

# Sparse Representation and Efficient Inference for SDSS Spectra

Joseph Richards

jwrichar@stat.cmu.edu

Department of Statistics  
Carnegie Mellon University

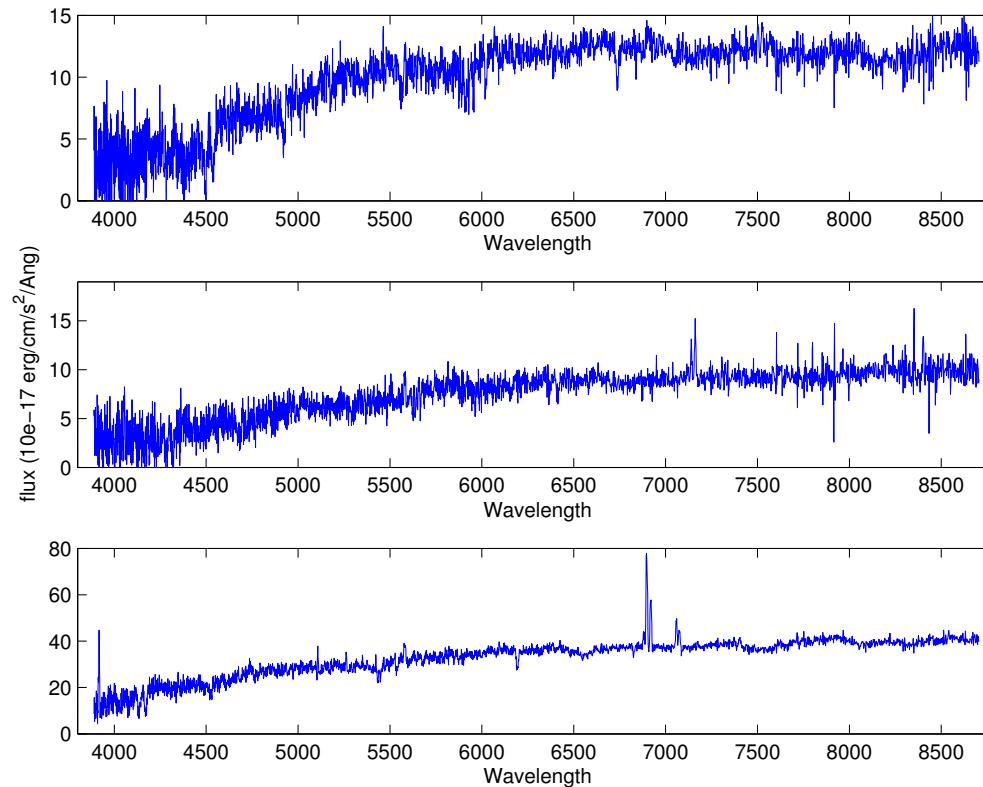
INternational Computational Astrostatistics  
[www.incagroup.org](http://www.incagroup.org)

# Motivation

- Astronomical spectra are data in very high dimensional space

# Motivation

- Astronomical spectra are data in very high dimensional space
- Sloan Digital Sky Survey (SDSS) spectra: 3850 wavelength bins per spectrum



# Motivation

- Astronomical spectra are data in very high dimensional space
- Our Goal: Estimate physically-important parameters for a large database of galaxies through their spectra

# Outline

- Methods
  - Principal Components Analysis
  - Diffusion Map
  - Adaptive Regression
- Applications
  - Redshift prediction
  - Age and Metallicity estimation
- Future Work

# Intro: Dimensionality Reduction

- In high dimensions, determining relationships between data points is difficult (curse of dimensionality)
- Explore whether there is a simple, lower-dimensional representation of the data.
- Make statistical inferences with this lower-dimensional parameterization of the data.
- In reducing dimensionality,
  - What properties of the original data set do we wish to retain?

# Methods: Principal Components Analysis (PCA)

- For data  $X$  ( $n \times p$ ), compute  $\hat{\Sigma} = \text{Cov}(X)$ . Perform spectral decomposition of  $\hat{\Sigma}$ :

