Exploiting Sparse Non-Linear Structure in Astronomical Data

Ann B. Lee
Department of Statistics and Department of Machine Learning, Carnegie Mellon University

Joint work with P. Freeman, C. Schafer, and J. Richards

High-Dimensional Inference for Large Data Sets

- Massive amounts of data collected in astronomical surveys.
- Typically, High dimensions (p), Huge data sets (n). Complex and noisy data.
- THE QUESTION IS: how do we extract useful information and make reliable predictions and estimates?
  - Visualization, outlier detection, density estimation
  - Regression/classification
- "Curse of dimensionality" (computational/statistical)

High Dimensional Inference: Two Important Considerations

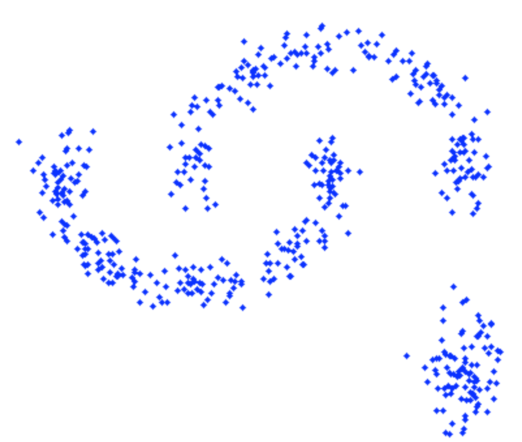
1. To find a good initial basis or set of attributes.
   - E.g. Construct ON basis. Dimensionality reduction by projection

$$\boxed{T_k : R^p \to R^k, \quad x \to T_k x = (x \cdot w_1, x \cdot w_2, \ldots, x \cdot w_k)}$$

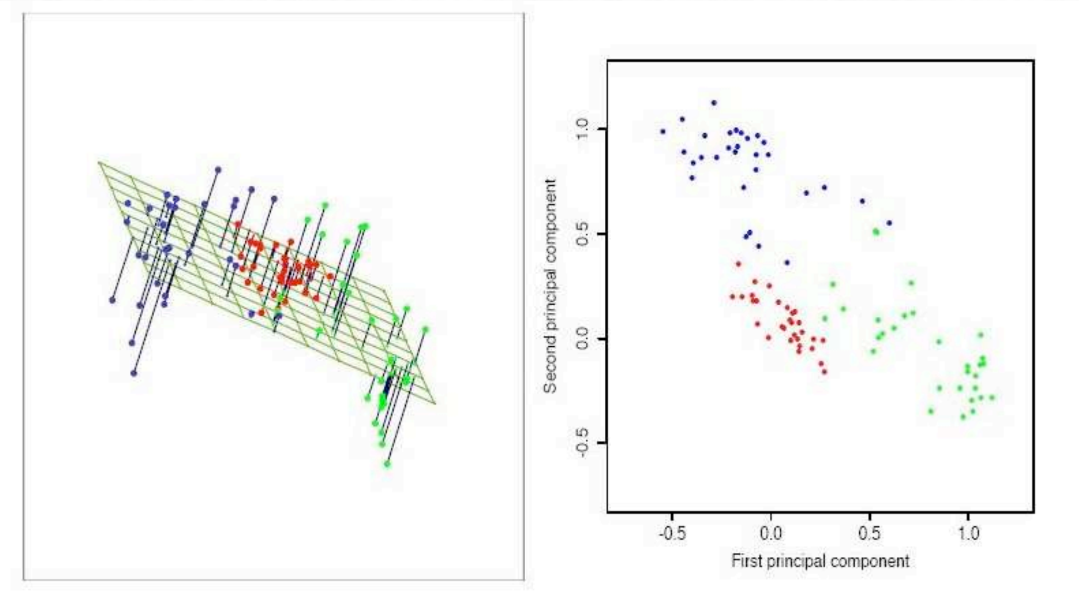
2. To find a parameterization and metric that reflects the underlying geometry of the data.

