Nonparametric Procedures that Exploit Sparse (Nonlinear) Structure in High Dimensions

Ann B. Lee
Department of Statistics
Carnegie Mellon University

Joint work with Rafael Izbicki, Peter Freeman, Chad Schafer

Statistical Challenges in Modern Astronomy

- Chris Genovese: "not just the amount/size but also the complexity and richness of the data are increasing"
- Due to both the physical processes and the measurement techniques that generated the data

Claim

- To fully take the richness and complexity of the data into account, we need to

  1. use nonparametric models

  2. work with high-dimensional data objects $\mathbf{x}$ (e.g. entire spectra, images, light curves, etc)

  3. move beyond regression/classification to estimating probability distributions of complex, high-dimensional objects $\mathbf{x}$

Claim

- To fully take the richness and complexity of the data into account, we need to

  1. use nonparametric models

  2. work with high-dimensional data objects $\mathbf{x}$ (e.g. entire spectra, images, light curves, etc)

  3. move beyond regression/classification to estimating probability distributions of complex, high-dimensional objects $\mathbf{x}$

Claim

- To fully take the richness and complexity of the data into account, we need to

  1. use nonparametric models

  2. work with high-dimensional data objects $\mathbf{x}$ (e.g. entire spectra, images, light curves, etc)

  3. move beyond regression/classification to estimating probability distributions of complex, high-dimensional objects $\mathbf{x}$

Let's put this in the context of astronomy...

Claim

- To fully take the richness and complexity of the data into account, we need to

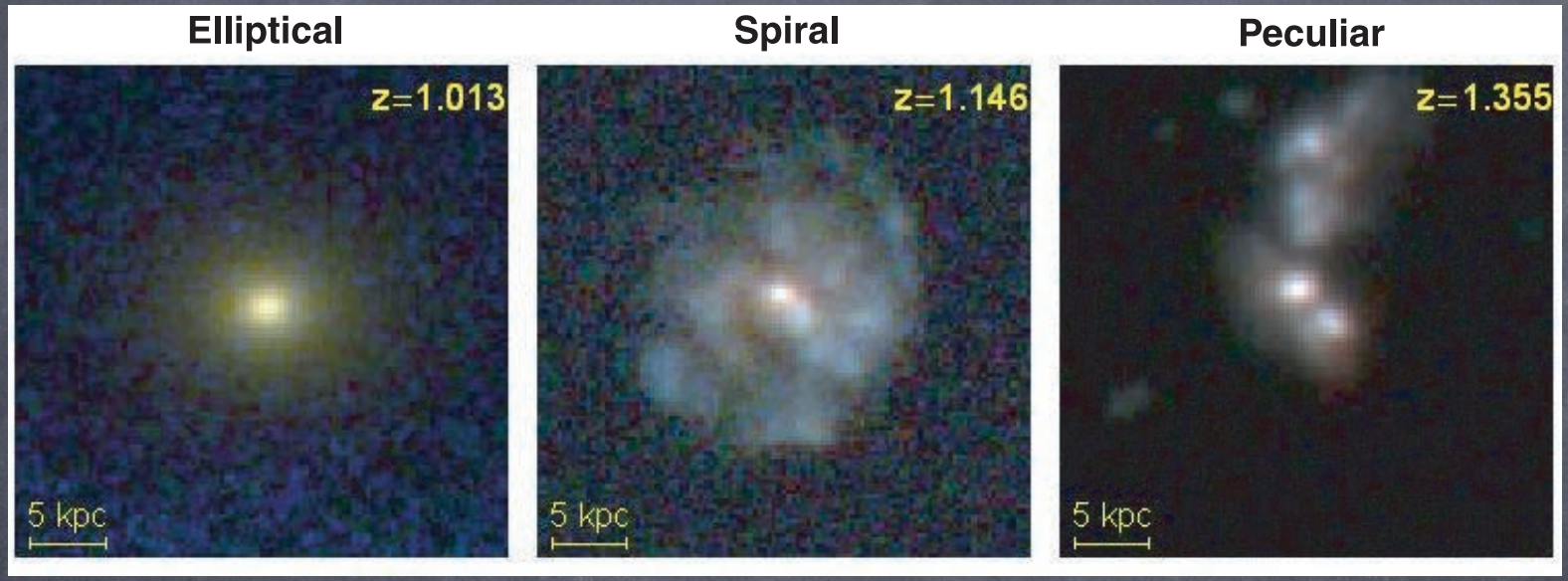
  1. use nonparametric models

  2. work with high-dimensional data objects $\mathbf{x}$ (e.g. entire spectra, images, light curves, etc)

  3. move beyond regression/classification to estimating probability distributions of complex, high-dimensional objects $\mathbf{x}$

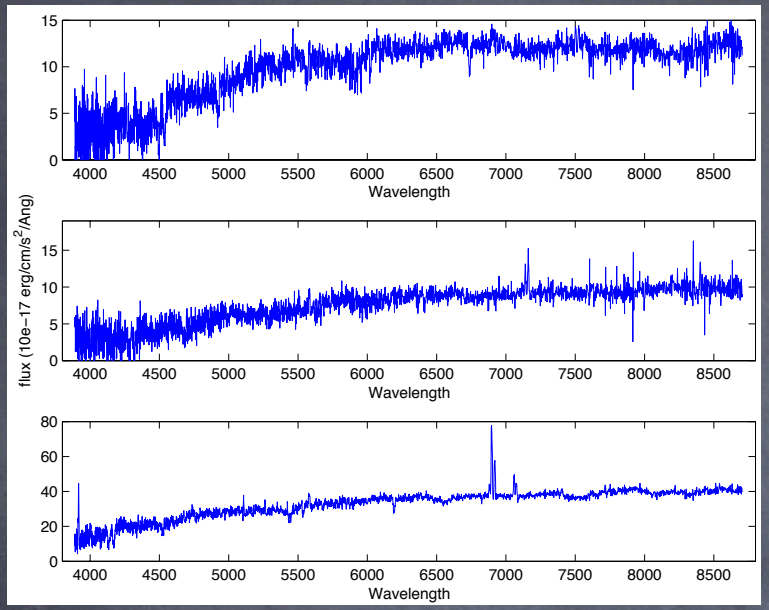
Let's put this in the context of astronomy...

Ex 1: Redshift Prediction



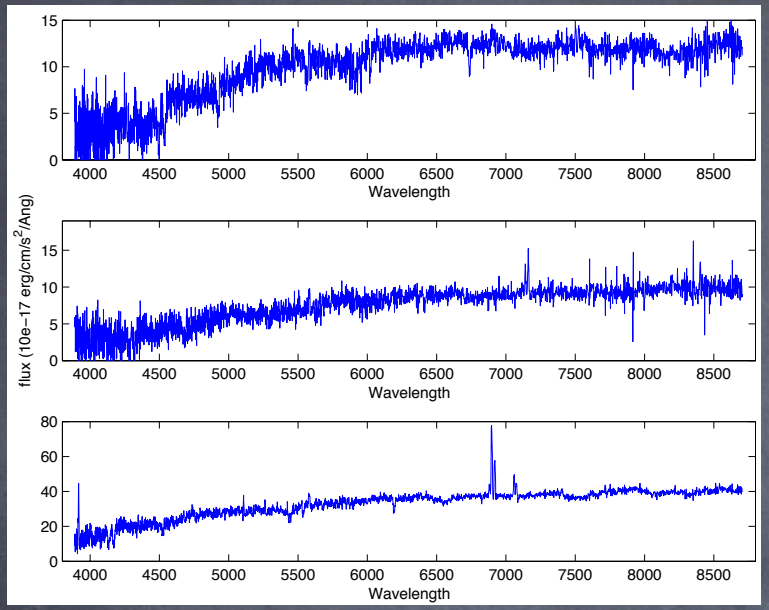
More distant objects are seen further back in time ---
but, we cannot measure distances directly....
Redshift = proxy for distance

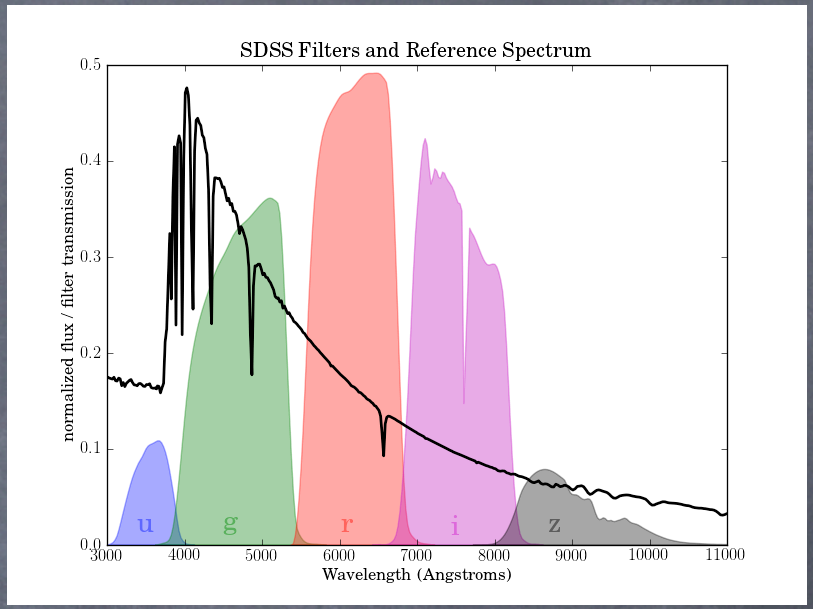
Redshift Prediction: Spectroscopy vs Photometry



- Left: High-resolution galaxy spectra
- Spectroscopy resource intensive ⇒ More than 99 percent of today's galaxy observations are instead from photometry.
- Right: Photometry (broad-band filters).
  Challenge -- Accurately estimate z using photometric data ✘.

Redshift Prediction: Spectroscopy vs Photometry





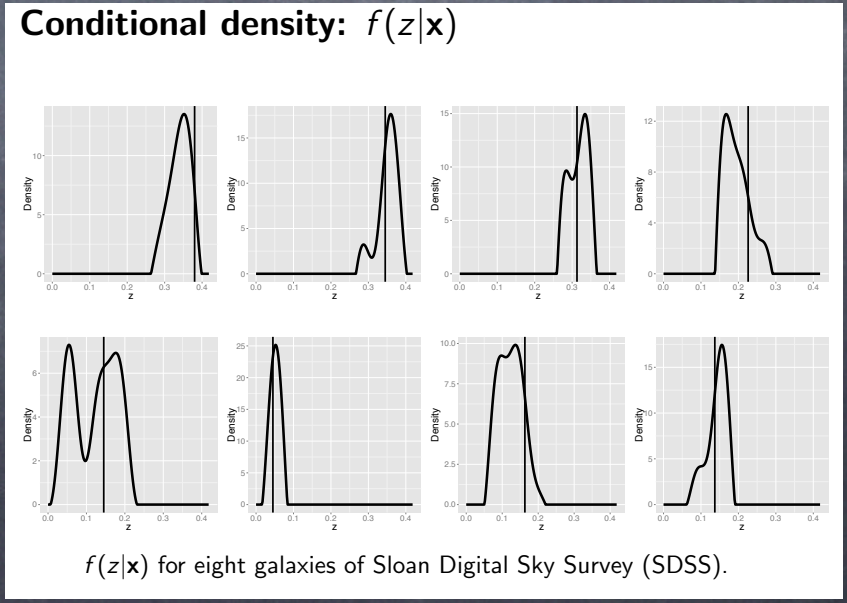
- Left: High-resolution galaxy spectra
- Spectroscopy resource intensive ⇒ More than 99 percent of today's galaxy observations are instead from photometry.
- Right: Photometry (broad-band filters)
  Challenge -- Accurately estimate z using photometric data ✘.

redshift (predictor) photometric covariates

Beyond Regression E(z|x) to Estimating f(z|x)

- Because of degeneracies and observational limitations =>
  Need the full distribution f(z|x) of the response z given x
  to quantify uncertainty in the predictions

Fig. Examples of photometric density estimates f(z|x) for SDSS galaxies. Complicated asymmetric and multimodal distributions, not easily summarized by means E(z|x) and variances V(z|x).



$f(z|\mathbf{x})$ for eight galaxies of Sloan Digital Sky Survey (SDSS).