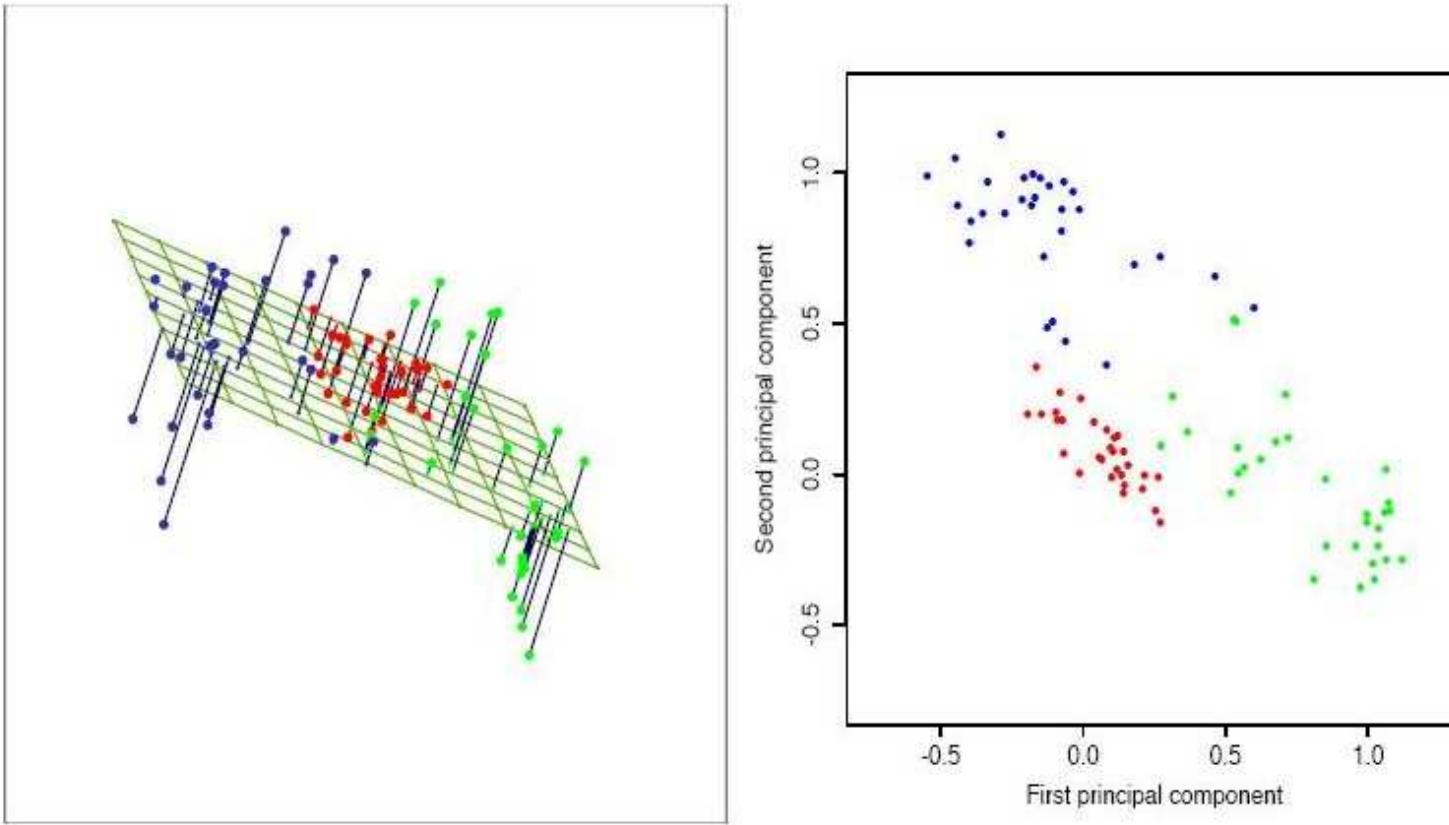
$$U^T \hat{\Sigma} U = \Lambda$$

$\Lambda$  is the diagonal matrix of eigenvalues of  $\hat{\Sigma}$

$U$  is a matrix of the corresponding eigenvectors

- Project the data  $X$  onto the eigenvectors corresponding to the largest  $m \ll p$  eigenvalues
- PCA projects original data to the  $m$ -dimensional hyperplane that maximizes variance of the new data
- low-dimensional PCA representation **optimally captures all Euclidean distances between original data points**

# PCA Example



*Elements of Statistical Learning,*  
Hastie, Tibshirani, and Friedman, pg. 488

# Drawbacks to PCA

- PCA is a **global** technique:
  - 1) fails to capture local heterogeneous features within a datapoint
  - 2) preserves all (Euclidean) distances between datapoints
- Is **linear**; projects data onto a hyperplane
- It is not robust to noise or outliers

# Drawbacks to PCA

- PCA is a **global** technique:
  - 1) fails to capture local heterogeneous features within a datapoint
  - 2) preserves all (Euclidean) distances between datapoints
- Is **linear**; projects data onto a hyperplane
- It is not robust to noise or outliers
- **Better strategy:** capture the intrinsic geometry of the data set (i.e. the connectivity of the data)

# Methods: Diffusion Maps

- Goal: learn the **intrinsic** (lower-dimensional) geometry of a dataset in  $\mathbb{R}^p$
- In high dimensions, distance measures are usually only relevant locally.

# Methods: Diffusion Maps

- Goal: learn the **intrinsic (lower-dimensional) geometry** of a dataset in  $\mathbb{R}^p$
- In high dimensions, distance measures are usually only relevant locally.
- Diffusion Map **Idea**:
  - Take a distance measure that makes sense locally, e.g.  $s(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$
  - Use  $s$  to define a distance metric,  $D(\mathbf{x}, \mathbf{y})$ , that reflects the connectivity of data points  $\mathbf{x}$  and  $\mathbf{y}$ .
  - **Preserve this distance when reducing dimensionality.**

# Diffusion Maps: Formulation

- Start with a local weight matrix,

$$w(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{s^2(\mathbf{x}, \mathbf{y})}{\epsilon}\right)$$

where  $\epsilon$  is a tuning parameter.

- The **diffusion kernel** is

$$p_1(\mathbf{x}, \mathbf{y}) = \frac{w(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{z}} w(\mathbf{x}, \mathbf{z})}$$

- $p_1(\mathbf{x}, \mathbf{y})$  can be interpreted as the transition probability of a Markov random walk on the data.