Natural occurring data often have a low intrinsic dimensionality.

Underlying dimension, underlying basis functions?



Principal Component Maps = Classical Multi-Dimensional Scaling (MDS)



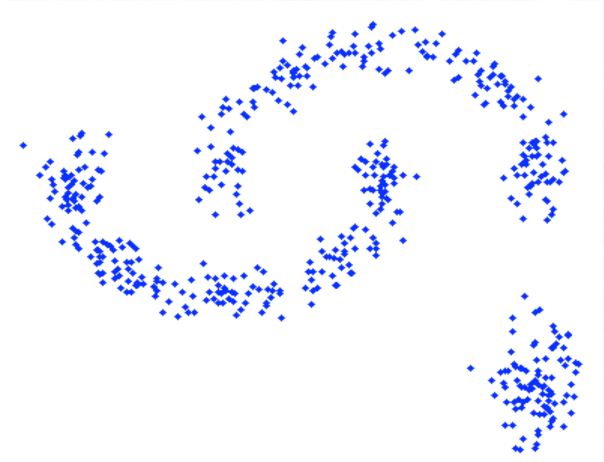
Elements of Statistical Learning,
Hastie, Tibshirani, and Friedman, pg. 488

$$\text{Minimize } \delta = \sum_{i,j}(d_{ij}^2 - \widehat{d_{ij}^2}), \quad \text{where } \mathrm{d}_{ij}^2 = \|x_i - x_j\|^2$$

How about more complex data structures?

Non-Linear Dimension Reduction and Data Parameterization via Spectral Connectivity Analysis

(Diffusion Maps, Euclidean Commute Time Maps and other metric-based random walk formulations of spectral kernel methods)



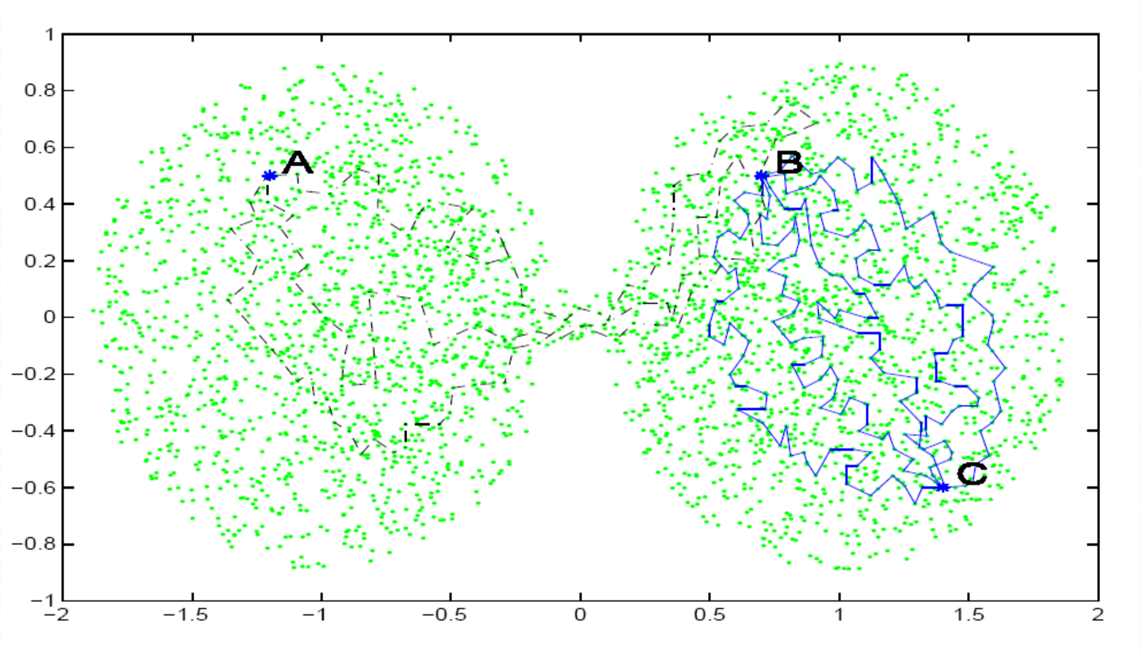
- Starting points: set of objects and similarity measure that makes sense locally (flexibility), e.g