[Izbicki and Lee, 2015]

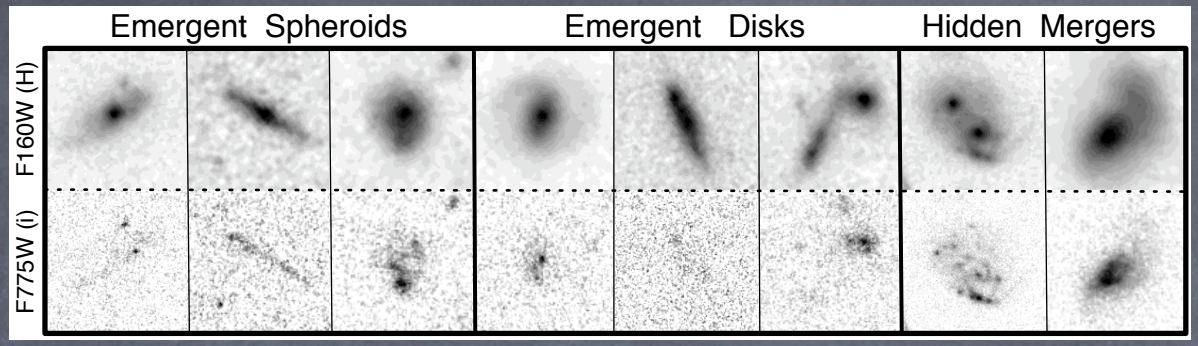
Ex 2: Evolution of Galaxy Morphology

Traditional picture:
The Canonical Hubble Sequence

The key towards answering these questions may lie in the data we already have. Not surprisingly, ***the existing coarse structural classifications fail to capture the complex assembly processes responsible for the morphological transformation of galaxies over the past 10 billion years.*** We need more precise quantitative tools to track galaxy structure and its evolution through time. The classifications of "bulge", "disk", and "other" are not sufficient, and essentially throw away vast amounts of information. And we are on the brink of even more big data. SDSS III is being followed by the deeper Pan-STARRS, Dark Energy Survey, and SDSS IV, with LSST on the horizon. HST will be followed by the infrared sensitive JWST and wide-field WFIRST, pushing our knowledge of the high-redshift universe. ALMA will trace detailed gas structures at resolutions and distances greater than HST's reach. The time has come to invest

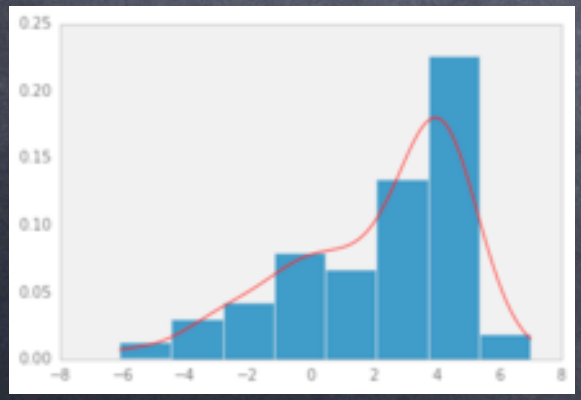
- quote from J. Lotz, Space Telescope Science Institute (2013)

Beyond Classification to Estimating f(x|z)



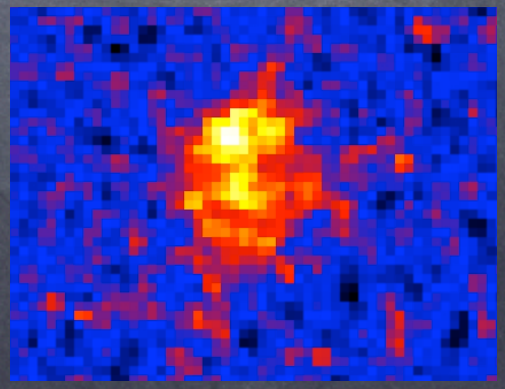
- How does the full distribution $\mathrm{P(image\,|\,z)}$ of galaxy morphologies evolve with time/redshift z? This is an extremely challenging (conditional) density estimation problem.

pdf
f(x)



x

x =



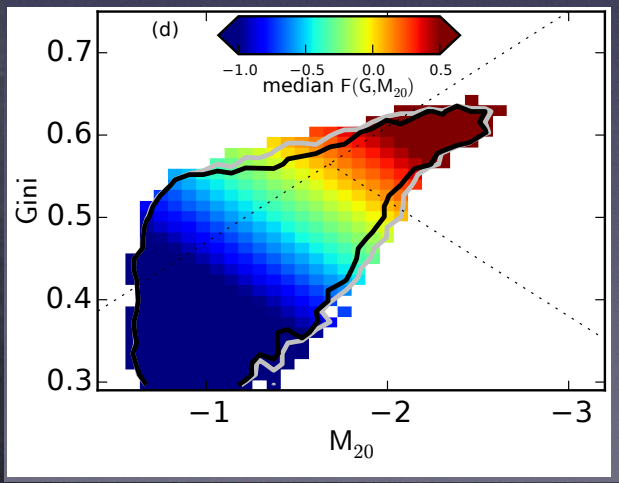
84 X 84 pixels = 7056 variables

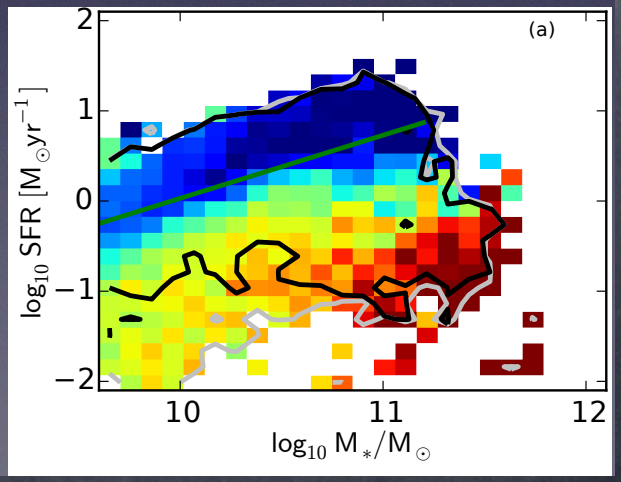
Ex 3: Learning the relationship between physical and observable properties

With regard to our second statistical aim, suppose that $\mathbf{x} \sim F$ represents the summary statistics for a population of simulated galaxies, while $\mathbf{y}$ represents the stellar masses, star-formation rates, etc., for those same galaxies. Given $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ for a set of n galaxies, we wish to learn the relationship between $\mathbf{x}$ and $\mathbf{y}$; for example, we may want to estimate the conditional density

$$f(\mathbf{y}|\mathbf{x}) = f(M_\star, SFR, \ldots \,|\, C, A, G, M_{20}, M, I, D)\,,$$

for the whole population of simulated galaxies, where the symbol | indicates that the quantities to the right are fixed.

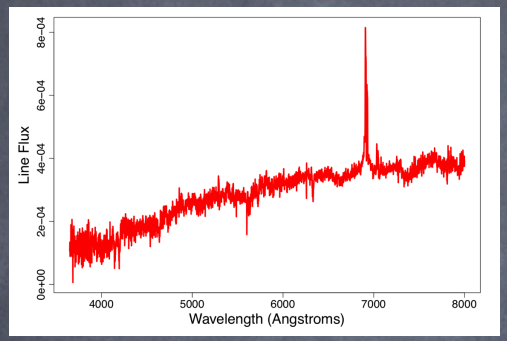


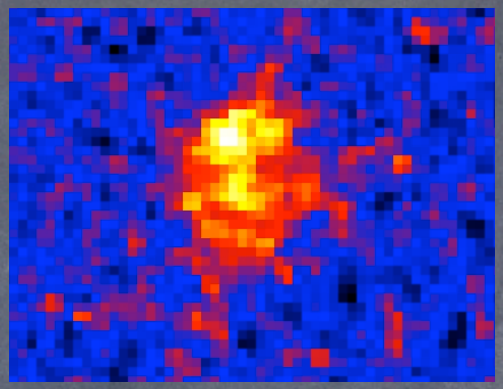


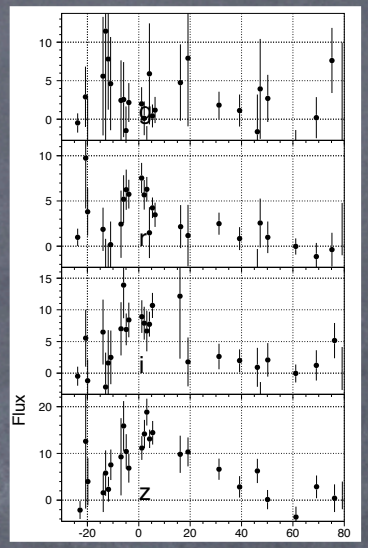
From: Peter Freeman

General Statistical Setting

X=







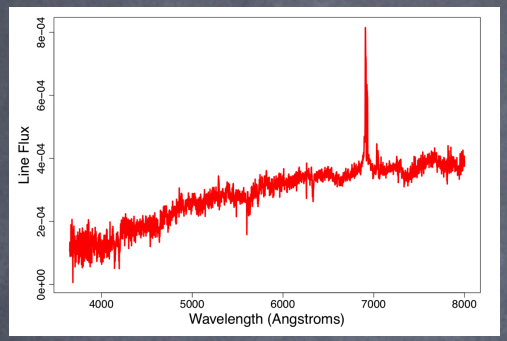
$$\mathcal{D} = \{(X_1, Z_1), \ldots, (X_n, Z_n), X_{n+1}, \ldots, X_{n+m}\},$$
$$\text{where } X_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d, \; Z_i \in R$$

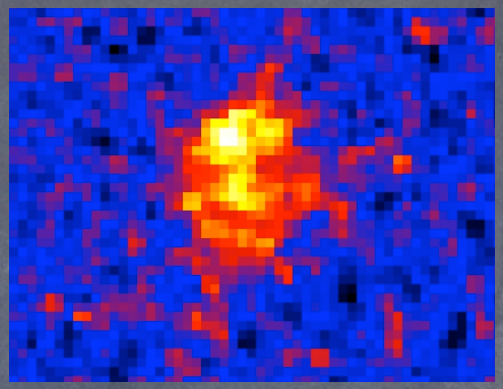
Prediction: E[Z|**x**], f(z|**x**)

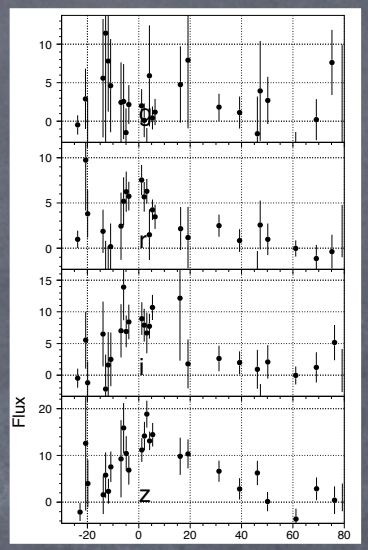
Density estimation: f(**x**), f(**x**)/g(**x**), f(**x**|z;θ)

General Statistical Setting

X=







$$\mathcal{D} = \{(X_1, Z_1), \ldots, (X_n, Z_n), X_{n+1}, \ldots, X_{n+m}\},$$
$$\text{where } X_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d, \; Z_i \in R$$

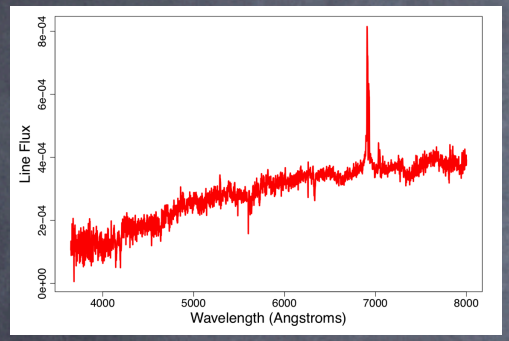
Prediction: E[Z|**x**], f(z|**x**)

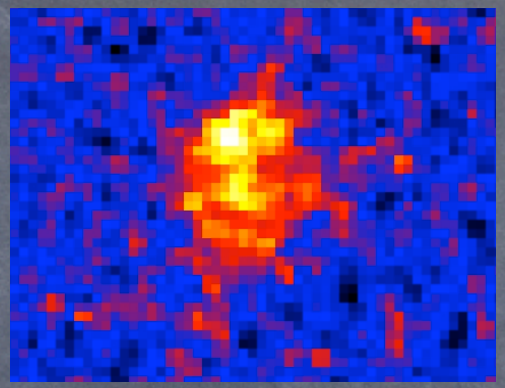
Density estimation: f(**x**), f(**x**)/g(**x**), f(**x**|z;θ)

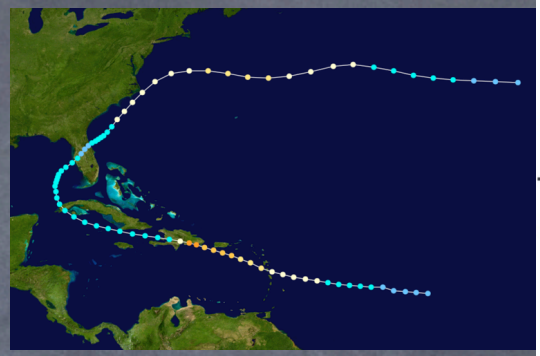
How do we estimate such functions in high dimensions?