# Diffusion Maps: Formulation

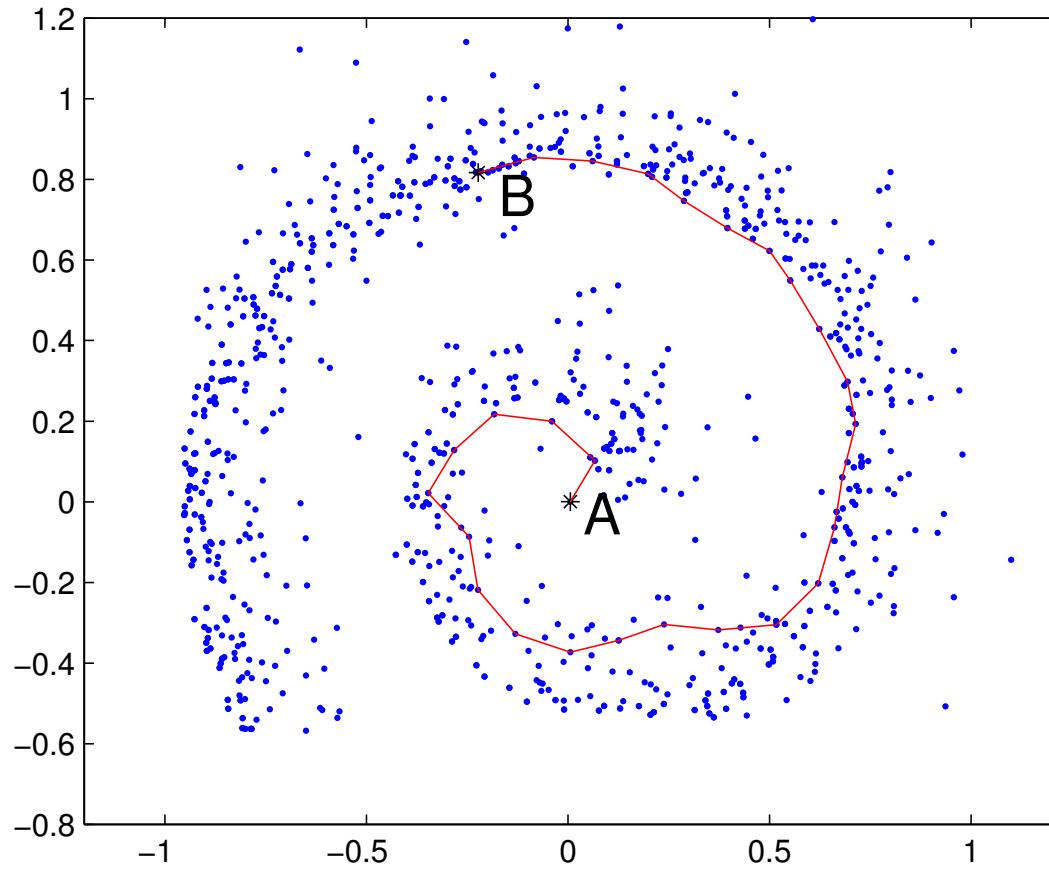
- Running this Markov chain forward in time,  $t$ , is equivalent to propagating the local influence of each data point with its neighbors
- Define the **diffusion distance** (Coifman and Lafon [2006])  $D_t$  for time  $t > 0$  as

$$D_t^2(\mathbf{x}, \mathbf{y}) = \|p_t(\mathbf{x}, \cdot) - p_t(\mathbf{y}, \cdot)\|_2^2 = \sum_{\mathbf{z} \in \Omega} (p_t(\mathbf{x}, \mathbf{z}) - p_t(\mathbf{y}, \mathbf{z}))^2$$

$D_t(\mathbf{x}, \mathbf{y})$  is closely related to  $p_t(\mathbf{x}, \mathbf{y})$

- The diffusion distance,  $D_t$ , between 2 points is related to their **connectivity**.

# Ex: Diffusion vs. Euclidean Distances



Diffusion distance more accurately describes the amount of dissimilarity between A and B (Lafon and Lee [2006]).

# Diffusion Maps: Dimension Reduction

- Perform spectral decomposition of  $P^t$ :

$$p_t(\mathbf{x}, \mathbf{y}) = \sum_{j \geq 0} \lambda_j^t \psi_j(\mathbf{x}) \phi_j(\mathbf{y})$$

where as in PCA, we retain the  $m$  eigenvectors corresponding to the  $m$  largest nontrivial eigenvalues

- The  $m$ -dimensional **diffusion map** of the data is

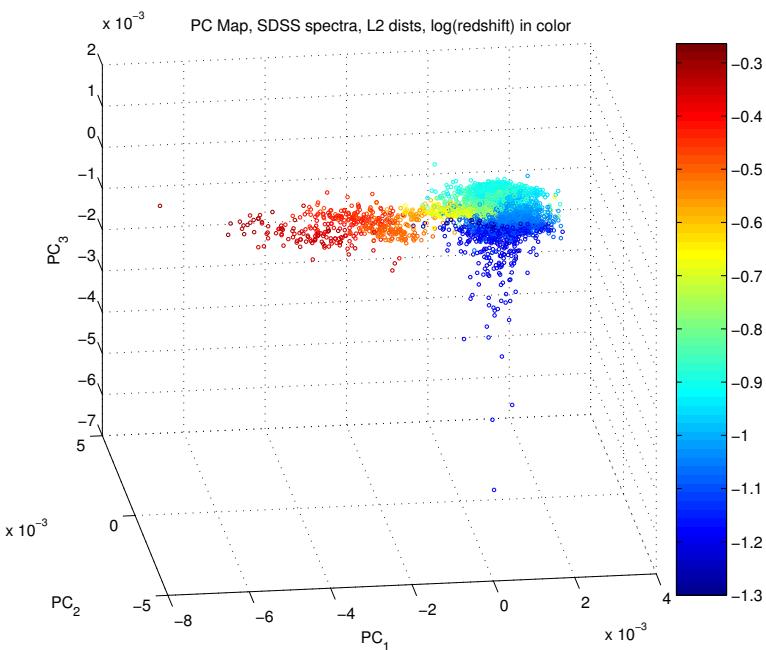
$$\Psi_t : \mathbf{x} \mapsto (\lambda_1^t \psi_1(\mathbf{x}), \lambda_2^t \psi_2(\mathbf{x}), \dots, \lambda_m^t \psi_m(\mathbf{x}))$$

where  $D_t^2(\mathbf{x}, \mathbf{y}) \simeq \|\Psi_t(\mathbf{x}) - \Psi_t(\mathbf{y})\|^2$

