utilize the latter insight to formulate a general regression and risk estimation framework for high-dimensional inference.

We may write any function r satisfying $\int r(\mathbf{x})^2 dx < \infty$, where $\mathbf{x} \in \mathcal{X}$, as

$$r(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x}) \,, \tag{4}$$

where the sequence of functions $\{\psi_1, \psi_2, \cdots\}$ forms an orthonormal basis. The choice of basis functions is traditionally *not* adapted to the geometry of the data, or the set $\mathcal{X}$. Standard choices are, for example, Fourier or wavelet bases for $\mathbf{L}^2(\mathbb{R}^p)$, which are constructed as tensor products of one-dimensional bases.

In regression, we are given n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, with the task of predicting the response $Y = r(X) + \epsilon$ at a new data point $X = \mathbf{x}$, where ϵ represents random noise. (In §4, the response Y is the redshift, z, and X is a complete spectrum.) In *non-parametric regression by orthogonal functions*, one assumes that $r(\mathbf{x})$ is given according to equation (4). An estimator of $r(\mathbf{x})$ typically has the form

$$\widehat{r}(\mathbf{x}) = \sum_{j=1}^{J} \widehat{\beta}_j \psi_j(\mathbf{x}) \ , \tag{5}$$

where $J \leq m$ and $\{\psi_j\}$ is a fixed basis. The primary goal is to minimize the prediction risk (i.e., expected error), commonly quantified by the mean-squared error (MSE), $R(J) = \mathbb{E}(Y - \widehat{r}(X))^2$, where the expectation averages everything that is random, including the randomness in the evaluation points X. (The risk and an appropriate value for J is then estimated from the data by, for example, cross-validation.) A secondary goal is "sparsity"; more specifically, among the estimators with a small risk, we choose representations with a smaller J.

We propose a new *adaptive* framework where the basis functions reflect the intrinsic (low-dimensional) geometry of the data. Rather than using a generic tensor-product basis for the high-dimensional space $\mathbb{R}^p$, we construct a data-driven basis for the lower-dimensional embedding $\mathcal{X}$ where the data lie. Let $\{\psi_1, \psi_2, \cdots, \psi_m\}$ be the orthogonal eigenfunctions computed by PCA or diffusion maps. Our regression function estimate $\widehat{r}(\mathbf{x})$ is then given by equation (5), where the different terms in the series expansion represent the fundamental eigenmodes of the data. Our claim is that this method will lead to efficient inference in high dimensions, as we are effectively performing regression in a lower-dimensional space $\mathcal{X}$. Furthermore, the use of eigenmodes in both the data parameterization and in the regression formulation provides an elegant, unifying framework for analysis and prediction.

4. Redshift Prediction

We apply the formalism presented in §§2-3 to the problem of redshift prediction in SDSS spectra. Physically similar objects residing at similar redshifts will have similar continuum shapes as well as absorption lines that occur at similar wavelengths, and thus the Euclidean distances between their spectra will be small. Thus adaptive regression provides a natural means by which to predict redshifts. Furthermore, it is computationally efficient, making its use appropriate for large databases such as the SDSS; one can use these predictions to inform more computationally expensive techniques by narrowing down the relevant parameter space (e.g., the redshift range or the set of templates in cross-correlation techniques). Adaptive regression also provides a useful tool for, e.g., quickly identifying anomalous data points (e.g., objects misclassified as galaxies), galaxies that have relatively rare features of interest, and galaxies whose SDSS redshift estimates may be incorrect.

We perform PCA and diffusion mapping for a sample of 3846 SDSS galaxy spectra (data from 10 arbitrarily chosen plates of SDSS DR6; Adelman-McCarthy et al. 2008). In our analysis, we (a) ignore the first 100 and last 250 pixels of each spectrum; (b) do not consider spectra with more than 10% of its remaining pixels flagged as "bad" pixels; and (c) replace data in the vicinity of prominent atmospheric lines at 5577Å, 6300Å, and 6363Å with the sample mean of the nine closest pixels on either side of each line. Aperture considerations lead us to analyze only data with SDSS redshift estimates $z_{\rm SDSS} \geq 0.05$, and we mask out emission lines because their highly variable strengths strongly bias distance calculations.

