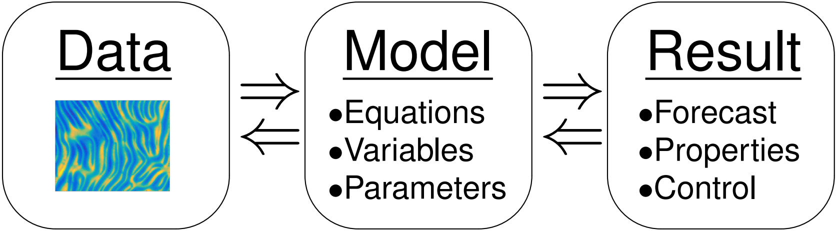
What geometries can we learn from data?

Tyrus Berry
Dept. of Mathematical Sciences, GMU

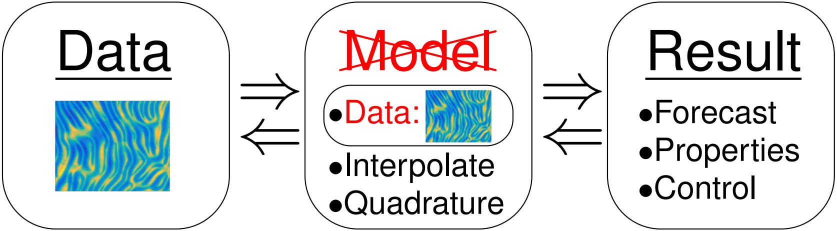
Johns Hopkins University
Nov. 15, 2017

PARAMETRIC MODELING



- **Model error:**
  - Trade off resolution and complexity
  - Stationarity/Homogeneity of parameters
- **Assimilate Data:** Fit Parameters/Variables
  - Lumps together noise and model error

NONPARAMETRIC MODELING



- **Tools:** For functions $f \in \mathcal{H}$ determined by values $\vec{f}_i = f(x_i)$
  - Interpolate: $f(x) = \sum_j \langle f, \varphi_j \rangle \varphi_j(x)$
  - Quadrature: $\langle f, \varphi_i \rangle \approx \sum_i f(x_i)\varphi(x_i)$
  - Operator Representation: $\mathbf{A}_{jk} = \langle \varphi_j, \mathcal{A}\varphi_k \rangle$
- All require a basis $\{\varphi_j\}$!

DIFFUSION FORECAST

- Autonomous SDE: $dx = a(x)\,dt + b(x)\,dW_t$
- Density solves Fokker-Planck PDE: $\frac{\partial}{\partial t} p = \mathcal{L}^* p$
- Shift map: $S(f)(x_i) = f(x_{i+1})$ estimates $\mathbb{E}[S(f)] = e^{\tau \mathcal{L}} f$

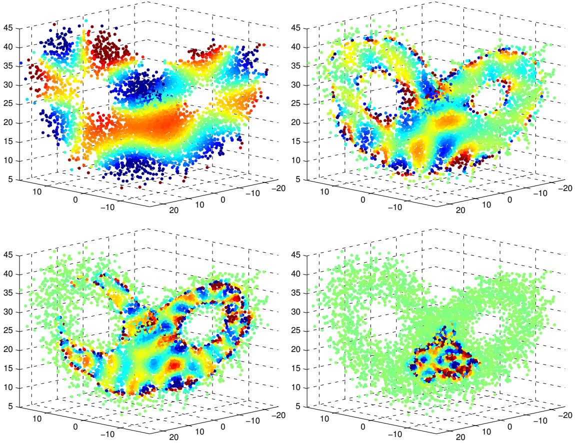
$$p(x,t) \quad \xrightarrow{\text{Diffusion Forecast}} \quad p(x,t+\tau) = e^{\tau \mathcal{L}^*} p(x,t)$$

$$\Big\downarrow \langle p, \varphi_j \rangle \qquad\qquad\qquad \Big\uparrow \textstyle\sum_j c_j \varphi_j q$$

$$\vec{c}(t) \quad \xrightarrow{A_{lj} \equiv \mathbb{E}[\langle \varphi_j, S\varphi_l \rangle_q]} \quad \vec{c}(t+\tau) = A\vec{c}(t).$$