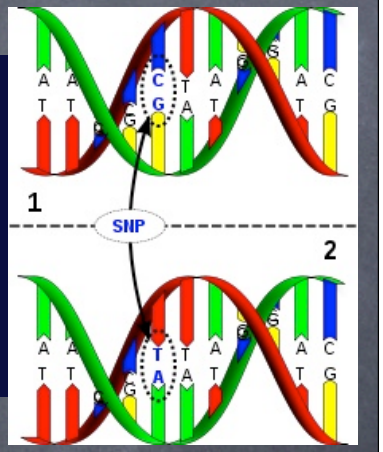
Key: Exploit Sparse Structure in Astronomical Data

X=







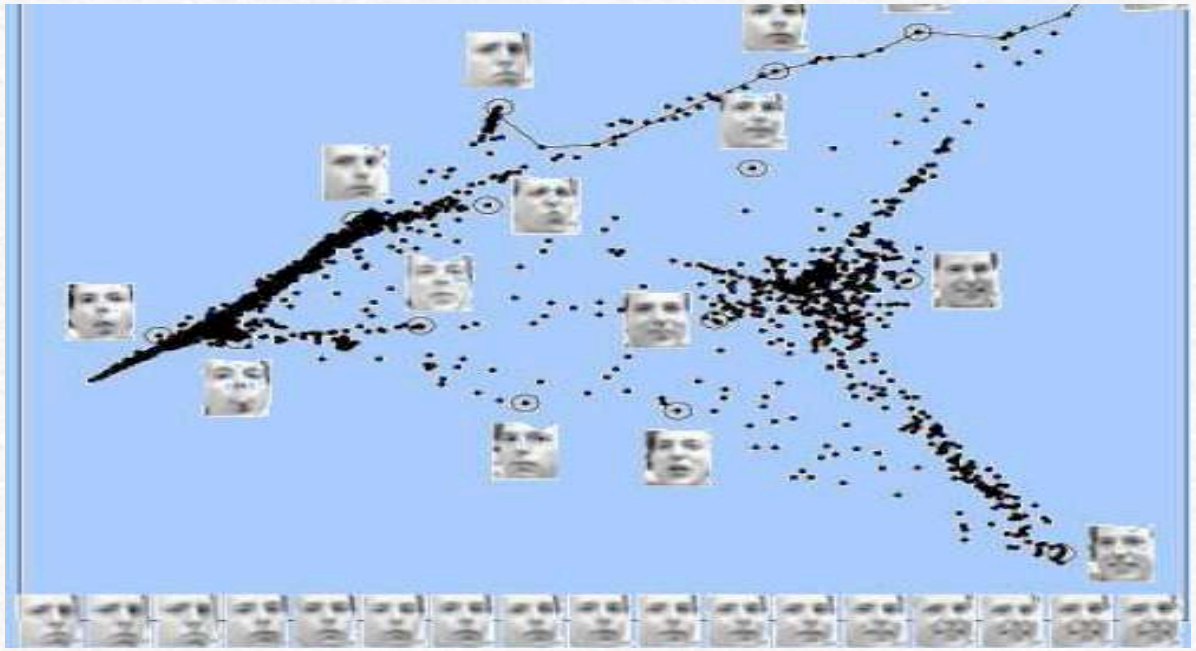


Assume data $X_i \sim P_X$

- The ambient dimension may be large but most of the sample space is empty.

"Sparsity" = P_X places most of its mass on a subset of the state space with small Lebesgue measure

Ex: Low-Dimensional Structure in Images of Faces



Each picture $X=(x_1,...,x_n)$ pixel values. High-dimensional data!
Low-dimensional structure due to certain constrained deformations of the face.

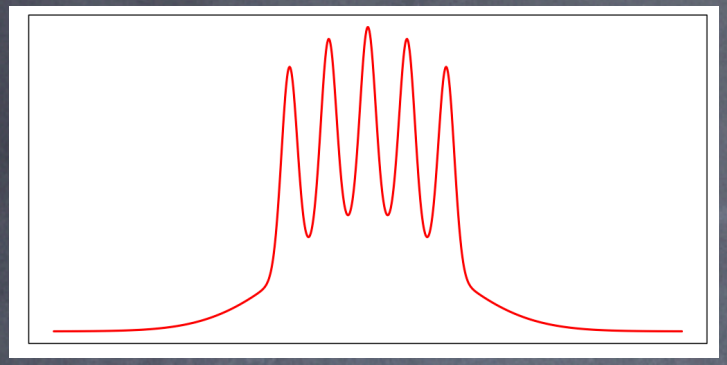
Q: How to construct nonparametric methods that adapts to the intrinsic data geometry?

The Spectral Series Method

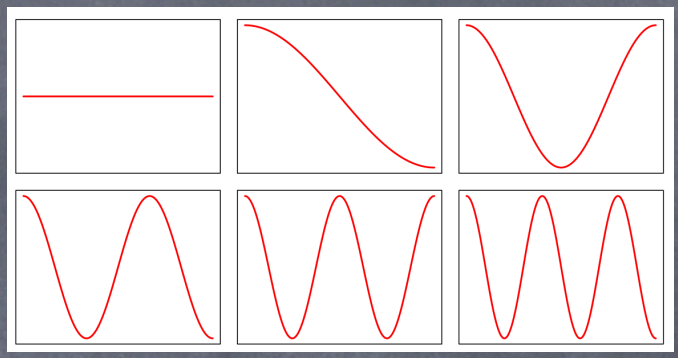
- Spectral Kernel Methods + Orthogonal Series

"Classical" Orthogonal Series Estimation (1D)

Unknown density f(x)

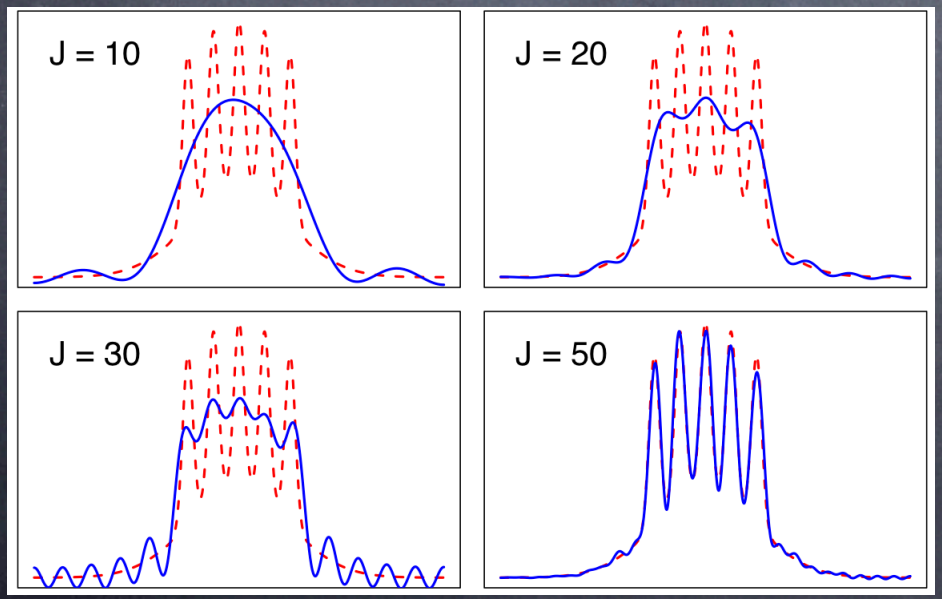


Fourier basis $\{\varphi_j(x)\}$



If $f \in L_2(0,1)$, then $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$

where $\int_0^1 \phi_i(x)\phi_j(x)dx = \delta_{i,j}$, and $\beta_j = \int_0^1 \phi_j(x) f(x) dx = \mathbb{E}[\phi_j(X)]$.



Density estimate

$$\widehat{f}(x) = \sum_{j=1}^{J} \widehat{\beta}_j \phi_j(x)$$

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(x_i)$$

Currently not known how to extend to higher than 2-3 dims.

We Propose "Spectral Basis"

- Gaussian kernel $w(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-d^2(\mathbf{x},\mathbf{y})}{4\epsilon}\right)$

data objects (e.g. entire images, spectra, light curves)

- Normalize:

$$k(\mathbf{x}, \mathbf{y}) = \frac{w(\mathbf{x}, \mathbf{y})}{\int w(\mathbf{x}, \mathbf{y}) dP(\mathbf{y})},$$

$$K(f)(\mathbf{x}) = \int k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) dP(\mathbf{y})$$

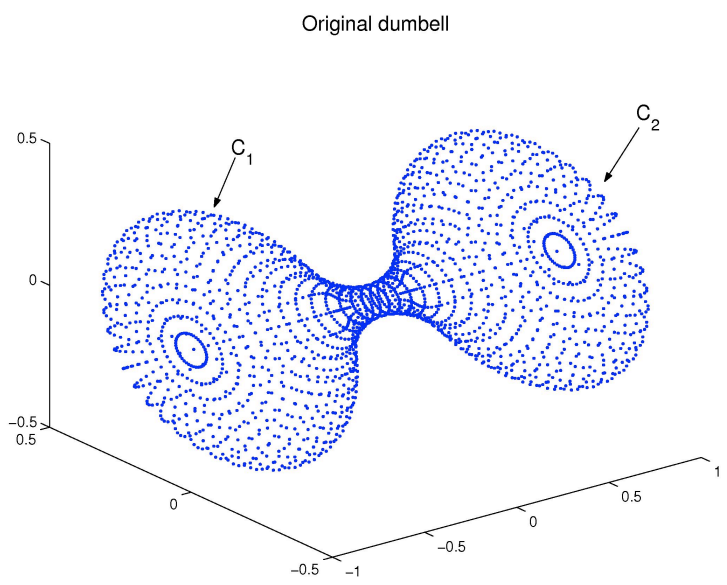
- Consider the eigenfunctions of K

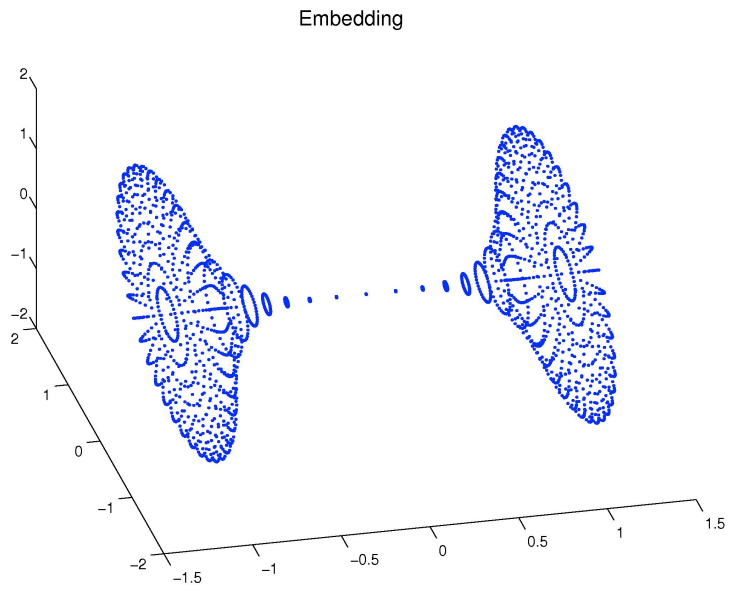
$$K(\psi_j)(\mathbf{x}) = \lambda_j \psi_j(\mathbf{x})$$

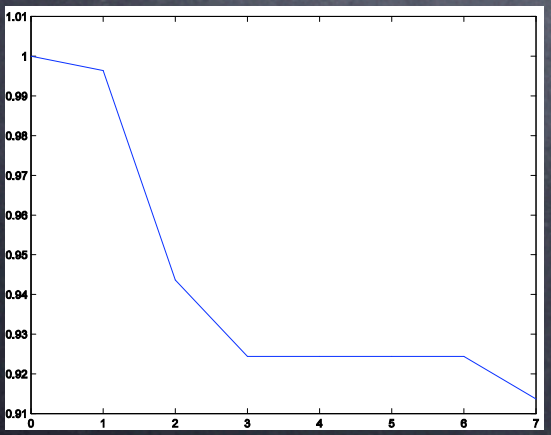
$$\{\psi_j(\mathbf{x})\}_{j \in \mathbb{N}}$$

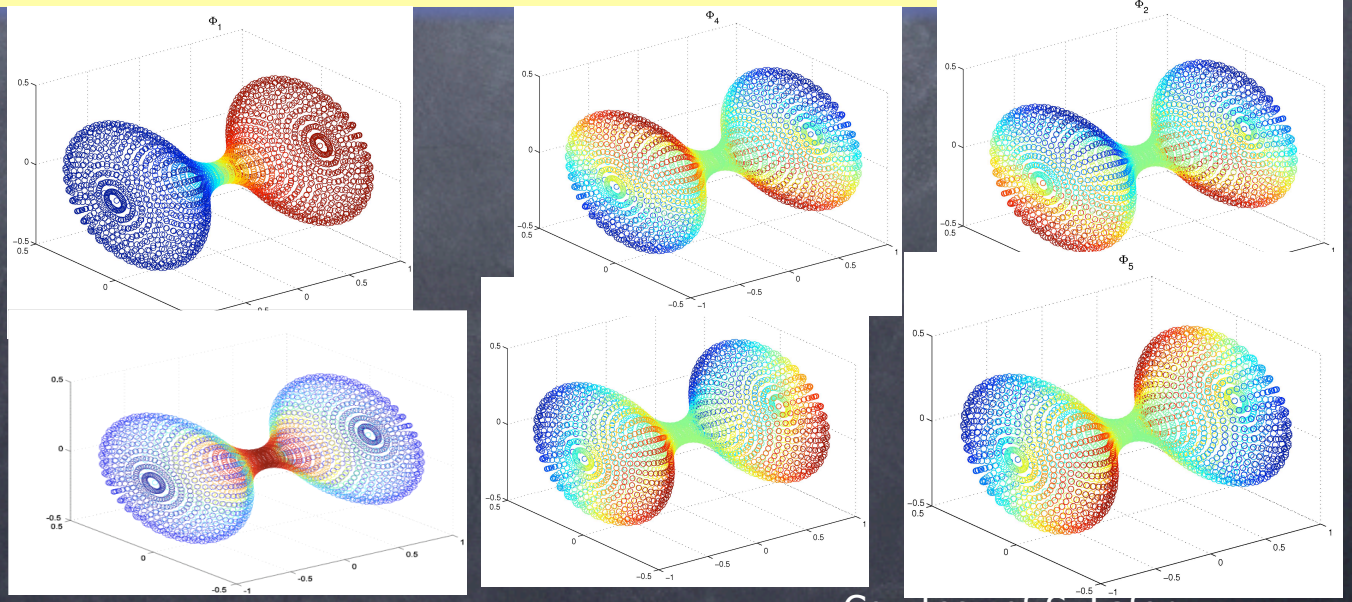
"Spectral Basis"