Results show that both PCA and diffusion maps perform well in recovering redshift. In Figure 2 we plot the embedding of the 2796 galaxies in our sample with SDSS confidence level[1] (CL) > 0.99 in the first three PC and diffusion map coordinates. In both maps we find that the structure of this reparameterization of the original data corresponds in a simple way to $\log_{10}(1 + z_{\rm SDSS})$.

Regression of $z_{\rm SDSS}$ on the PC and diffusion map eigenmodes reveals the real advantage of the diffusion map method in this problem. In our analysis, to eliminate the effects that poorly estimated SDSS redshifts may have on our results, we only consider galaxies with SDSS z CL > 0.99. In Figure 4 it is shown that for any number of eigenmodes, we generally achieve a lower cross-validation prediction risk ($\hat{R}_{CV}$, an unbiased estimate of R; see, e.g., Wasserman 2006) from regressing on diffusion map basis functions than from regressing on PC basis functions. The low-dimensional diffusion map representation of our data captures the trend in z better than the PC representation.

Finally, we use the regression model trained on the 2796 z CL > 0.99 galaxy spectra to predict redshifts for the other 1050 spectra. The optimal regression model, i.e., the model that minimizes cross-validation prediction risk, is the diffusion map model using $J = 43$ eigenfunctions. Note that

[1]SDSS confidence levels are functions of the strengths of observed lines and thus should not be interpreted probabilistically.

since our original data were in 3500 dimensions, our optimal model has achieved a 98.8% reduction in dimensionality. Table 1 shows parameters for the optimal diffusion map and PC regression models.

In Figure 5 we plot our predictions for the z CL ≤ 0.99 galaxies against SDSS z estimates using the optimal diffusion map model. In that same figure, we plot predictions for z CL > 0.99 from 10-fold cross-validation (i.e., in each fold, the model is trained on 90% of the data and predictions made for the other 10%). Most of our predictions are in close correspondence with the SDSS estimates. There are 56 outliers at the 4σ level. Manual inspection of these spectra indicate that roughly half are problematic (e.g., clearly misclassified spectra and spectra exhibiting anomalous features, etc.; see Figure 3). The remainder appear to closely match SDSS templates 29 (luminous red galaxy) and 25 (galaxy); since these templates match the vast majority of SDSS galaxies, this may simply indicate that the widths of our prediction intervals in Figure 5, which are constructed based on assumptions of normality, are underestimated.

5. Summary

We present a fast and powerful eigenmode-based framework for estimating physical parameters in databases of high-dimensional astronomical data. In most applications, Principal Component Analysis (PCA) is used as a data-explorative tool for dimensionality reduction, with no formal methods and statistical criteria for regression, risk estimation and selection of relevant eigenvectors. Here we propose a statistically rigorous, unified framework for regression and data parameterization. We apply the methodology to predict redshift for a sample of SDSS galaxy spectra, and also compare the use of the proposed method with PCA versus a non-linear eigenmap technique called "diffusion maps." We find that the prediction error for the diffusion-map-based approach is markedly smaller than that of a similar framework based on PCA. Our techniques are also more robust than commonly used template matching methods because they learn the structure of the entire high-dimensional data set. Statistical inferences are based on this structure, instead of considering each data point separately in an object-by-object matching algorithm. Work in progress extends this approach to photometric redshift estimation and to the estimation of the intrinsic parameters (e.g., mean metallicities and ages) of galaxies.

The authors would like to thank Jeff Newman for helpful conversations. This work was supported by NSF grant #0707059.