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \quad s(x_i, x_j) = \|x_i - x_j\|$$

- Basic idea: Integrate local information from overlapping neighborhoods into global representation

Non-Linear Dimension Reduction and Data Parameterization via Spectral Connectivity Analysis

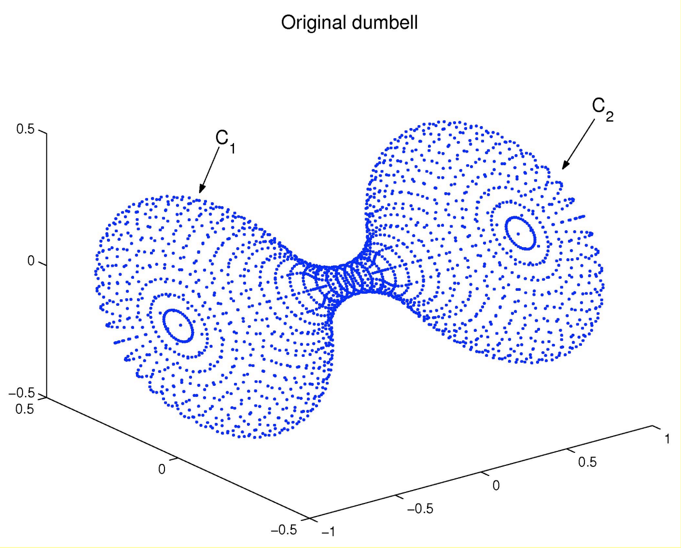
- Create a distance d(x,y) that measures "connectivity" or how easy information "flows" from x to y (Markov chain on your data)
- Find x'=f(x) and y'=f(y) so that $d(x,y) \approx \|x' - y'\|$
- Use only the first few components of x' and y' --- eigenvectors of Markov transition matrix

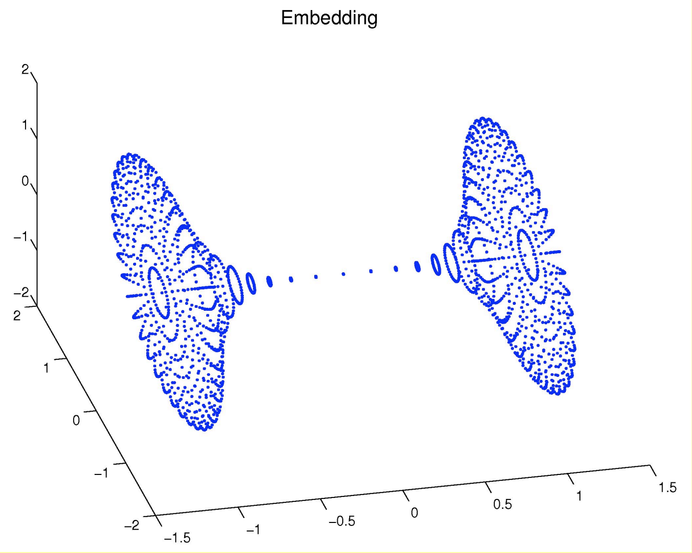


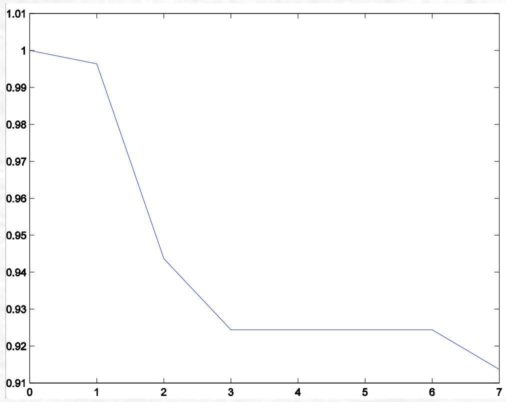
Integrates all paths of length t connecting x and y. Extremely robust to noise!