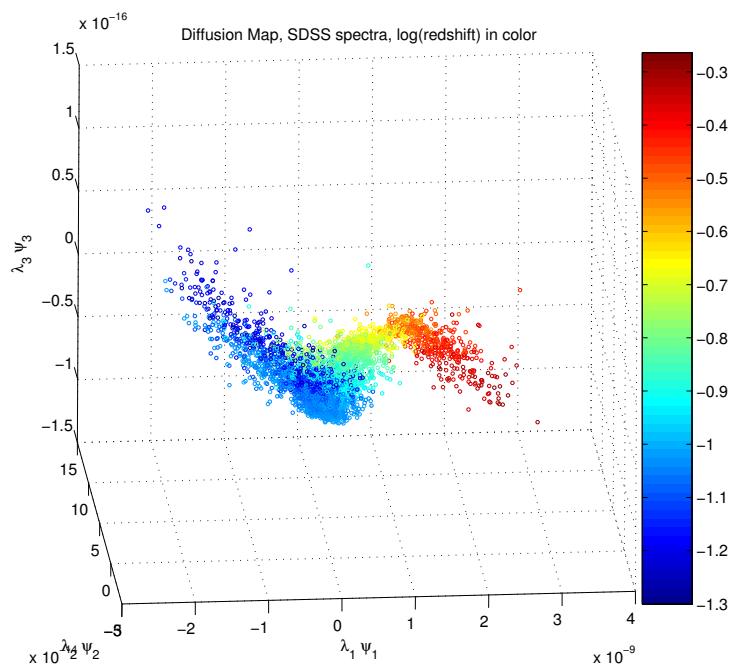
- Instead of capturing Euclidean distances (PCA), low-dim. representation captures diffusion distances

# Results: $z$ prediction

Sample of 3846 SDSS galaxy spectra  
Colored by  $\log(z_{SDSS})$



PC Map



Diffusion Map

# Methods: Adaptive Regression

- Any smooth regression function  $r$  can be written as

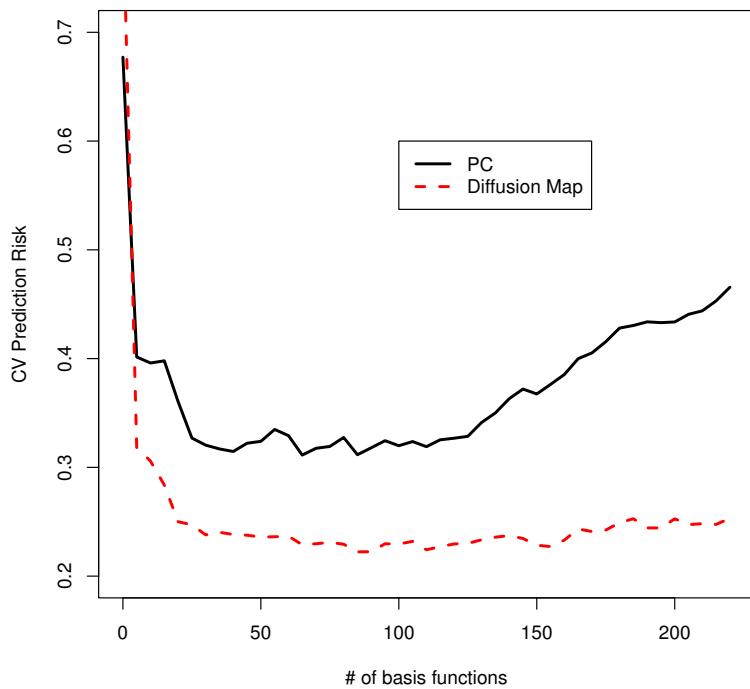
$$r(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x})$$

where  $\{\psi_1, \psi_2, \dots\}$  form an orthonormal basis,  $\mathbf{x} \in \mathbb{R}^p$

- Generally, choice of  $\psi_j$  is arbitrary (e.g. high-dimensional cosine basis, wavelets, curvelets, etc.)
- Regression function estimate:  $\hat{r}(\mathbf{x}) = \sum_{j=1}^J \hat{\beta}_j \psi_j(\mathbf{x})$
- Adaptive framework:** use the learned orthogonal function estimates  $\{\psi_1, \dots, \psi_m\}$  for  $\mathcal{X} \subset \mathbb{R}^p$  from PCA or diffusion maps.