Example: Hour-glass surface



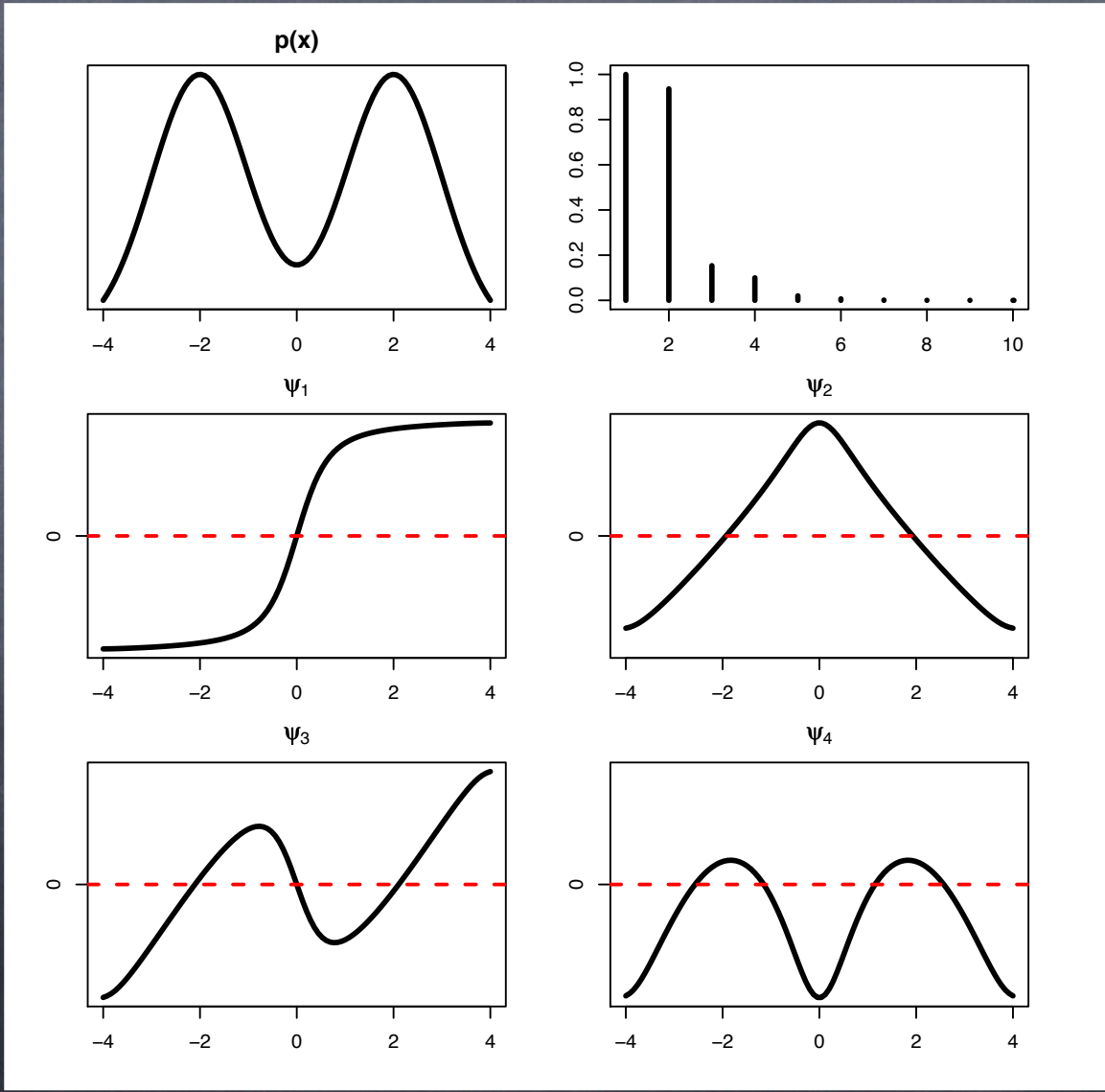




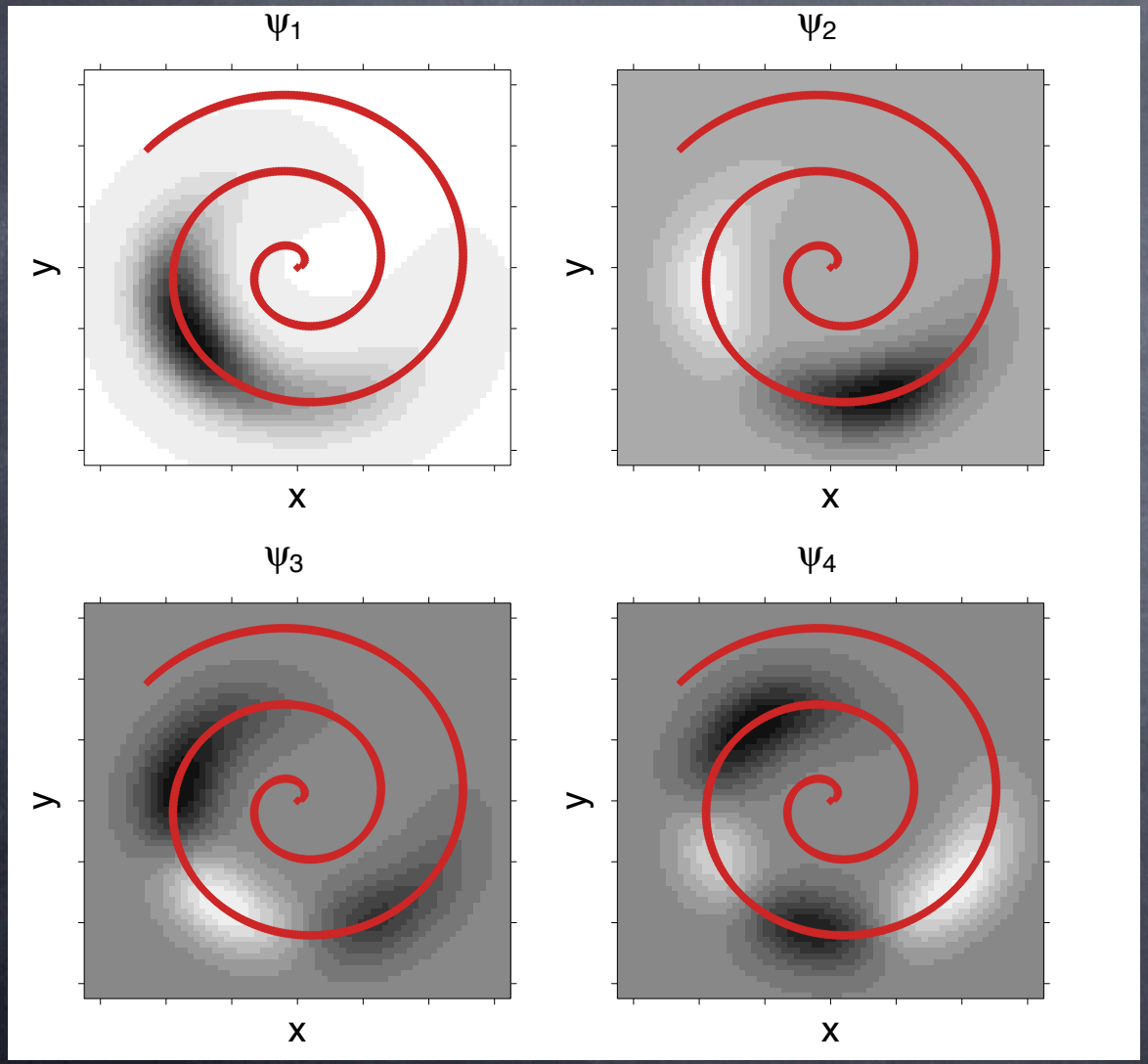


Courtesy of S. Lafon

- Spectral Basis: If a distribution has a few well-defined clusters, then the first few eigenfunctions behave like indicator functions for these clusters. The rest of the eigenfunctions provide smooth Fourier-like basis functions within each cluster.



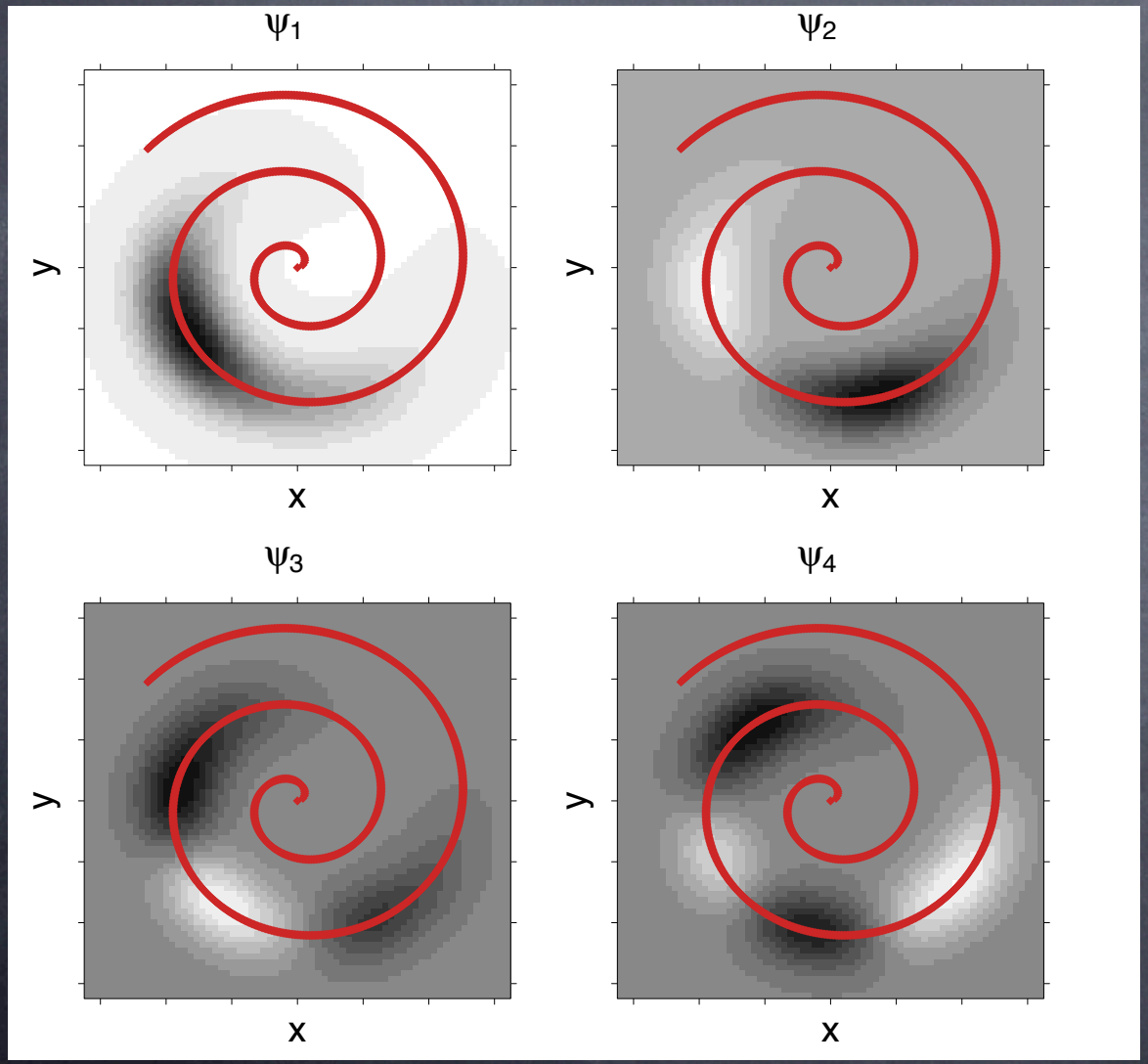
Spectral Basis -- Orthogonal Basis Adapted to the Intrinsic Data Geometry



Contour plots of the top eigenfunctions for data on a spiral.

$$\int_{\mathcal{X}} \psi_i(x)\psi_j(x)dP_X(x) = \delta_{i,j}$$

Spectral Basis -- Orthogonal Basis Adapted to the Intrinsic Data Geometry



Contour plots of the top eigenfunctions for data on a spiral.

$$\int_{\mathcal{X}} \psi_i(x)\psi_j(x)dP_X(x) = \delta_{i,j}$$

high-dimensional object

$g(\mathbf{x})$ smooth $\Rightarrow g(\mathbf{x}) \approx \sum_{j=1}^{J} \beta_j \psi_j(\mathbf{x})$

No need of tensor products

Next

- Two examples of high-dimensional inference beyond regression/classification:

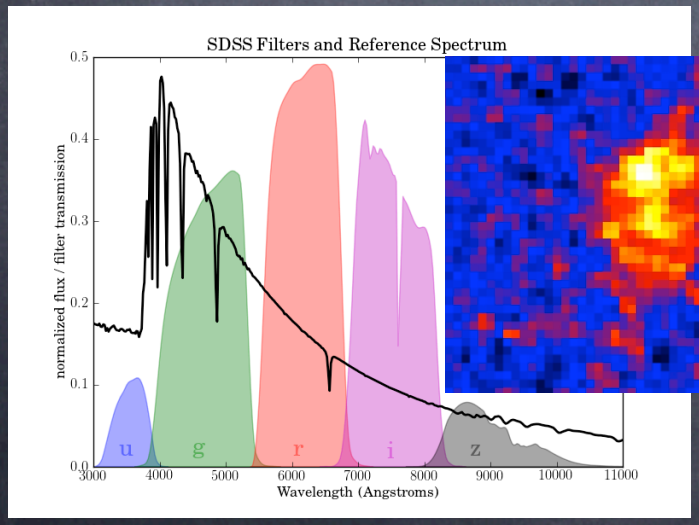
  1. Photo-z estimation: Estimating $f(z|\mathbf{x})$ in a regression setting.