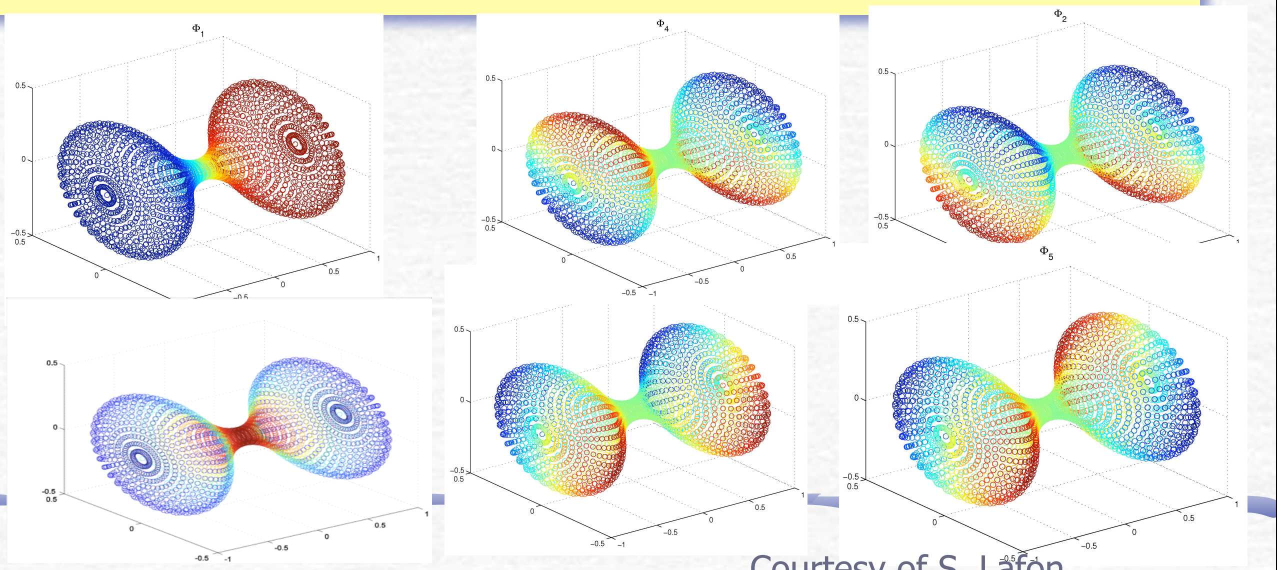
$$D_m^2(x,z) = \|p_m(x,\cdot) - p_m(z,\cdot)\|^2_{1/\phi_0}$$