# Results: $z$ prediction

CV Pred. Risk versus  $J$

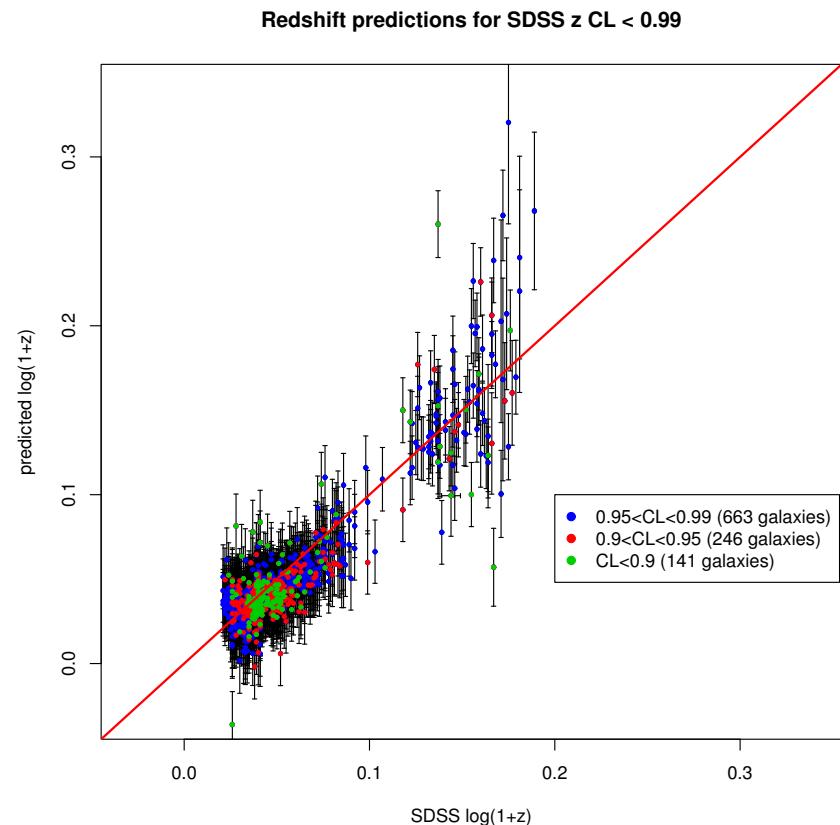
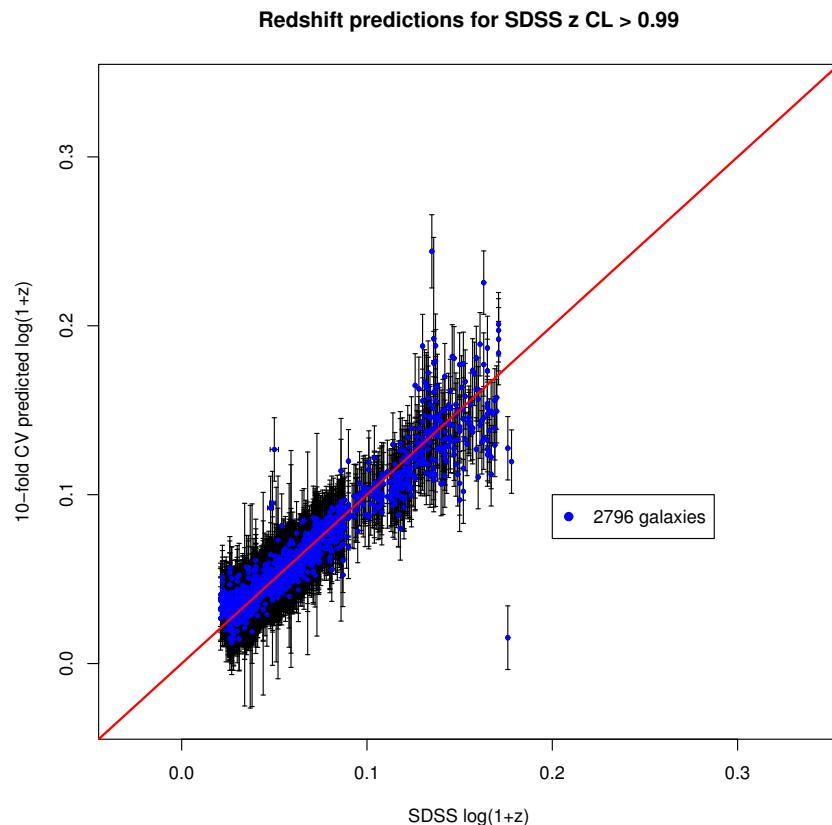


- Find PC and diffusion map coefficients for all 3846 spectra
- Regress only on spectra with  $z_{SDSS}$  CL > 0.99 (2796, 73%)
- Choose diffusion map  $\epsilon$  that minimizes CV prediction risk.