  2. Approximate likelihood computation:
     Estimating $f(\mathbf{x}|\theta)$ using output ($\mathbf{x}$ = entire image) from a simulation model.

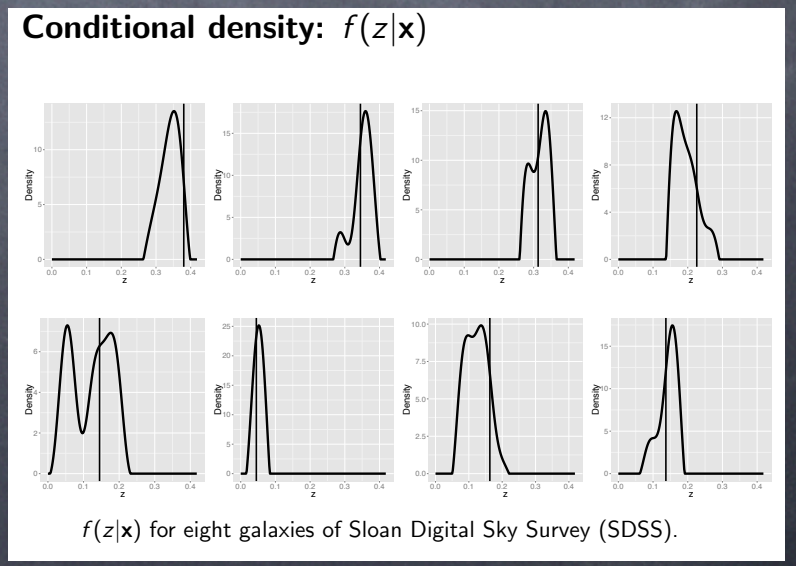
I. Photo-z Estimation

$$\mathcal{D} = \{(X_1, Z_1), \ldots, (X_n, Z_n), X_{n+1}, \ldots, X_{n+m}\},$$

- z = "true" redshift (spectroscopically confirmed)
- **x** = photometric covariates
- Goal: Estimate f(z|**x**) instead of just r(**x**)=E[Z|**x**].



Photometry



$f(z|\mathbf{x})$ for eight galaxies of Sloan Digital Sky Survey (SDSS).

Estimates of f(z|x) from photometry

CDE via Spectral Series

spectral basis

$f(z|\mathbf{x}) = \sum_{i,j} \beta_{i,j} \Psi_{i,j}(z, \mathbf{x})$, where $\Psi_{i,j}(z, \mathbf{x}) = \phi_i(z)\psi_j(\mathbf{x})$.

Orthogonality: $\langle \phi_k, \phi_\ell \rangle = \delta_{k,\ell}$ and $\langle \psi_k, \psi_\ell \rangle_{P_\mathbf{X}} = \delta_{k,\ell}$.

Hence, $\beta_{i,j} = \iint f(z|\mathbf{x}) \Psi_{i,j}(z, x) dP(\mathbf{x}) dz = \mathbb{E}[\Psi_{i,j}(Z, \mathbf{X})]$.

CDE via Spectral Series

spectral basis

$f(z|\mathbf{x}) = \sum_{i,j} \beta_{i,j} \Psi_{i,j}(z, \mathbf{x})$, where $\Psi_{i,j}(z, \mathbf{x}) = \phi_i(z)\psi_j(\mathbf{x})$.

Orthogonality: $\langle \phi_k, \phi_\ell \rangle = \delta_{k,\ell}$ and $\langle \psi_k, \psi_\ell \rangle_{P_\mathbf{X}} = \delta_{k,\ell}$.

Hence, $\beta_{i,j} = \iint f(z|\mathbf{x}) \Psi_{i,j}(z, x) dP(\mathbf{x}) dz = \mathbb{E}[\Psi_{i,j}(Z, \mathbf{X})]$.

- Spectral Series conditional density estimator

$$\widehat{f}(z|\mathbf{x}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \widehat{\beta}_{i,j} \widehat{\Psi}_{i,j}(z, \mathbf{x}),$$

$$\widehat{\beta}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \widehat{\Psi}_{i,j}(z_k, \mathbf{x}_k), \quad \widehat{\Psi}_{i,j}(z_k, \mathbf{x}_k) = \phi_i(z_k)\widehat{\psi}_j(\mathbf{x}_k),$$

Fast Tuning of Parameters

- To tune parameters, minimize estimated loss on a validation set {(X_1',Z_1'),..(X_n',Z_n')}.

Loss function:

$$L(\widehat{f}, f) = \iint \left(\widehat{f}(z|\mathbf{x}) - f(z|\mathbf{x})\right)^2 dP(\mathbf{x})dz$$

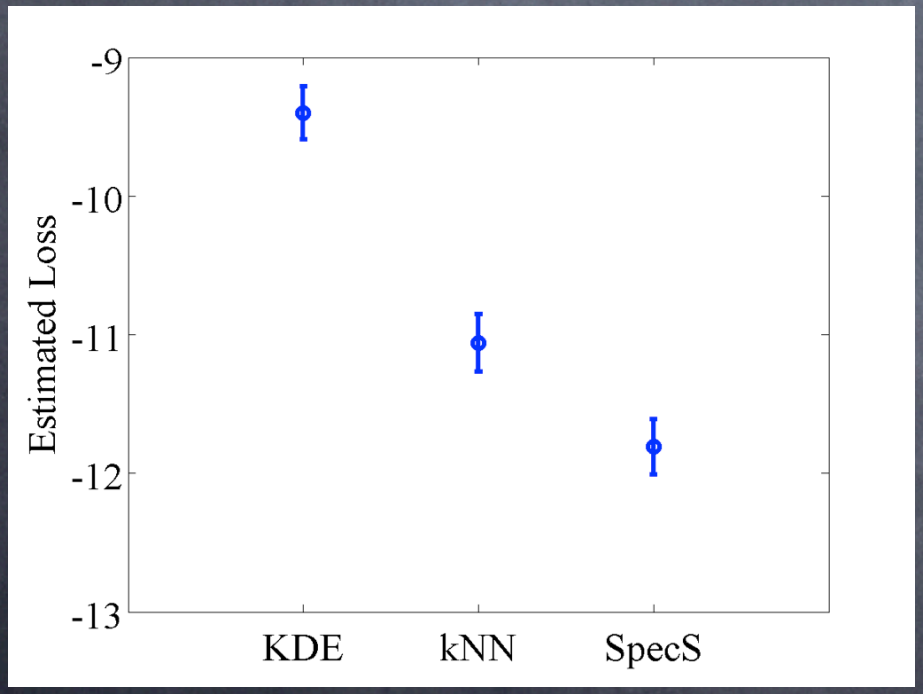
Estimated loss (up to a constant):

$$\widehat{L}(\widehat{f}, f) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{J} \widehat{\beta}_{i,j}\widehat{\beta}_{i,m}\widehat{W}_{j,m} - 2\frac{1}{n'}\sum_{k=1}^{n'} \widehat{f}(z'_k|\mathbf{x}'_k) + C,$$

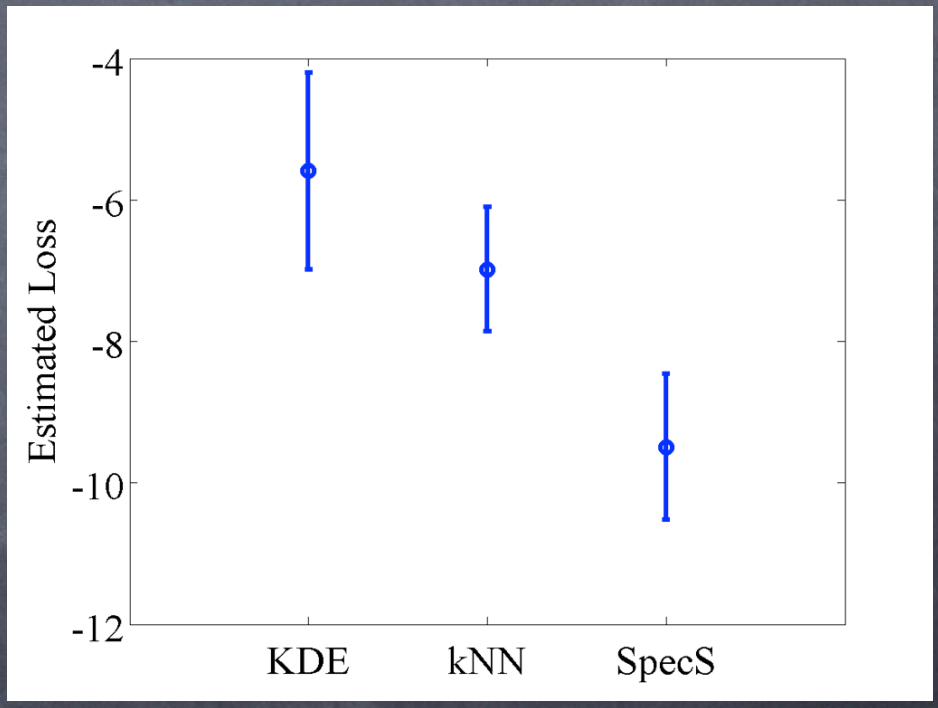
where $\widehat{W}_{j,m} = (n')^{-1}\sum_{k=1}^{n'} \widehat{\psi}_j(\mathbf{x}'_k)\widehat{\psi}_m(\mathbf{x}'_k).$

Estimating f(z|x): Performance of Different Estimators

Set 1 [Sheldon et al., 2012]:
10,000 galaxies from multiple surveys with d=10 covariates.

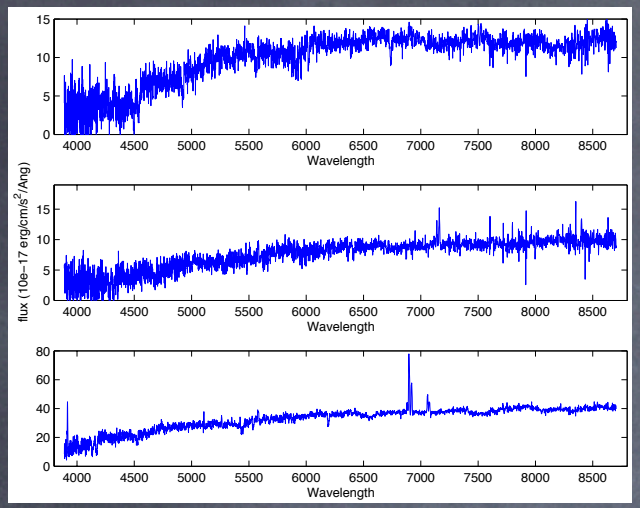


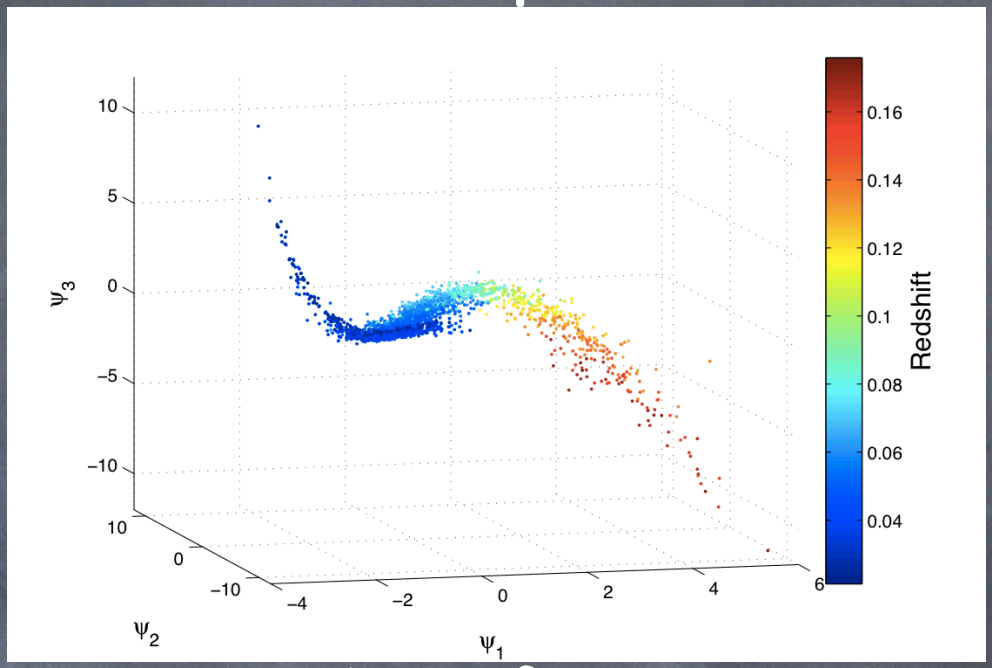
Set 2 [T. Dahlen 2013]:
752 galaxies from COSMOS with d=37 covariates



- Spectral Series have a significantly smaller loss than a nonparametric kernel smoother and Nearest Neighbors.