Example: Hour-glass surface



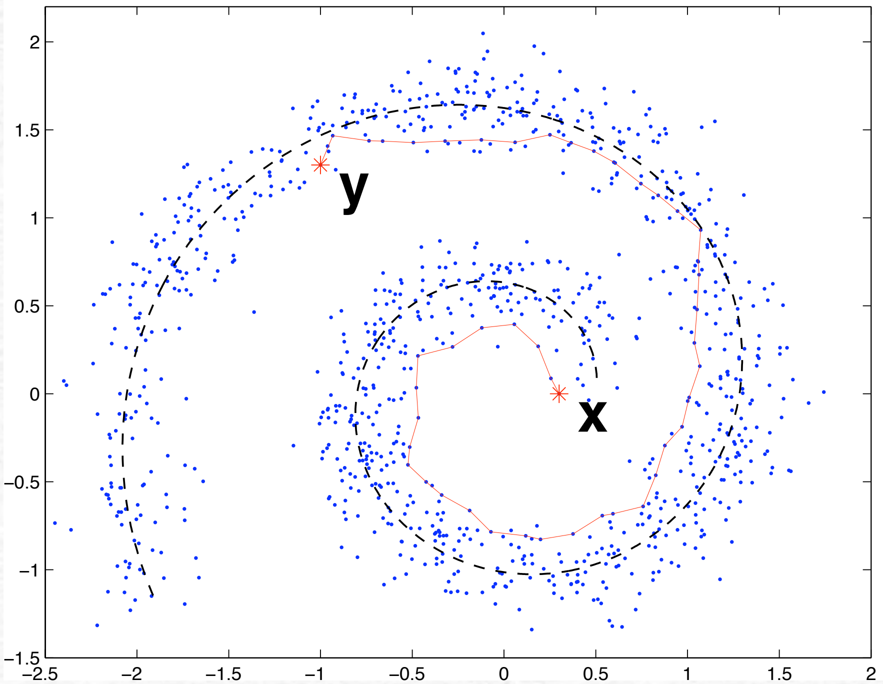


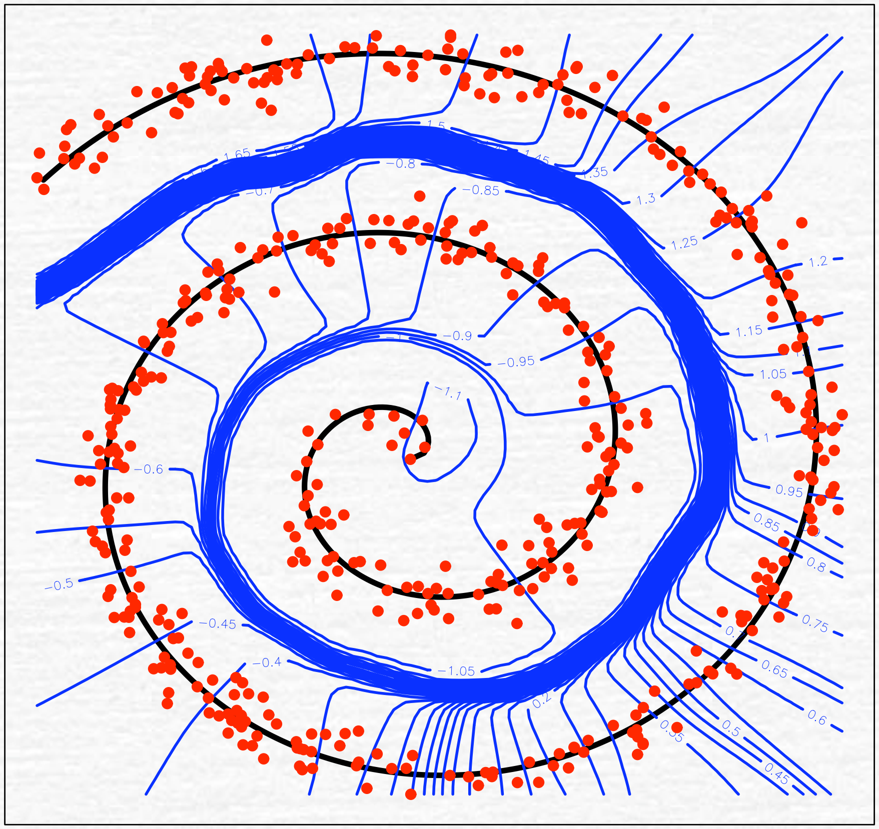




Courtesy of S. Lafon

Unraveling the underlying structure by SCA



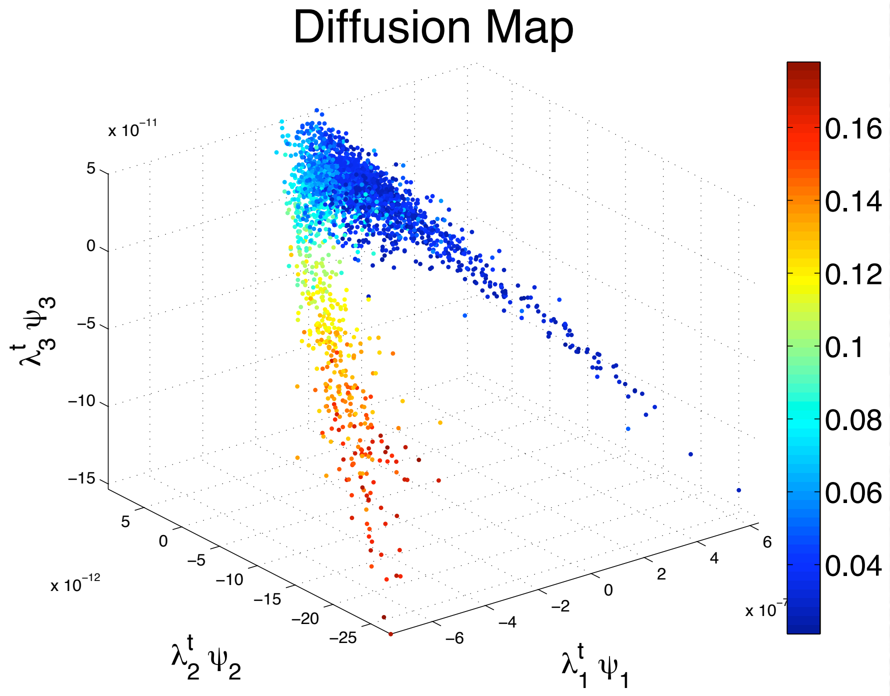