REFERENCES

Adelman-McCarthy, J. K., et al. 2008, ApJS, 175, 297

Boroson, T. A., & Green, R. F. 1992, ApJS, 80, 109

Coifman, R. R., & Lafon, S. 2006, Appl. Comput. Harmon. Anal., 21, 5

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, AJ, 110, 1071

Jian-Nan, Z., Fu-Chao, W., Li, A-Li, L., & Yong-Heng, Z. 2006, ChJAA, 30, 176

Lafon, S., & Lee, A. 2006, IEEE Trans. Pattern Anal. and Mach. Intel., 28, 1393

Li, C., Wang, T.-G., Zhou, H.-Y., Dong, X.-B., & Cheng, F.-Z. 2005, AJ, 129, 669

Madgwick, D. S., et al. 2003, ApJ, 599, 997

Re Fiorentin, P., et al. 2007, A&A, 467, 1373

Rogers, B., Ferreras, I., Lahav, O., Bernardi, M., Kaviraj, S., & Yi, S. K. 2007, MNRAS, 382, 750

Vanden Berk, D. E., et al. 2006, AJ, 131, 84

Wasserman, L. W. 2006, All of Nonparametric Statistics (New York:Springer)

Yip, C. W., et al. 2004, AJ, 128, 585

Yip, C. W., et al. 2004, AJ, 128, 2603

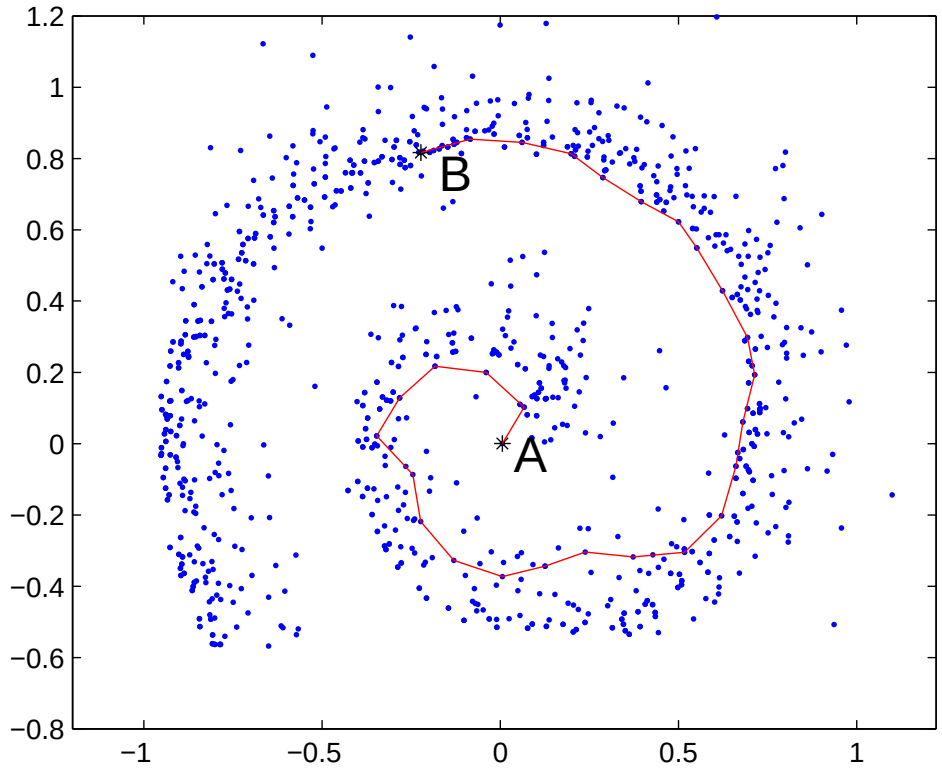

Fig. 1.— An example of a one-dimensional manifold embedded in two dimensions. The path from A to B is representative of the diffusion distance between A and B, and is a better representation of dissimilarity between them than the Euclidean distance.