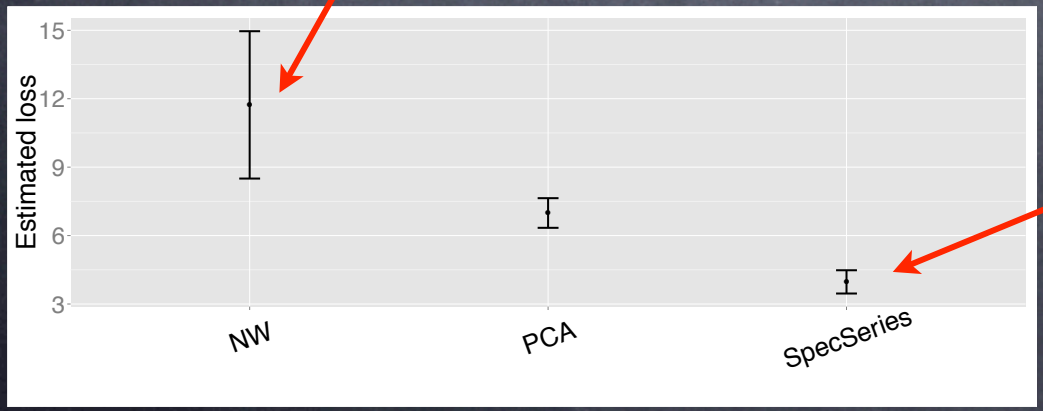
"Proof of Concept": Redshift Prediction Using High-Resolution Spectra





Trad non-parametric regression (and kNN)

Spectral basis for visualization "diffusion map"

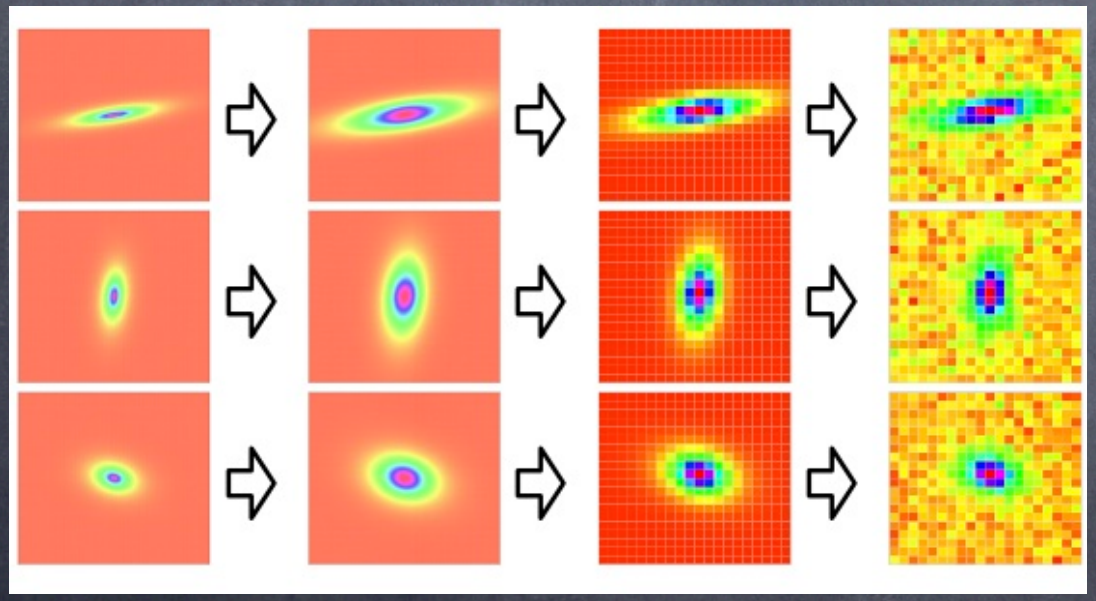


Spectral Series Regression

II. Approximate Likelihood Computation

- For some complex systems, the only real "theory" available may be in the form of a simulation model

Fig: Galaxy images generated by GalSim (blurring, pixelation, noise)



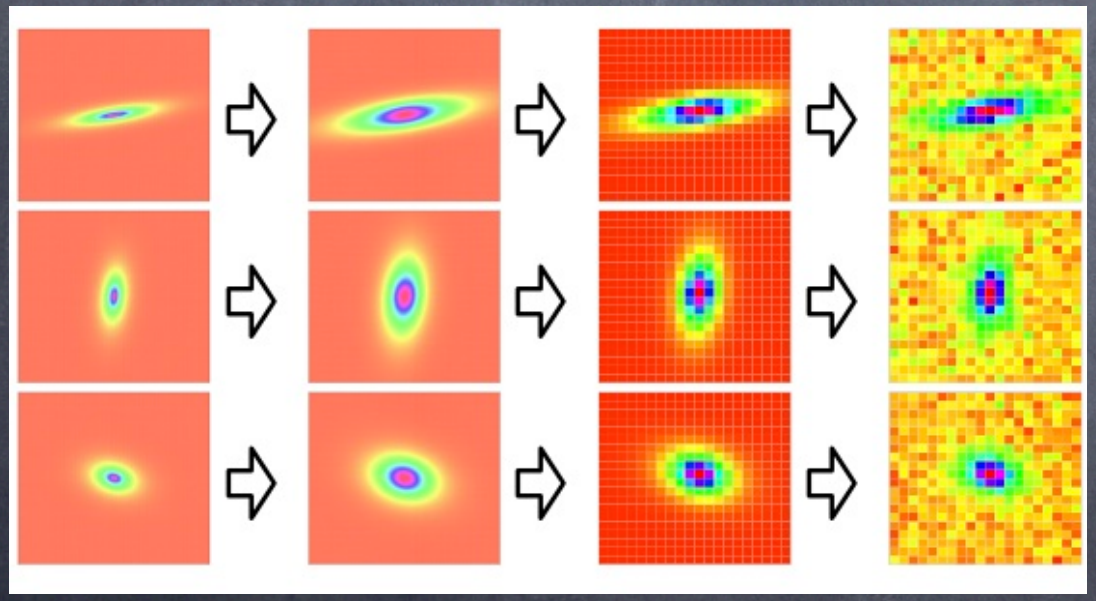
θ: shear

$\mathbf{x}$: entire image (low-res)

II. Approximate Likelihood Computation

- For some complex systems, the only real "theory" available may be in the form of a simulation model

Fig: Galaxy images generated by GalSim (blurring, pixelation, noise)



θ: shear

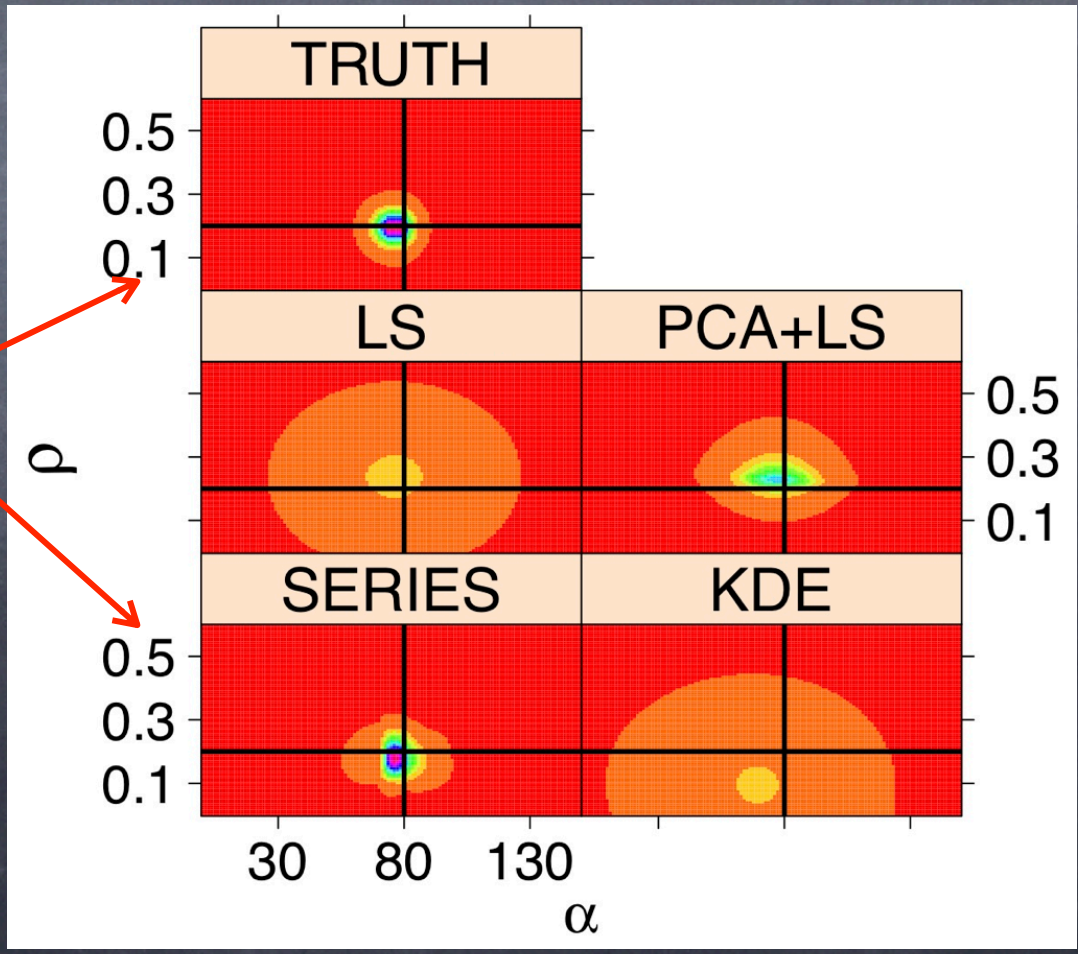
$\mathbf{x}$: entire image (low-res)

Given a simulated sample $(x_1,\theta_1),\ldots,(x_n,\theta_n)$, can estimate the likelihood function $L(\theta) \propto f(\mathbf{x}|\theta)$ nonparametrically via Spectral Series. No prior dimension reduction and no MCMC.

Approximate Likelihood Computation (cont'd)

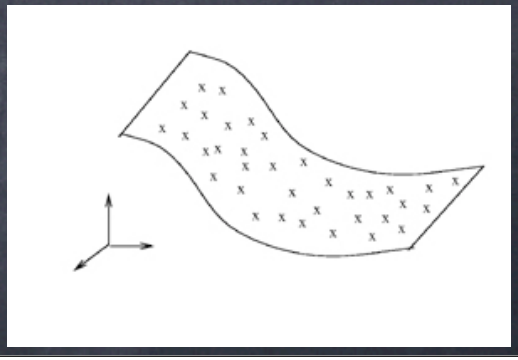
- Shear: rotation angle α, axis ratio ρ
- Contours of the estimated likelihood for different methods:

The spectral series estimator (bottom left) comes close to the true distribution (top)



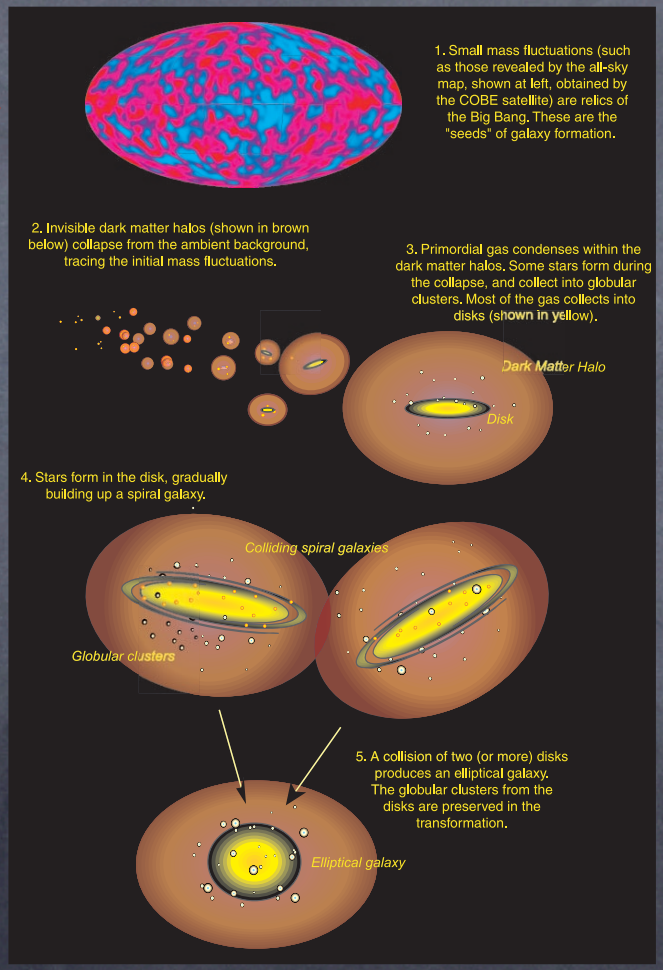
Key Points of Talk

- To take advantage of the richness of modern data and simulation models, need to work with high-dimensional data objects x and probability distributions.
- Propose a new Fourier-based approach ("Spectral Series") that
  - exploits the intrinisic geometry of the data, and
  - that can be used to estimate general functions on complex objects x; e.g. conditional densities f(z|**x**) in a regression setting, and likelihood functions f(**x**|θ),



Work in Progress/Open Problems: Calibrating Complex Models

Theory In (Simulation)



Abraham & van den Bergh (2001)

Ensemble of Galaxies Out

How, Exactly?

Observed Data

- Compare distributions in high dimensions
- Estimate parameters in the model

References

- Izbicki and Lee, ``Nonparametric Conditional Density Estimation in a High-Dimensional Regression Setting", JCGS 2015.
- Izbicki, Lee and Schafer, ``High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation", AISTATS 2014, arXiv:1404.7063
- Lee and Freeman, "Exploiting Low-Dimensional Structure in Astronomical Data for Improved Statistical Inference", SCMA V, 2011.
- Lee and Wasserman, "Spectral Connectivity Analysis", JASA 2010.
- Coifman et al, "Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data", PNAS 2005.