Use underlying basis functions for visualization, clustering, regression, sampling strategies, quick searches in large databases, etc.

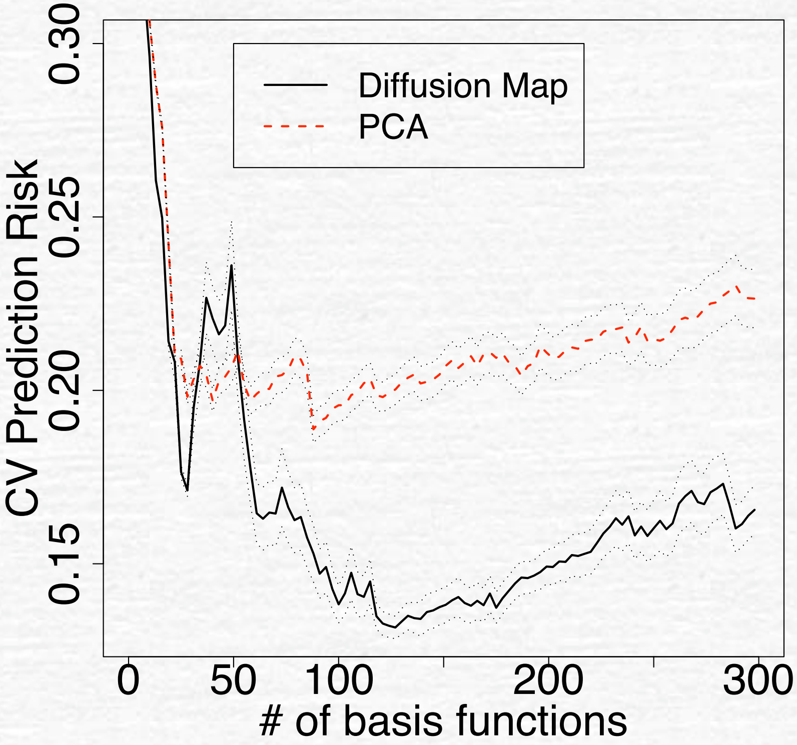
Next: Examples for astrophysical problems

Red Shift Estimation from SDSS spectra

[J. Richards et al, 2009]