# Results: $z$ prediction

Predicted  $\log(1 + z)$  vs.  $\log(1 + z_{SDSS})$

Optimal model: 20 diffusion map coordinates



$z_{SDSS}$  CL > 0.99

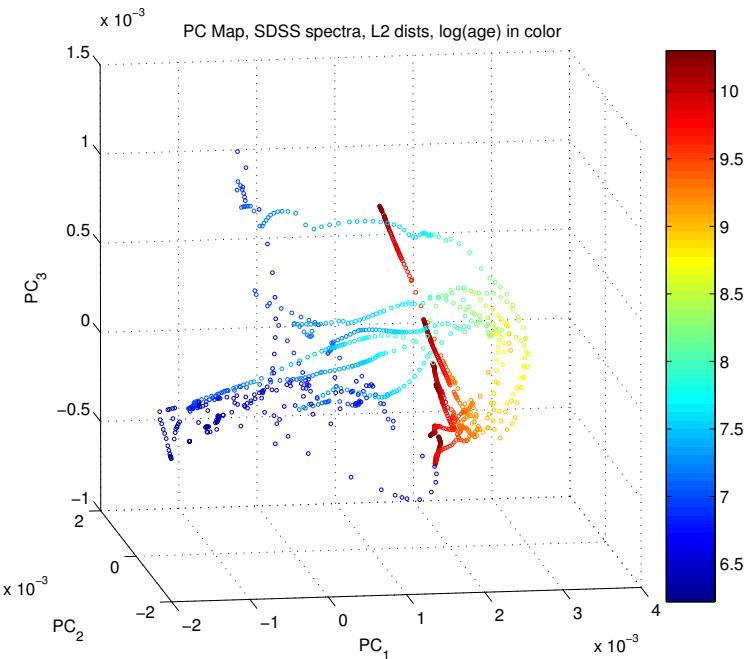
$z_{SDSS}$  CL  $\leq$  0.99

# Problem 2: Age & Metallicity

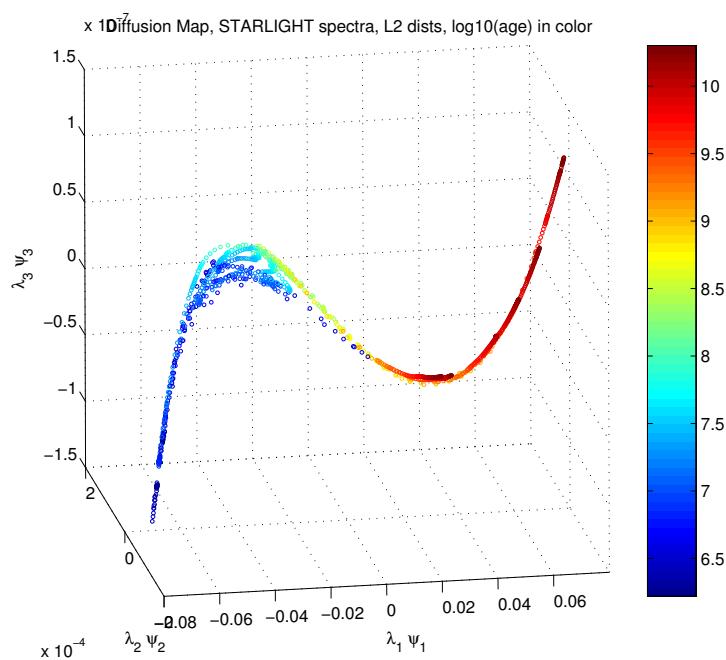
- Our main goal: Estimate age and metallicity of SDSS galaxies
- Training set: 1146 synthetic spectra from Bruzual and Charlot [2003] population synthesis models (Cid Fernandes et al. [2005], [www.starlight.ufsc.br](http://www.starlight.ufsc.br))
- For each synthetic spectrum, we know age &  $Z$ .

# Maps for synthetic spectra

Maps for simulated spectra, color is  $\log(\text{age})$



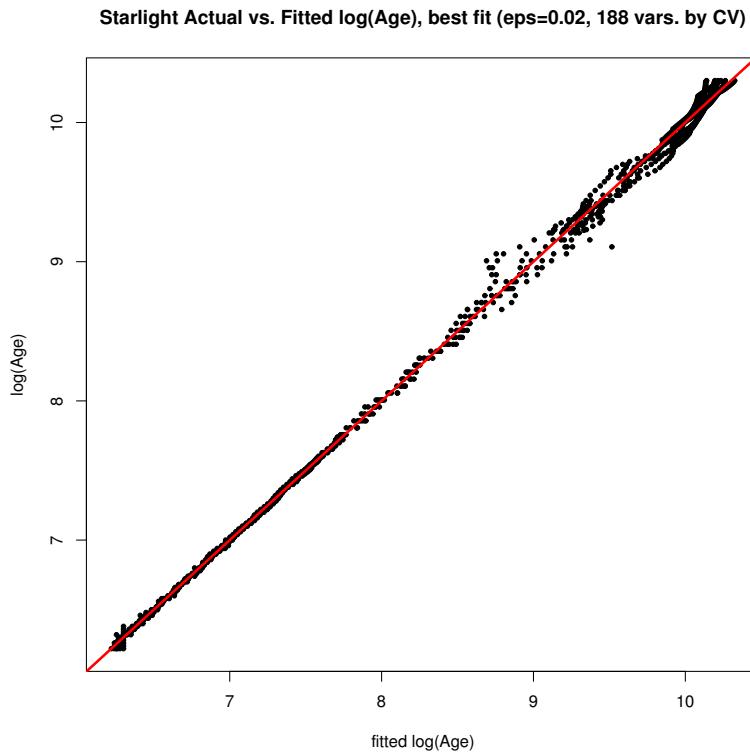
PC Map



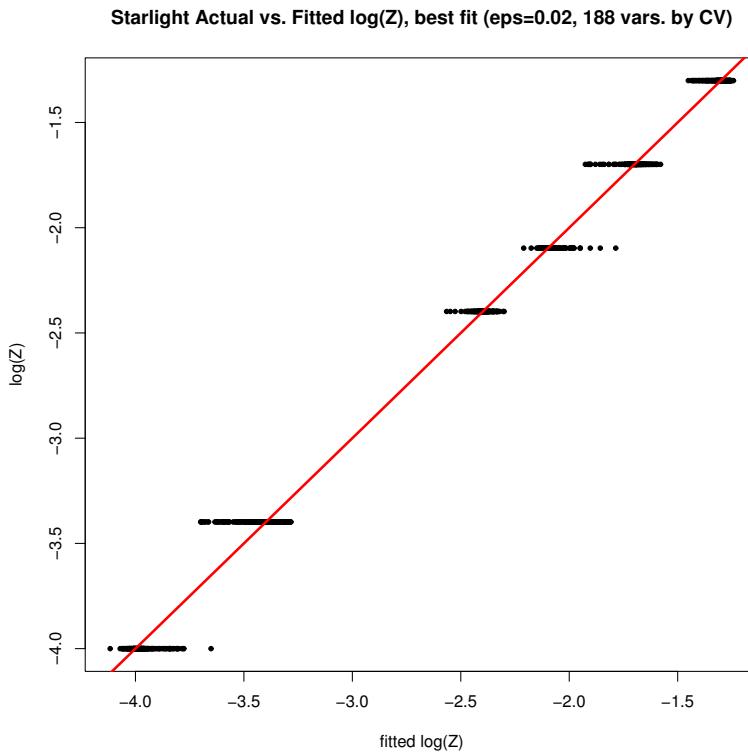
Diffusion Map

# Results: Age, Z prediction

Age



Metallicity



# Future Work

- Redshift Prediction
  - Apply redshift prediction techniques to more SDSS data
  - Predict redshift using continuum-subtracted spectra
- Age & Metallicity Estimation
  - Predict ages and metallicities for SDSS galaxies
  - Compare results to template-matching method of Cid Fernandes et al. [2005]