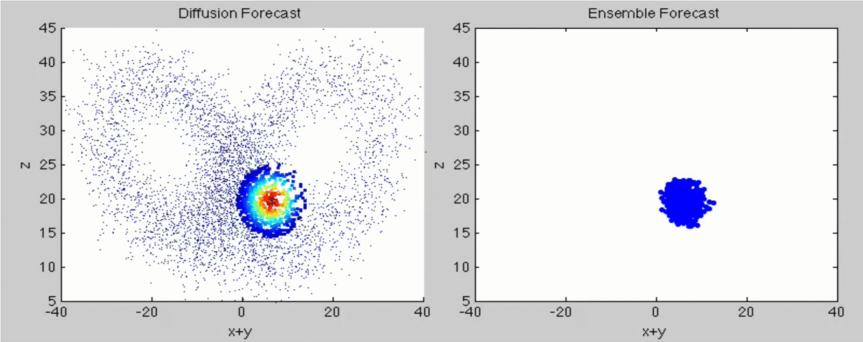
MANIFOLD LEARNING ⇒ CUSTOM 'FOURIER' BASIS

- **Optimal basis:** Minimum variance $A_{lj} \equiv \mathbb{E}[\langle \varphi_j, S\varphi_l \rangle_q]$



DIFFUSION FORECAST EXAMPLE

No Model Perfect Model



ROADMAP

- What is manifold learning? $\Rightarrow$ Estimate Laplacian, Δ
- How to find the Laplacian? $\Rightarrow$ Graph Laplacian, **L**
- Convergence $\mathbf{L} \to \Delta$ and overcoming limitations
- New graph construction based on key result (TDA)
- Future directions

WHAT IS MANIFOLD LEARNING?

- Geometric prior: Data on Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$
- **Goal**: Represent all the information about a manifold
- A smooth embedded manifold $\mathcal{M} \subset \mathbb{R}^m$ inherits:
  - A metric tensor $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ (inner product)
  - g completely determines the geometry of $\mathcal{M}$
  - A volume form $dV(x) = \sqrt{\det(g_x)}\, dx^1 \wedge \cdots \wedge dx^d$
- Laplace-Beltrami operator, Δ, is equivalent to g
  - $\Delta f = \mathrm{div} \circ \nabla = \frac{1}{\sqrt{|g|}} \partial_i g^{ij} \sqrt{|g|} \partial_j f$
  - $g(\nabla f, \nabla h) = \frac{1}{2}(f\Delta h + h\Delta f - \Delta(fh))$

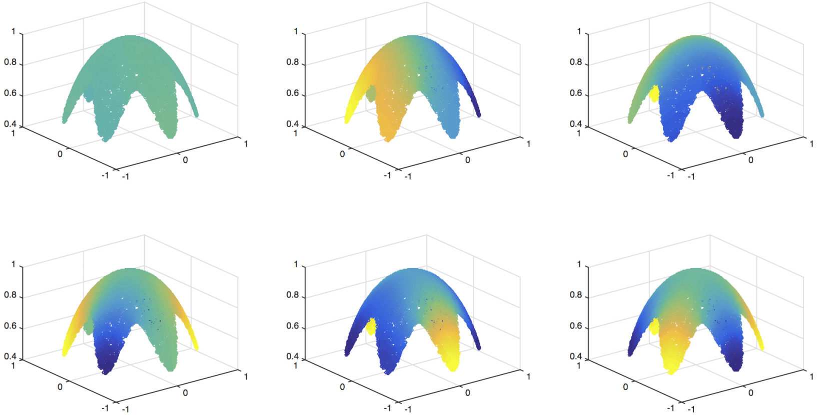
WHAT IS MANIFOLD LEARNING?