For preprints and upcoming papers: http:/www.stat.cmu.edu/~annlee

My email address: annlee@cmu.edu

EXTRA SLIDES START HERE

In Practice: Need to Estimate the Basis

$X_1, \ldots, X_n$: i.i.d. sample

1. Construct Gram matrix (or row-normalized matrix)

$$\begin{bmatrix} a(X_1, X_1) & a(X_1, X_2) & \cdots & a(X_1, X_n) \\ a(X_2, X_1) & a(X_2, X_2) & \cdots & a(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ a(X_n, X_1) & a(X_n, X_2) & \cdots & a(X_n, X_n) \end{bmatrix}$$

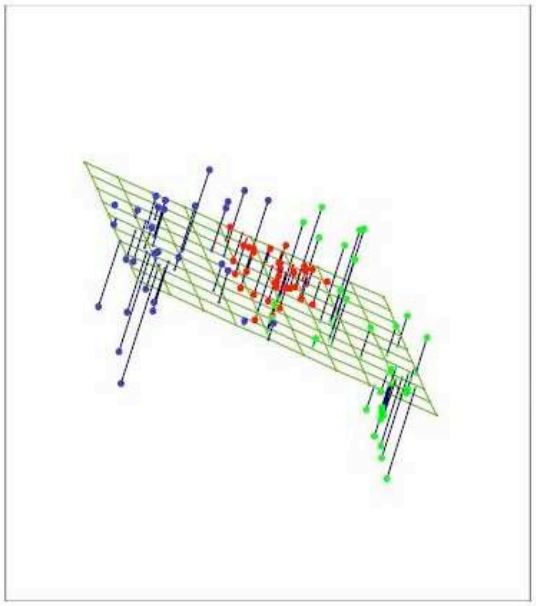
2. Compute eigenvectors (j=1,2,...)

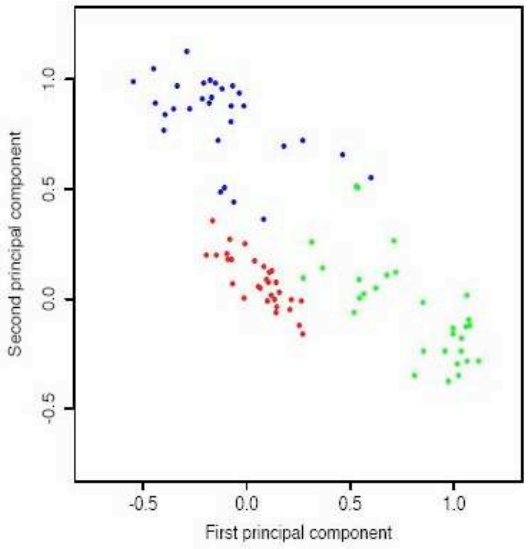
$$\left(\widetilde{\psi}_j(X_1), \ldots, \widetilde{\psi}_j(X_n)\right)$$

3. For $x \in \mathcal{X}$, "Nyström extension"

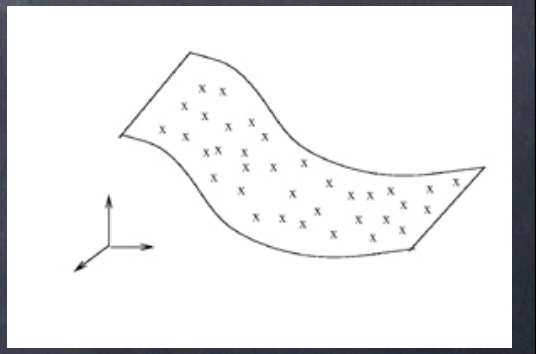
$$\widehat{\psi}_j(x) = \frac{1}{\widehat{\lambda}_j} \sum_{k=1}^{n} a(x, X_k) \widetilde{\psi}_j(X_k).$$

Recall: PCA is a linear data compression. Does not capture (nonlinear) geometries.



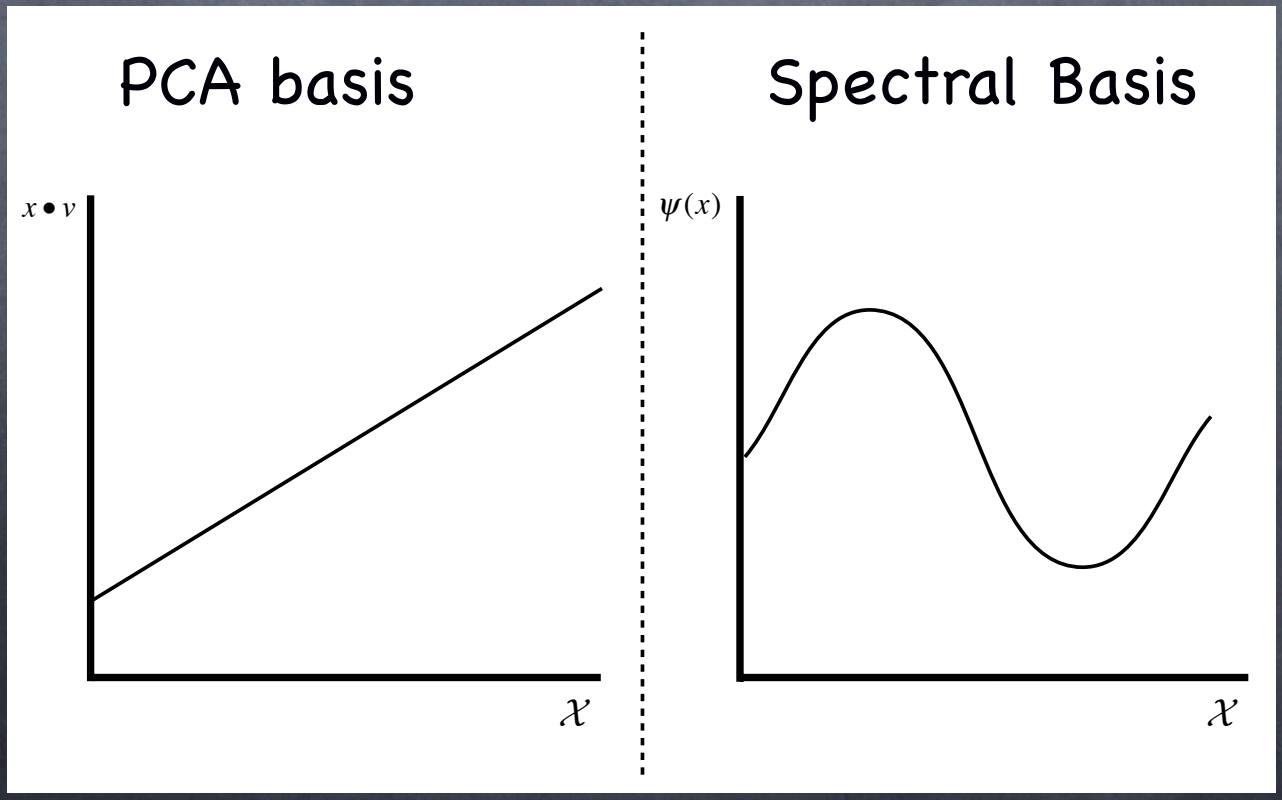


Elements of Statistical Learning,
Hastie, Tibshirani, and Friedman, pg. 488



Why Spectral Basis and Not Just PCA?

- In PCA, eigenfunctions are linear.
- For the "diffusion kernel", the eigenfunctions form a Fourier-like basis for estimating functions that are smooth relative P_X.

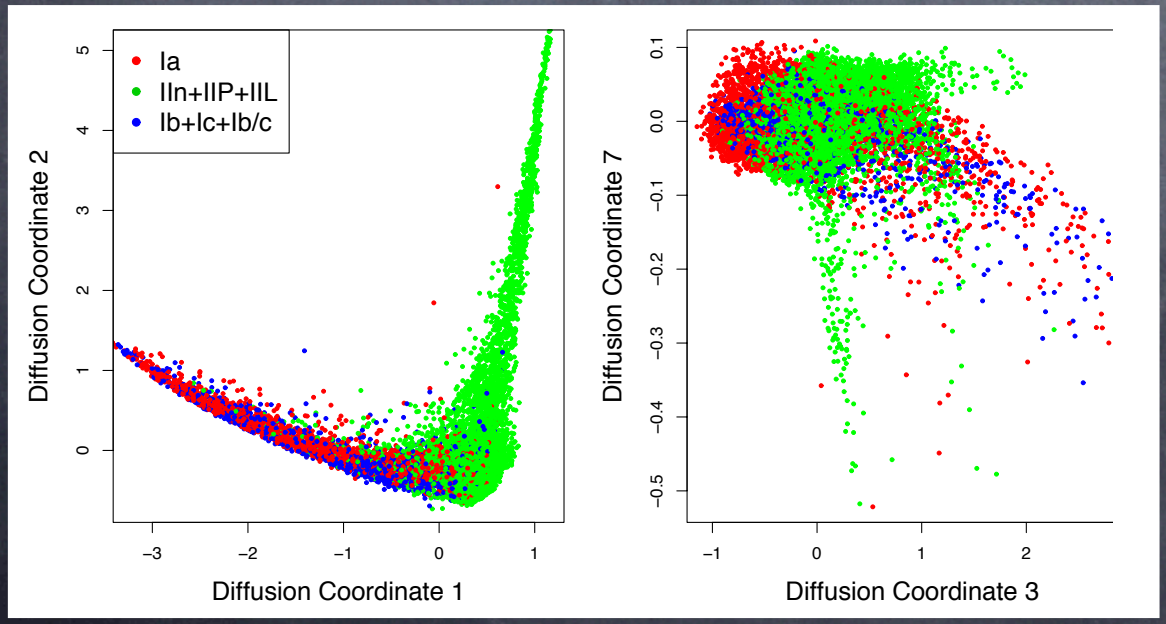


Dual Interpretation of Eigenfunctions

1. Coordinates of complex data objects

$$x \in \mathcal{X} \;\mapsto\; (\lambda_1\psi_1(x), \lambda_2\psi_2(x), \ldots)$$

- Useful for organizing and clustering data objects
- Dimension reduction (if only p<d coordinates retained)



Embedding of supernova light curves using spectral basis [Richards et al, 2011]

Dual Interpretation of Eigenfunctions

1. Coordinates of complex data objects

$$x \in \mathcal{X} \;\mapsto\; (\lambda_1 \psi_1(x), \lambda_2 \psi_2(x), \ldots)$$

- Useful for organizing and clustering data objects
- Dimension reduction (if only p<d coordinates retained)

2. Orthogonal basis for functions on the data

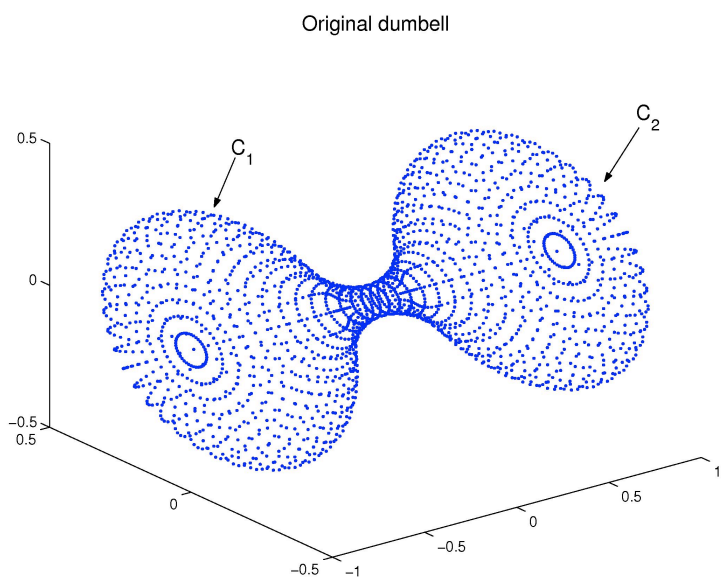
- Useful for non-parametric curve estimation (regression/classification/density estimation):

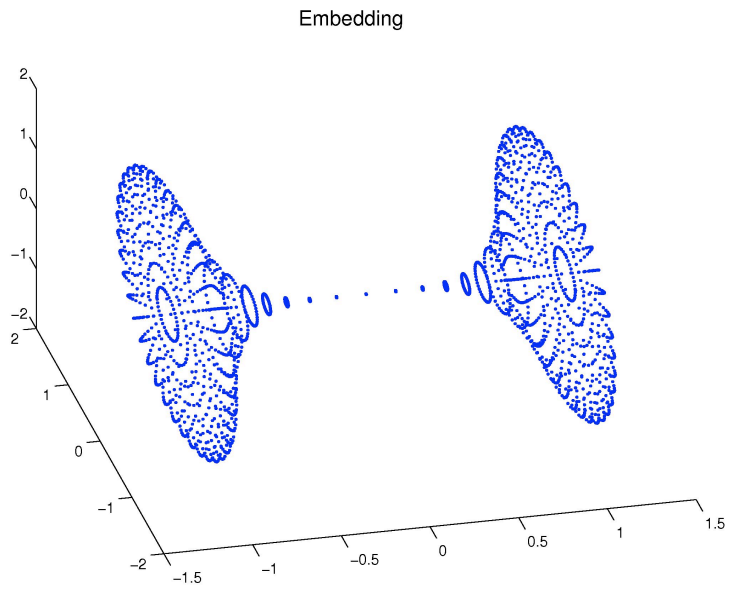
$$\text{If } f \in L^2(\mathcal{X}, P), \text{ then } f(x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x)$$
$$\text{where } \langle \psi_i, \psi_j \rangle_P = \delta_{i,j} \text{ and } \beta_j = \langle f, \psi_j \rangle_P.$$