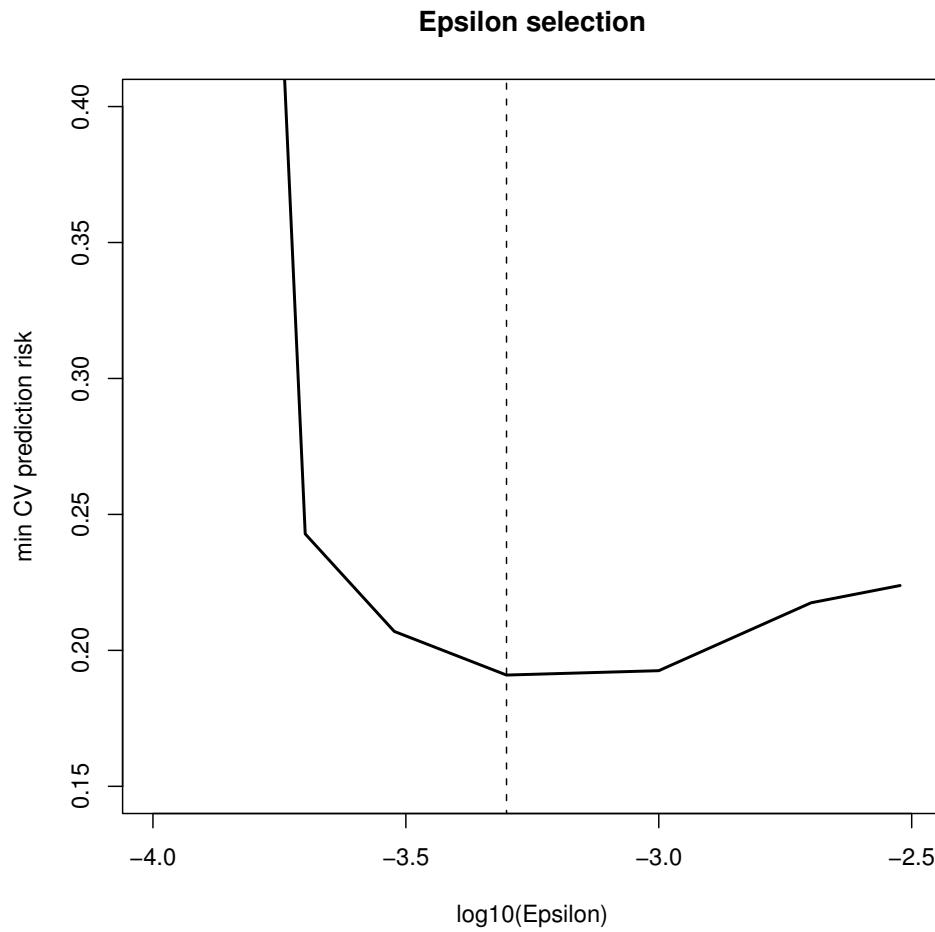
# Conclusions

1. Diffusion map is an **effective and efficient** dimensionality reduction technique
2. Outperforms linear methods (PCA) in regression tasks on real astronomical data sets
3. **Adaptive regression** used to perform statistical inference and choose optimal dimensionality
4. Methods **applicable for wide variety of astronomical data sets** (e.g. spectra, images, multi-wavelength data sets, etc., etc.) for variety of tasks (e.g. regression, classification, clustering, etc.)

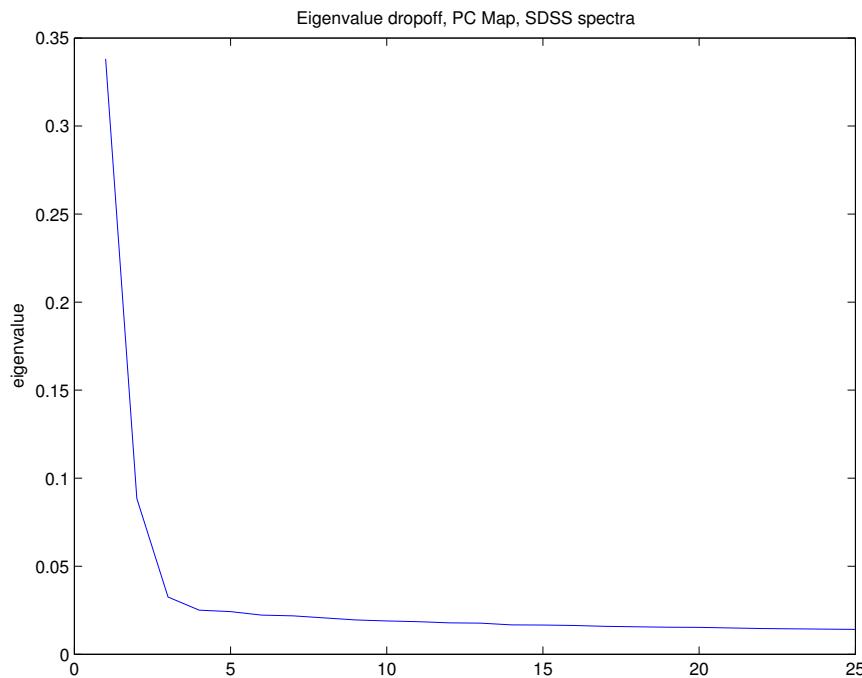
# References

- G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. *MNRAS*, 344:1000–1028, Oct. 2003.
- R. Cid Fernandes, A. Mateus, L. Sodré, G. Stasińska, and J. M. Gomes. Semi-empirical analysis of Sloan Digital Sky Survey galaxies - I. Spectral synthesis method. *MNRAS*, 358:363–378, Apr. 2005.
- R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006. URL  
<http://dx.doi.org/10.1016/j.acha.2006.04.006>.
- S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9), September 2006.