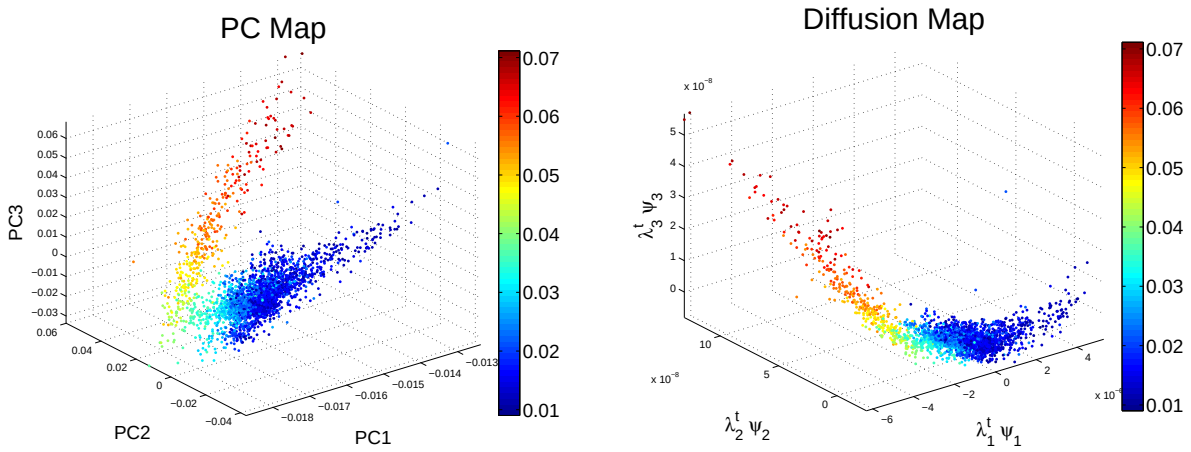

Fig. 2.— Embedding of our sample of 2796 SDSS galaxy spectra with SDSS z CL > 0.99 with the first 3 PC and diffusion map coordinates, respectively. The color codes for $\log_{10}(1 + z_{\rm SDSS})$ values. Both maps show a clear correspondence with redshift.