- **Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami**
- Eigenfunctions $\Delta\varphi_i = \lambda_i\varphi_i$ orthonormal basis for $L^2(\mathcal{M}, g)$
- Smoothest functions: φ_i minimizes the functional

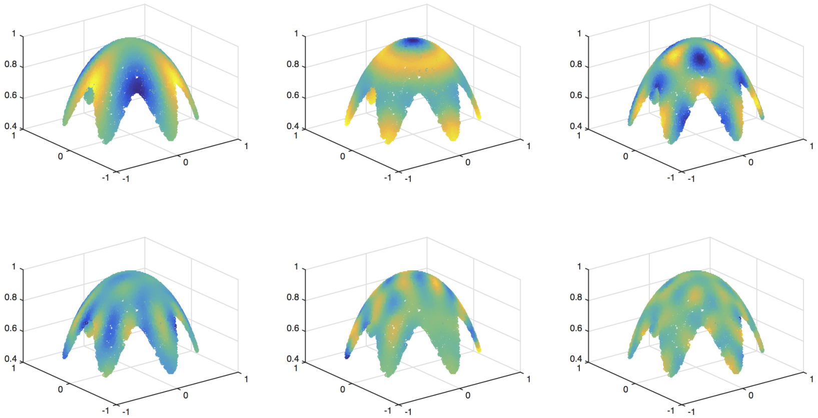
$$\lambda_i = \min_{\substack{f \perp \varphi_k \\ k=1,\ldots,i-1}} \left\{ \frac{\int_{\mathcal{M}} ||\nabla f||^2 \, dV}{\int_{\mathcal{M}} |f|^2 \, dV} \right\}$$

- Eigenfunctions of Δ are custom Fourier basis
  - Smoothest orthonormal basis for $L^2(\mathcal{M}, g)$
  - Can be used to define wavelet frame
  - Define the Sobolev spaces on $\mathcal{M}$

HARMONIC ANALYSIS ON MANIFOLDS



Harmonic Analysis on Manifolds



SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

- Data set $\Rightarrow$ *weighted graph*
- Edge Weights defined by a kernel function

$$K_\delta(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{4\delta^2}}$$

- Bandwidth δ determines localization
- 'Adjacency' matrix: $\mathbf{K}_{ij} = K(x_i, x_j)$
- 'Degree' matrix: $\mathbf{D}_{ii} = \sum_j \mathbf{K}_{ij}$
- Normalized graph Laplacian: $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{K}$

POINTWISE CONVERGENCE

Theorem: (Belkin & Niyogi, 2005, Singer, 2006)
For $\{x_i\}_{i=1}^N \subset \mathcal{M} \subset \mathbb{R}^m$ uniformly sampled on a compact manifold and for $\vec{f}_i = f(x_i)$ where $f \in C^3(\mathcal{M})$

$$\frac{1}{\delta^2}\left(\mathbf{L}\vec{f}\right)_i = \Delta f(x_i) + \mathcal{O}\left(\delta^2, \frac{1}{N^{1/2}\delta^{1+d/2}}\right)$$

$\delta =$ bandwidth
$N =$ number of points

RESTRICTIONS THAT HAVE BEEN OVERCOME:

- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

RESTRICTIONS THAT HAVE BEEN OVERCOME:

- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

DIFFUSION MAPS: ALLOWING ARBITRARY SAMPLING

- Data sampled from density $x_i \sim q$
- Matrix times vector converges to integral operator:

$$\left(\mathbf{K}\vec{f}\right)_i = \sum_{j=1}^{N} K_\delta(x_i, x_j) f(x_j) \xrightarrow{N\to\infty} \int_{\mathcal{M}} K_\delta(x_i, y) f(y) q(y)\, dV(y)$$

- 'Degree' matrix $\mathbf{D}_{ii} = \sum_j \mathbf{K}_{ij} = \sum_j K(x_i, x_j) \propto q(x_i)$
- Right normalization: $\hat{\mathbf{K}} = \mathbf{K}\mathbf{D}^{-1}$

$$\left(\hat{\mathbf{K}}\vec{f}\right)_i = \sum_{j=1}^{N} K_\delta(x_i, x_j) \frac{f(x_j)}{\mathbf{D}_{ii}} \xrightarrow{N\to\infty} \int_{\mathcal{M}} K_\delta(x_i, y) f(y)\, dV(y)$$

RESTRICTIONS THAT HAVE BEEN OVERCOME:

- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

LOCAL KERNELS

- A *local kernel* has exponential decay:

$$K_\delta(x, x + \delta z) < c_1 e^{-c_2 ||z||^2}$$

- **Theorem:** Symmetric local kernels converge to Laplacians
  - Every local kernel determines a geometry
  - Every geometry accessible by a local kernel
- Explain success of 'kernel methods' in data science:
  - KPCA: Kernel Principal Component Analysis
  - KSVM: Kernel Support Vector Machines
  - KDE: Kernel Density Estimation
  - RKHS: Reproducing Kernel Hilbert Spaces
  - Spectral Clustering (KPCA)

LOCAL KERNELS: NEW GEOMETRY

- Prototypical Kernels: For $\mathbf{A}(x) \in \mathbb{R}^{m\times m}$

$$K_\delta(x,y) = \exp\left(-\frac{1}{\delta^2}(y-x)^\top \mathbf{A}(x)^{-1}(y-x)\right)$$

- The matrix $\mathbf{A}(x)$ will change the geometry
- $\mathbf{A}(x)$ is an $m \times m$ matrix, but the metric $g_{ij}(x)$ is $d \times d$
- Restrict $\mathbf{A}(x)$ to the tangent space, let $\mathcal{I} : \mathbb{R}^m \to T_x\mathcal{M}$
- We define $\hat{\mathbf{A}}(x) = \mathcal{I}(x)\mathbf{A}(x)\mathcal{I}(x)^\top$
- The new geometry is $\hat{g}(x) = \hat{\mathbf{A}}(x)^{-1/2} g(x) \hat{\mathbf{A}}(x)^{-1/2}$

LOCAL KERNELS: INTRINSIC GEOMETRY

- For a general local kernel K_δ

$$\hat{\mathbf{A}}(x) \equiv \int_{T_x\mathcal{M}} zz^\top K_\delta(x, x+\delta z)\, dz$$

- Building $\Delta_\mathcal{N}$ where $F : \mathcal{N} \to \mathcal{M}$ is a diffeomorphism and

$$\hat{\mathbf{A}}(x) = dF(x)dF(x)^\top$$

- Assume sampling density result of conformal isometry
- For $\mathbf{A}(x) = q(x)^{2/d}\mathbf{I}$, $\mathcal{N}$ is conformal to $\mathcal{M}$ with volume

$$dV_\mathcal{N} = q\, dV_\mathcal{M}$$

- **Conformally Invariant Diffusion Map (CIDM)**: $\mathbf{A}(x) = q(x)^{2/d}$ is invariant to any conformal transformation of the data

RESTRICTIONS THAT HAVE BEEN OVERCOME:

- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

TANGIBLE MANIFOLDS

- Compute ambient distance $||x - y||_{\mathbb{R}^m}$
- Need localization in $d_{\mathcal{I}}(x,y) = \inf_\gamma \left\{ \int_0^1 |\gamma'(t)| \, dt \right\}$
- Assume ratio $R(x,y) = \frac{||x-y||_{\mathbb{R}^m}}{d_{\mathcal{I}}(x,y)}$ bounded away from zero
- We will use the exponential map to change variables
- Assume injectivity radius $\text{inj}(x)$ bounded away from zero

Definition: A manifold is uniformly tangible if there are positive lower bounds on $\text{inj}(x)$ and $\inf_{y\in\mathcal{M}} R(x,y)$ independent of x

Pointwise Consistency Part 1

- Matrix times vector converges to integral operator:

$$\left(\mathbf{K}\vec{f}\right)_i = \sum_{j=1}^{N} K_\delta(x_i, x_j) f(x_j) \xrightarrow{N\to\infty} \int_{\mathcal{M}} K_\delta(x_i, y) f(y)\, dV(y)$$

- Assume kernel has fast decay: $K_\delta(x,y) < e^{-||x-y||^2/\delta^2}$
- Localize integral, requires $R(x_i, y) = \frac{||x_i - y||}{d_I(x_i, y)} > 0$

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{\mathcal{M}\cap \exp_{x_i}(B_\delta(0))} K_\delta(x_i, y) f(y)\, dV(y) + \mathcal{O}(\delta^k)$$