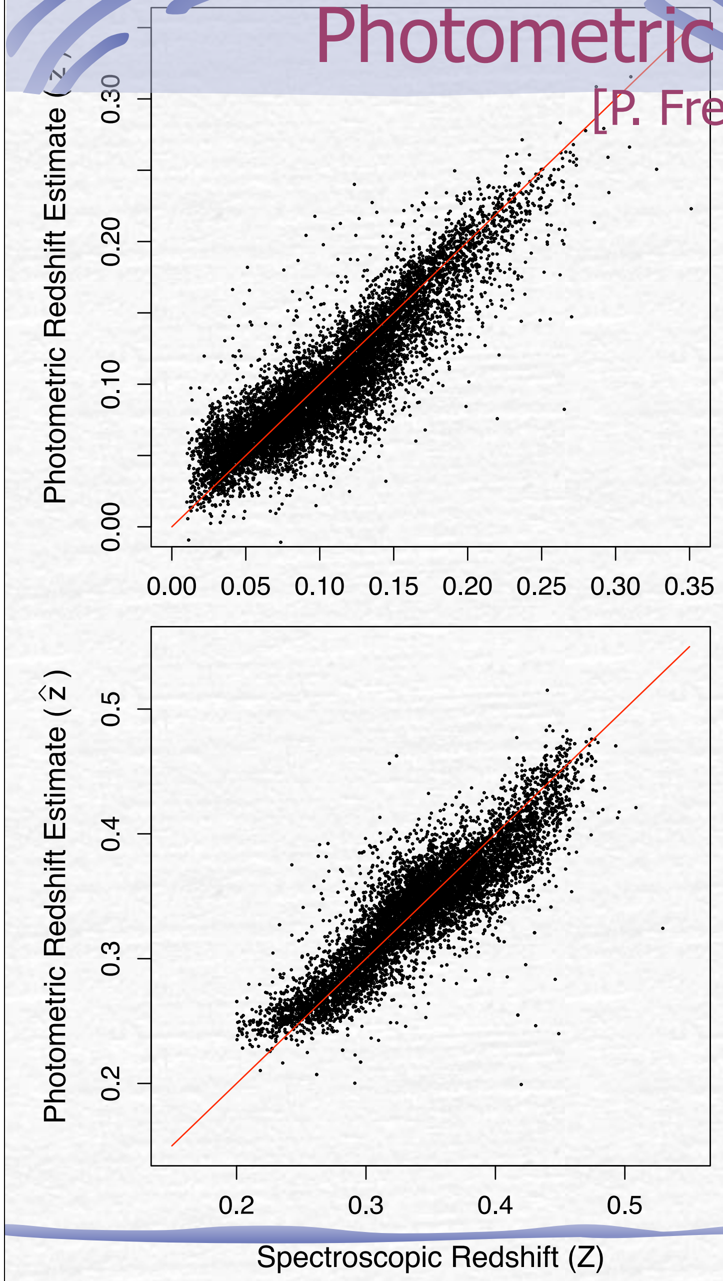
Embedding of a sample of 2793 SDSS spectra with SDSS z CL>0.99. Color codes for log(1+z).



Adaptive regression using orthogonal eigenfunctions: $r(x) = \sum_j \beta_j \psi_j(x)$

Photometric Redshift Estimation

[P. Freeman et al, 2009]