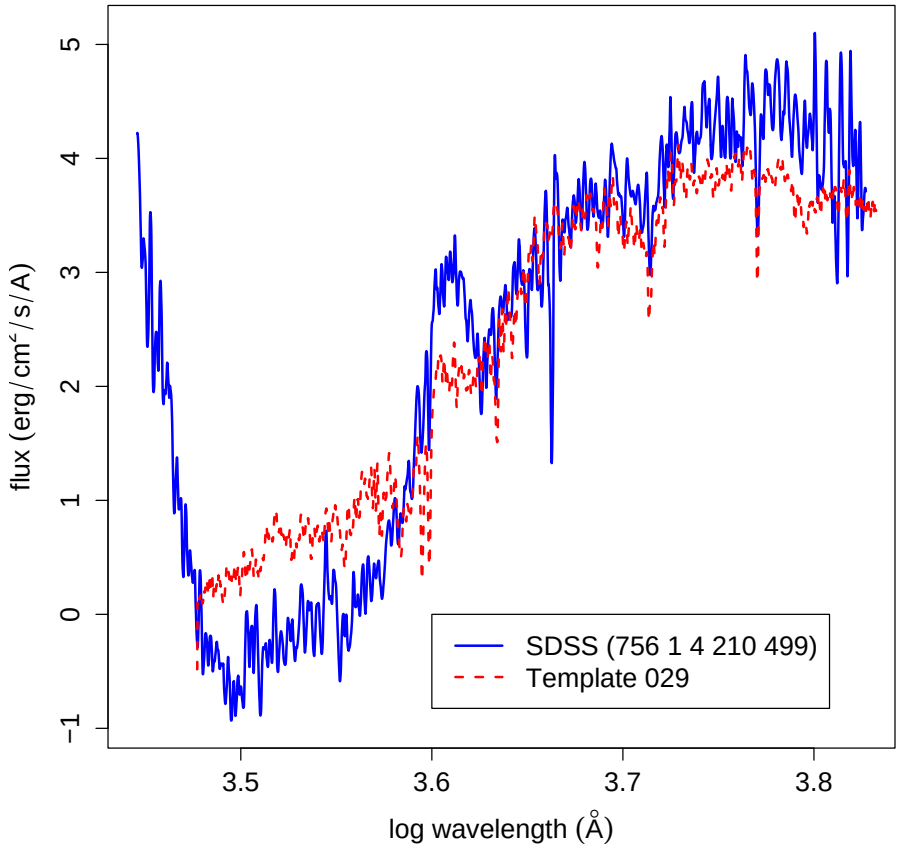

Fig. 3.— SDSS galaxy spectrum (with `OBJID`) identified as an outlier ($> 4\sigma$) by the diffusion map-based regression, overlaid with SDSS template 29, which provided the highest CL z_{SDSS} estimate in template cross-correlation. The spectrum exhibits two anomalous features: a sharp, unexplained rise at low wavelengths and a broad emission feature at ≈ 4100 Å.

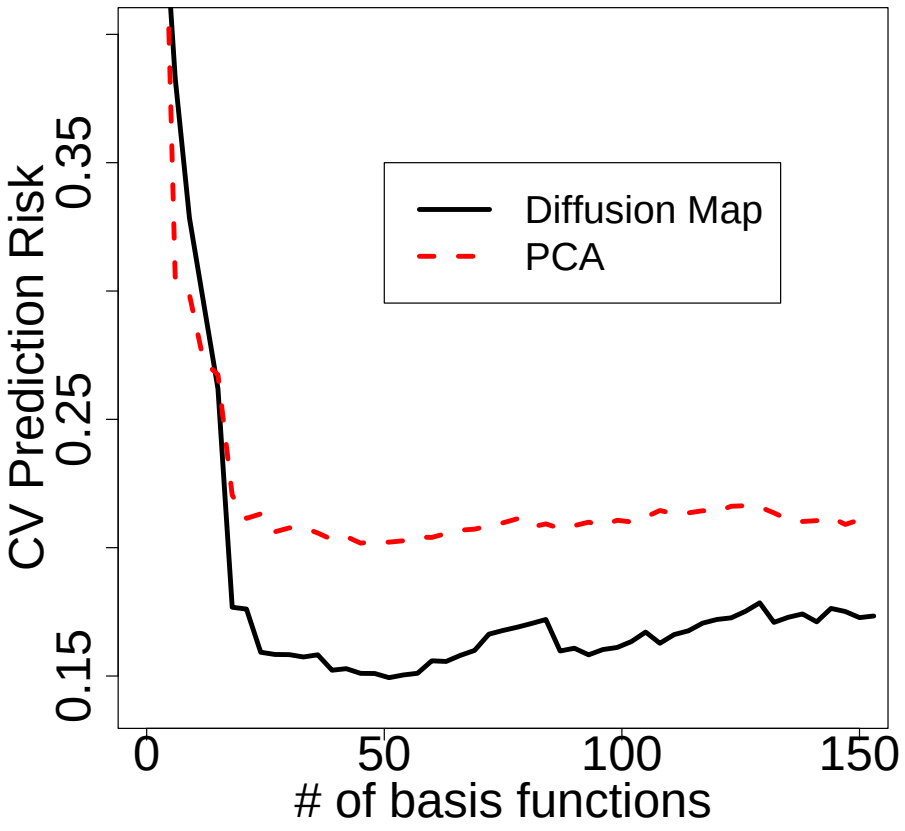

Fig. 4.— Risk estimates ($\hat{R}_{CV}$) for regression of z on diffusion map coordinates and PCs. Diffusion map is "sparser" and attains a lower risk for almost every number of coordinates in the regression. It also achieves a lower minimum risk as indicated by Table 1.