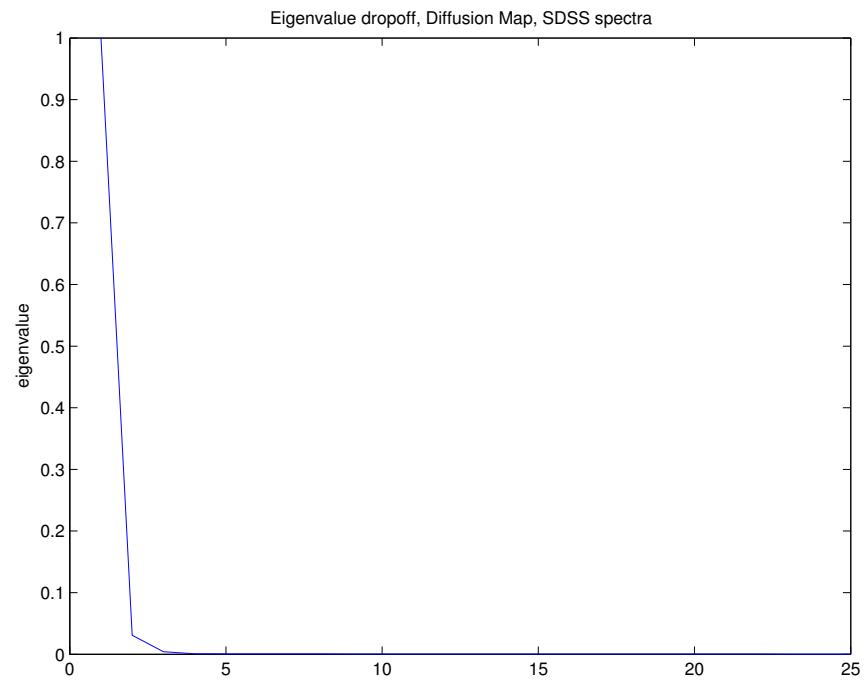
# Appendix: Epsilon Selection



# Appendix: Eigenvalue Dropoff



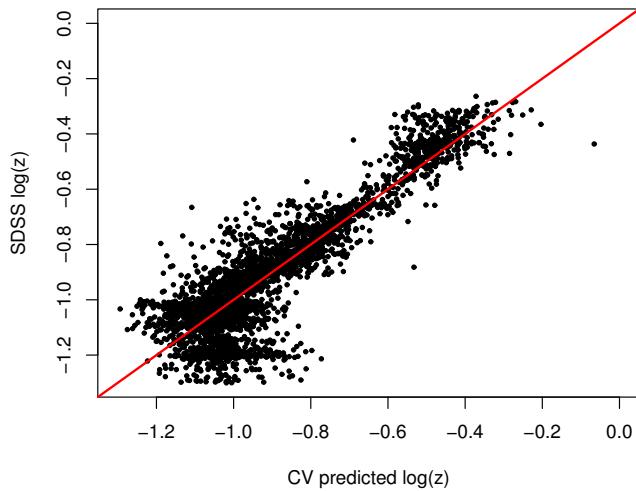
PCA



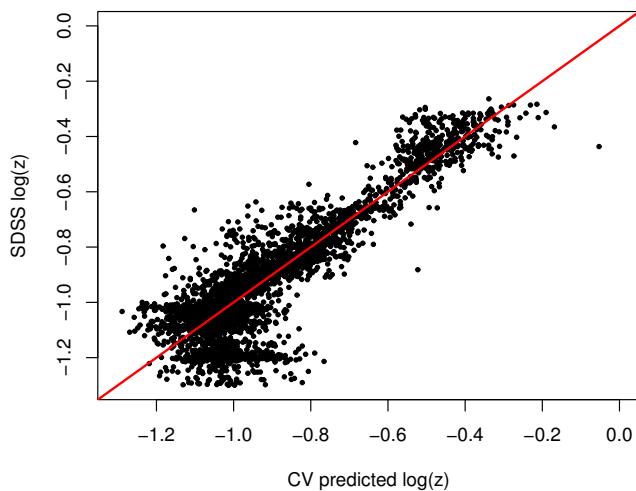
Diffusion Mapping

# Appendix: Kernel Regression

a) Linear regression 3 diffusion coords. (CV risk=36.6)



b) Kernel regression 3 diffusion coords. (CV risk=36.2)



# Appendix: SDSS-SL calibration

The following steps are performed to compare synthetic and SDSS spectra:

1. SDSS spectra are brought to rest-frame wavelengths (based on SDSS estimate of  $z$ )
2. Emission lines in SDSS spectra are masked out
3. SL spectra are rebinned to match SDSS binning
4. SL and SDSS spectra are normalized to sum to 1 in overlapping wavelength range
5.  $L_2$  distance computed between all SL and SDSS spectra
  - a. SDSS flux errors accounted for, or
  - b. SDSS spectra smoothed first