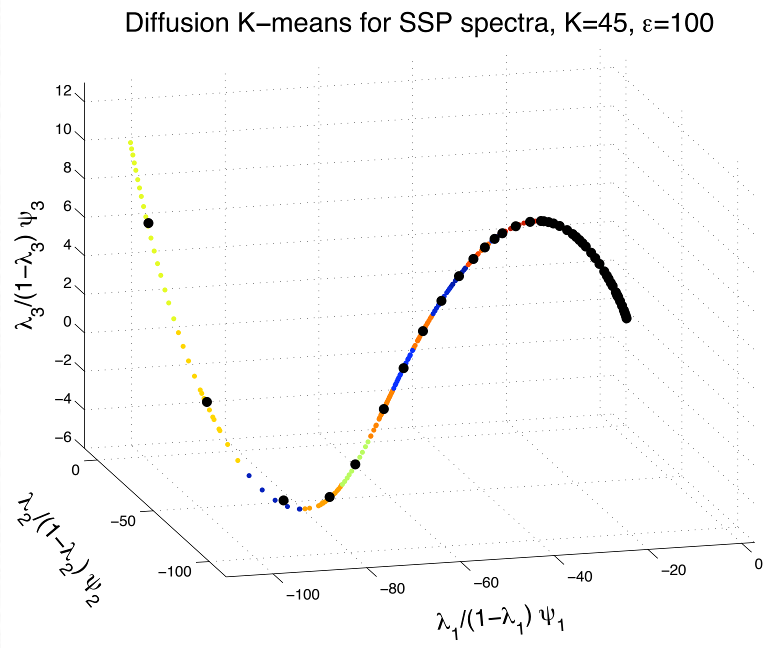
Top: predictions for MSG validation set.

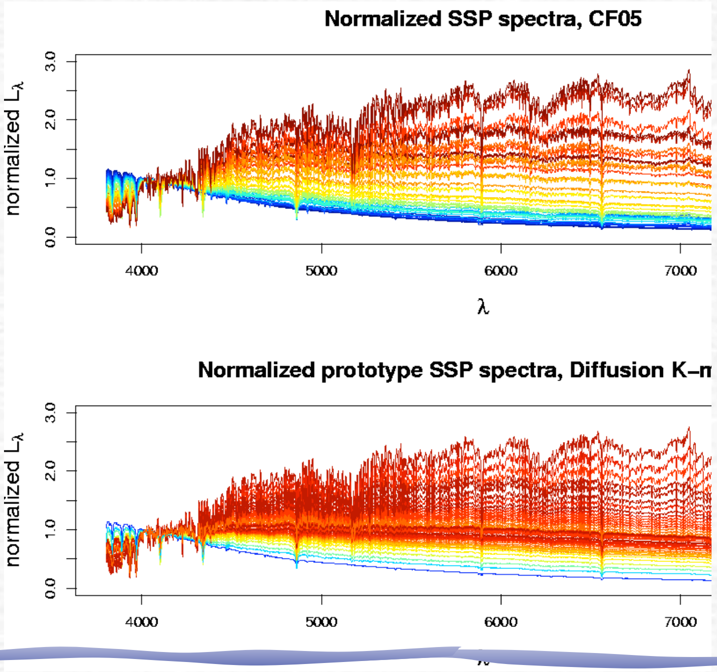
Train orthogonal series model on 9749 randomly selected objects. Estimate redshifts and basis functions for another 340,989 galaxies using the Nystrom extension

Adaptive regression using orthogonal eigenfunctions: $r(x) = \sum_j \beta_j \psi_j(x)$

Estimation of Star Formation History Using Galaxy Spectra [Cid Fernandes et al. 2004; J. Richards et al, 2009]

- Population synthesis: Model galaxy spectra as linear combinations of observable data from K simple stellar populations (SSPs)
- Cid Fernandes et al: “Elements of the base should span the range of spectral properties observed in sample galaxies and provide enough resolution in age (t) and metallicity (Z) to address the desired questions”
- Sampling strategy? How do we choose a grid of t and Z?





Basis spectra for CF05 and Diffusion K-means colored by log t

SUMMARY. Exploiting non-linear sparse structure in astronomical data

- Natural data often have low-dimensional structure
- In complex settings, linear methods may not be adequate.
- SCA learns the underlying non-linear geometry (basis functions). Can greatly improve data understanding, visualization, high-dimensional inference, sampling strategies and guide prior beliefs

- Open questions:
  - How useful are these techniques for astronomy?
  - Are they viable for immense data sets (billions of objects)?
  - Semi-supervised learning --- small set of labeled data; improved prediction by learning geometry from a large amount of unlabeled data (guide prior beliefs)

Figure 1: Unlabeled Data and Prior Beliefs