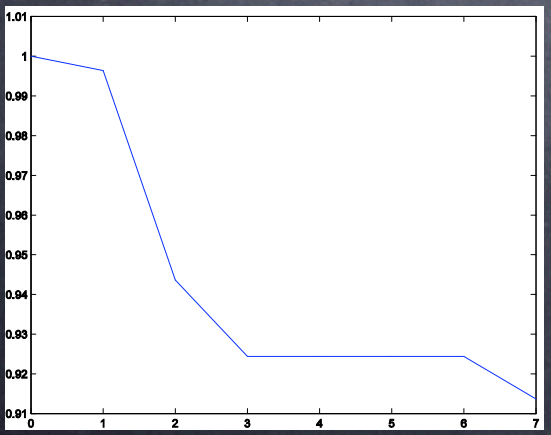
e.g. redshift

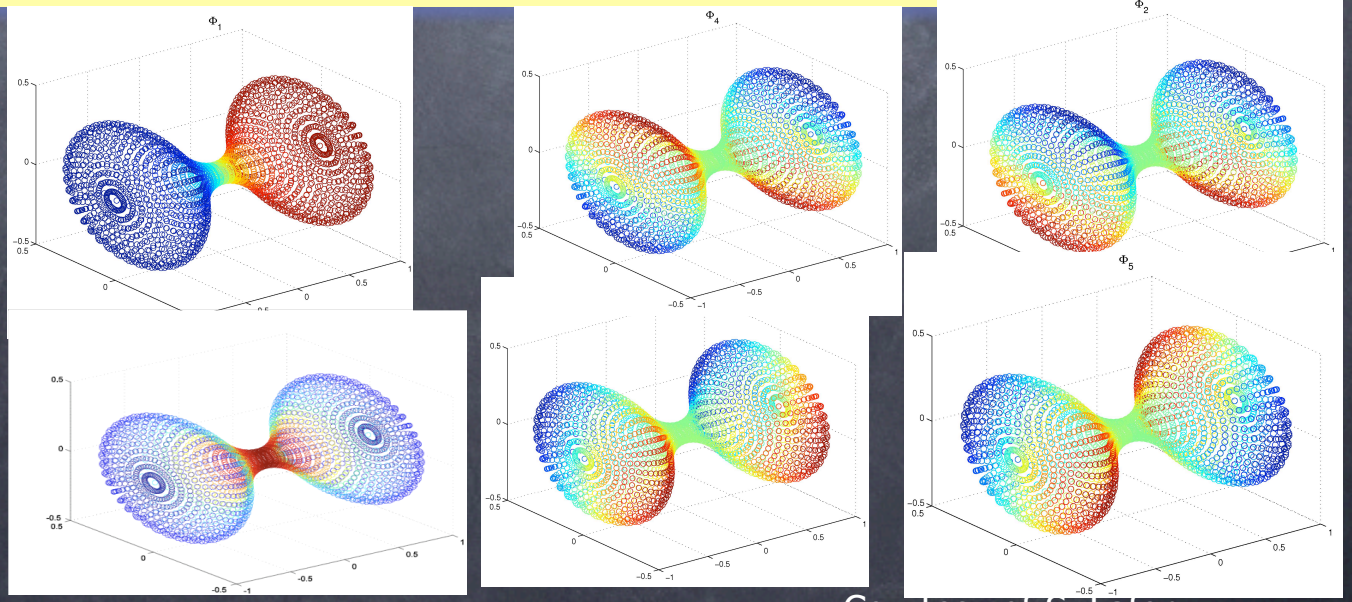
e.g. galaxy spectra

Example: Hour-glass surface





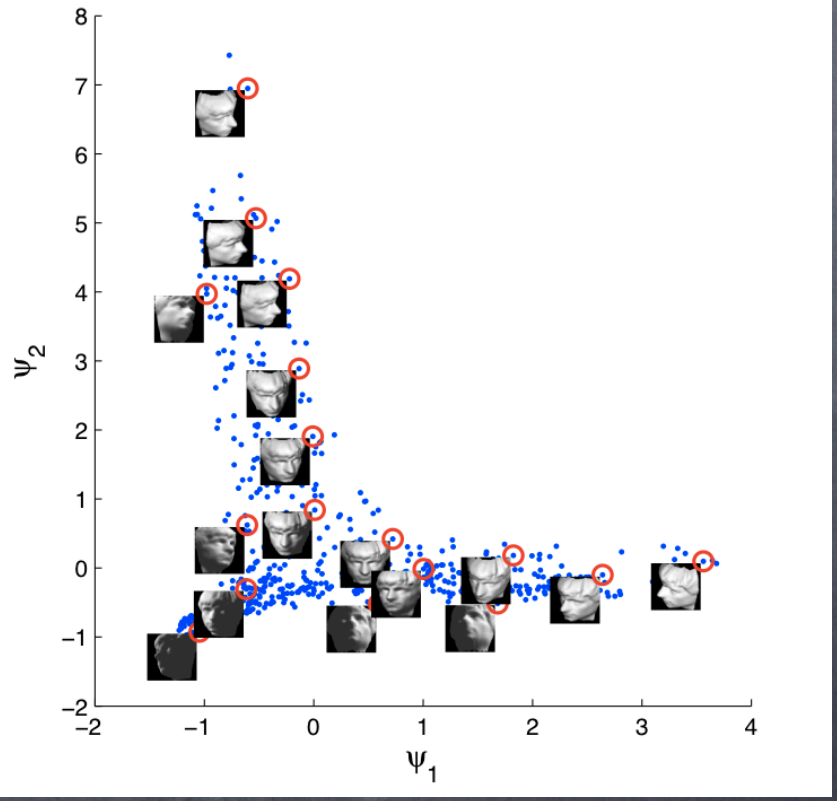




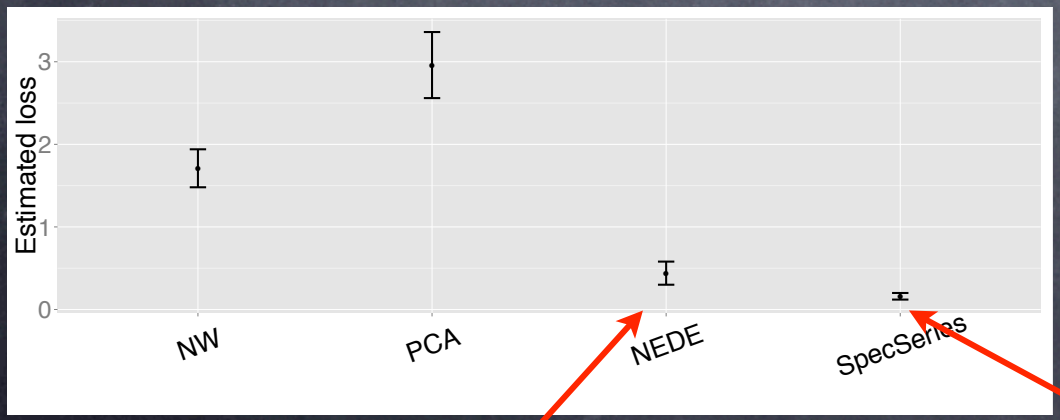
Courtesy of S. Lafon

Ex: Estimating Pose of Images of Faces

Isomap Face Database



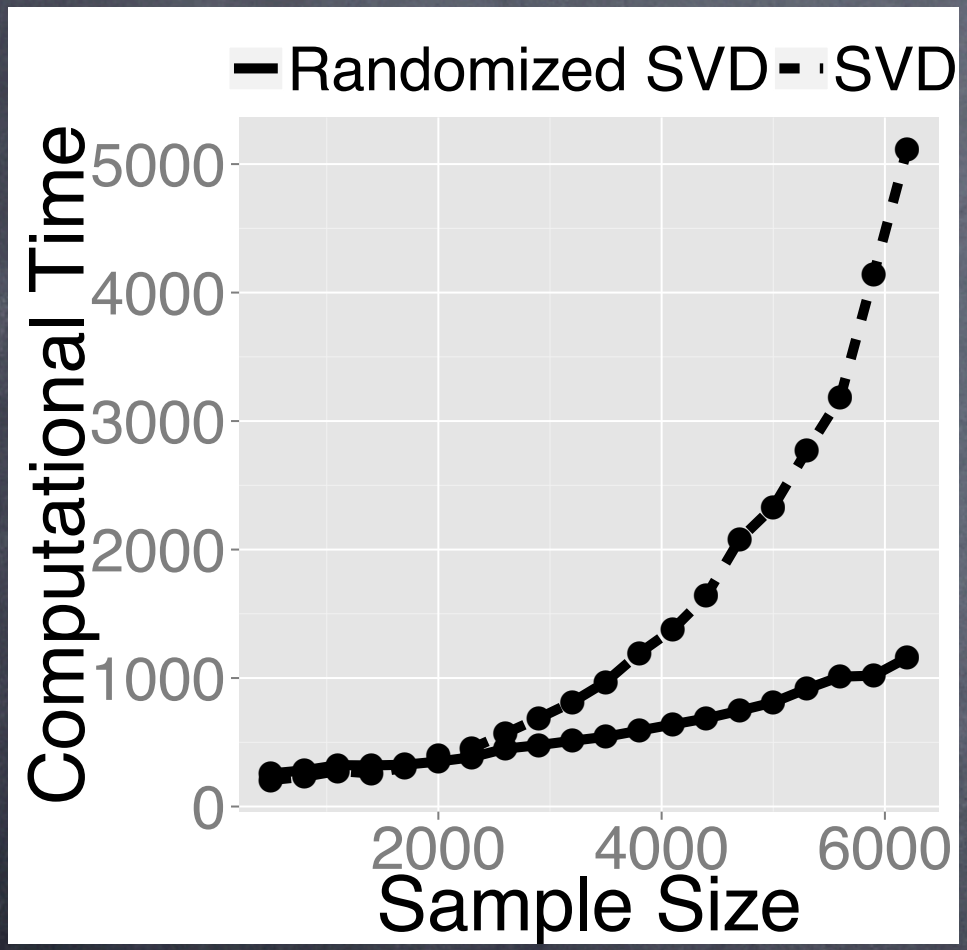
Spectral basis for visualization

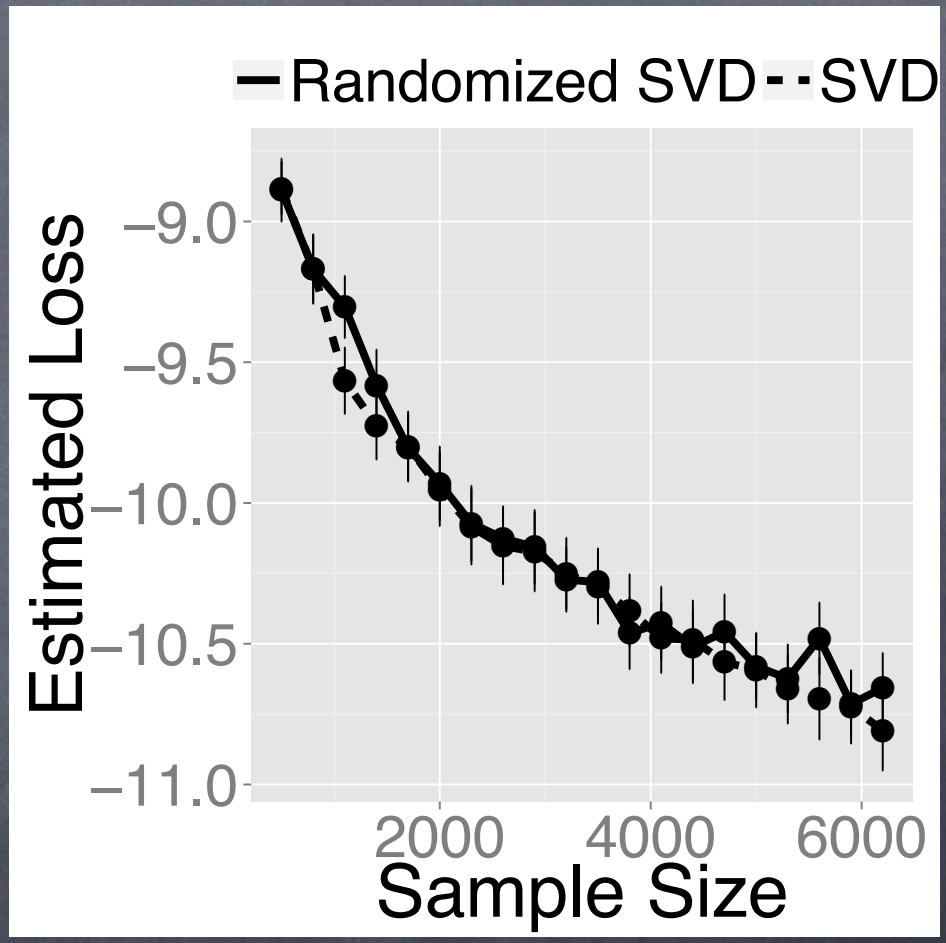


Manifold-based regression (30 mins)

Spectral Series Regression (seconds)

Scalability: Preliminary Results





Benefits of using the randomized SVD (Halko et al, 2011). Parallelizable algorithm.