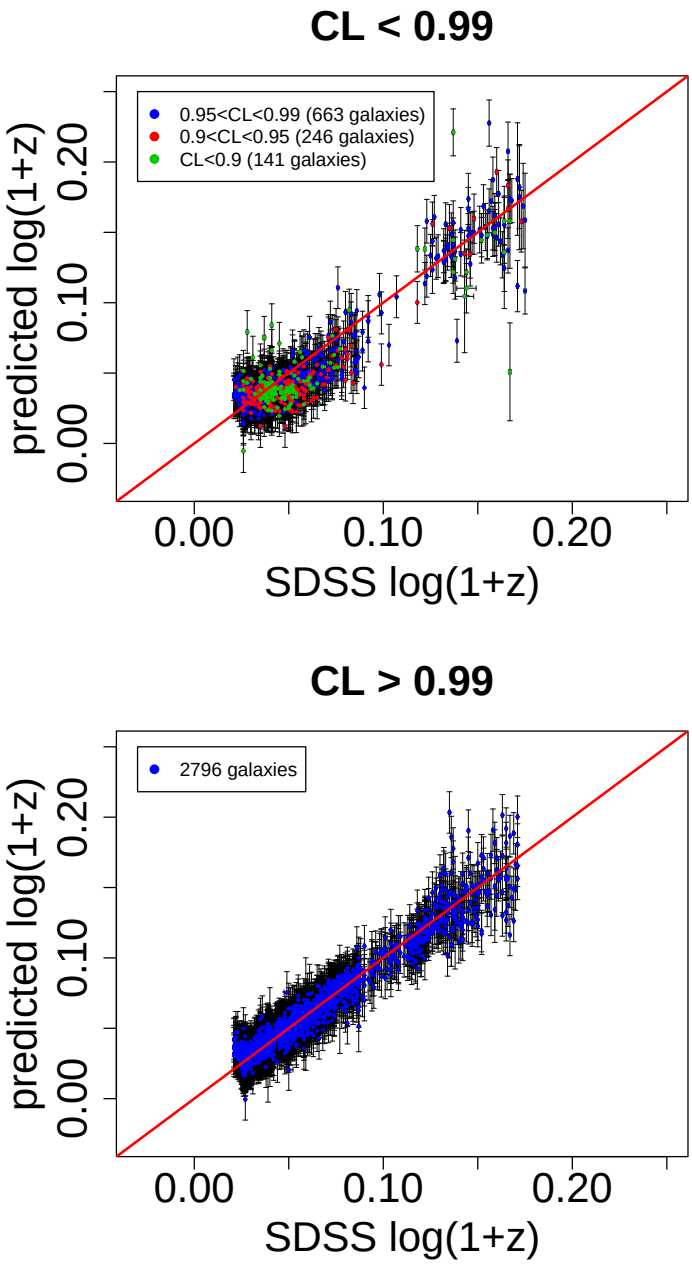

Fig. 5.— Redshift predictions using diffusion map coordinates for galaxies with SDSS z CL ≤ 0.99 (top) and z CL > 0.99 (bottom), each plotted against z_{SDSS}. Error bars represent 95% prediction intervals. For most galaxies, our predictions are in close correspondence with SDSS estimates.

Table 1. Parameters of Optimal Regression on $\log_{10}(1 + z_{\rm SDSS})$

	ϵ_{opt}	J_{opt}	$\hat{R}_{CV}(\epsilon_{opt}, J_{opt})$[a]	Number of Outliers 3σ	4σ	5σ
Diffusion Map	.0008	43	0.1488	141	56	22
PC	–	53	0.2024	147	58	29

[a]Prediction risk estimated via cross-validation; see equation (5) and subsequent discussion.