- Change variables to the tangent space $y = \exp_{x_i}(s)$:

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{B_\delta(0)} K_\delta(x_i, \exp_{x_i}(s)) f(\exp_{x_i}(s))\, ds + \mathcal{O}(\delta^k)$$

- Requires injectivity radius $\mathrm{inj}(x_i) > \delta > 0$

Pointwise Consistency Part 2

- Taylor expansion in normal coordinates:

$$f(\exp_x(s)) = f(x) + \nabla f(x) \cdot s + \frac{1}{2} s^\top H(f \circ \exp_x)(0) s$$

- Symmetric kernel $\Rightarrow$ Odd terms integrate to zero

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{||s||<\delta} \left(h\left(||s||^2\right) + \mathcal{O}(\delta^2 s_i^4) h'(||s||^2)/||s||^2 \right) \cdot$$
$$\left(f(x_i) + \delta \nabla f(x_i) \cdot s + \frac{\delta^2}{2} s^\top H(f \circ \exp_{x_i})(0) s \right) \, ds + \mathcal{O}(\delta^4)$$
$$= f(x_i) + m\delta^2 (f(x_i)\omega(x) + \Delta f(x_i)) + \mathcal{O}(\delta^4)$$

- Normalize: $\mathbf{D}^{-1}\mathbf{K}\vec{f} = \frac{\mathbf{K}\vec{f}}{\mathbf{K}\vec{1}} \to \vec{f} - m\delta^2 \overrightarrow{\Delta f} + \mathcal{O}(\delta^4)$
- **Consistency:** $\frac{1}{m\delta^2}(\mathbf{I} - \mathbf{D}^{-1}\mathbf{K})\vec{f} \to \overrightarrow{\Delta f} + \mathcal{O}(\delta^2)$

CONSISTENCY IS NOT ENOUGH!

- Extend to arbitrary sampling $x_i \sim q$ (Coifman & Lafon)
- **Variance:** $\mathbb{E}[((L\vec{f})_i - \mathbf{E}[(L\vec{f})_i])^2] = \mathcal{O}\left(\frac{q(x_i)^{3-4d}}{N\delta^{2+d}}\right)$
- Negative exponent: $3 - 4d < 0$
- As density q approaches zero the variance blows up!
- **Solution:** Variable bandwidth

VARIABLE BANDWIDTH KERNELS

We introduced the variable bandwidth kernel:

$$K_{\delta,\beta}(x,y) = K\left(\frac{||x-y||}{\delta\sqrt{q(x)^{\beta}q(y)^{\beta}}}\right)$$

Theorem (Berry and Harlim, ACHA, 2015):

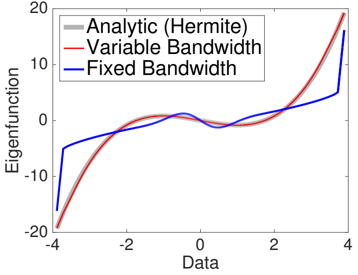
$$\mathbf{L}_{\delta,\alpha,\beta}\vec{f} = \Delta f + c_1 \nabla f \cdot \nabla \log q + \mathcal{O}\left(\delta^2, \frac{q^{-c_2}}{\sqrt{N}h^{1+d/2}}\right)$$

- Operator defined by: $c_1 = 2 - 2\alpha + d\beta + 2\beta$
- Variance determined by: $c_2 = 1/2 - 2\alpha + 2d\alpha + d\beta/2 + \beta$

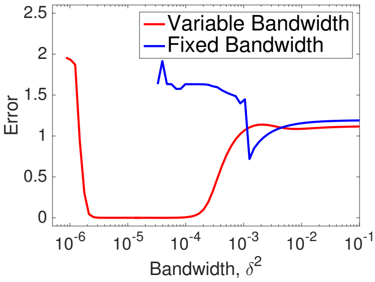
EXAMPLE: VARIABLE BANDWIDTH KERNEL

Gaussian data: Brownian motion in quadratic potential

Eigenfunctions (Hermite)



Error vs. Bandwidth



RESTRICTIONS THAT HAVE BEEN OVERCOME:

- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

SPECTRAL INGREDIENTS:

- Main Ingredient: Consistency of Spectral Clustering (Luxburg et al., Ann. Stat. 2008)
- Bias and variance, optimal bandwidth
- Proportional error in the eigenvalues
- Discrete Spectrum $\Leftarrow$ Finite volume

CONSISTENCY OF SPECTRAL CLUSTERING

- Extremely powerful theory by Luxburg et al.
- Problem: $\mathbf{L} = \mathbf{D} - \mathbf{K}$,
  - $\mathbf{K}$ represents compact integral operator, no problem
  - $\mathbf{D}\vec{f} = D(x)f(x)$ represents a multiplication operator!
- Their "Result 2" states that spectral convergence of unnormalized Laplacian requires eigenvalues no lie in the range of the degree function $D(x) = \lim_{N\to\infty} \sum_{j=1}^{N} K_\delta(x, x_j)$
- They neglect the constant c^{-1} required: $c^{-1}(\mathbf{D} - \mathbf{K}) \to \Delta$
- True degree function is $\mathcal{O}(\delta^{-2}) \to \infty$ as $N \to \infty$ and $\delta \to 0$

HOW TO CHOOSE δ?

Pointwise

$$\mathbb{E}\left[\frac{1}{c}L\right] = \Delta f + \mathcal{O}(\delta^2)$$

$$\mathrm{Var}\left[\frac{1}{c}L\right] = \mathcal{O}(N^{-1}\delta^{-m-2})$$

$$(\delta^2)^2 = N^{-1}\delta^{-m-2}$$

$$\delta^{m+6} = N^{-1}$$

$$\delta = N^{-\frac{1}{m+6}}$$

Spectral

$$\mathbb{E}\left[\frac{\vec{f}^{\,T}\frac{1}{c}L\vec{f}}{\vec{f}^{\,T}\vec{f}}\right] = \frac{\langle f, \Delta f\rangle_{L^2(\mathcal{M},g)}}{\langle f, f\rangle_{L^2(\mathcal{M},g)}} + \mathcal{O}(\delta^2, N^{-1})$$

$$\mathrm{Var}\left[\frac{\vec{f}^{\,T}\frac{1}{c}L\vec{f}}{\vec{f}^{\,T}\vec{f}}\right] = \mathcal{O}(N^{-2}\delta^{-m-2})$$

$$(\delta^2)^2 = N^{-2}\delta^{-m-2}$$

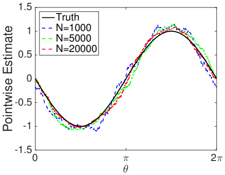
$$\delta^{m+6} = N^{-2}$$

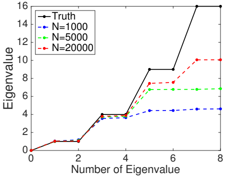
$$\delta = N^{-\frac{2}{m+6}}$$

HOW TO CHOOSE δ?

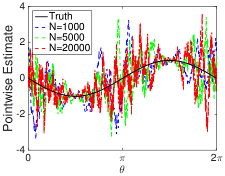
Example. $M = S^1$, $f(\theta) = \sin\theta$.

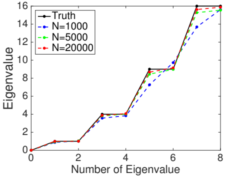
Pointwise and spectral approximations with $\delta = 3N^{-1/7}$





Pointwise and spectral approximations with $\delta = 3N^{-2/7}$

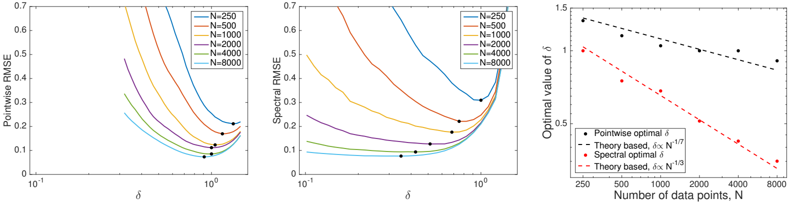




HOW TO CHOOSE δ?

Example. $M = S^1, f(\theta) = \sin\theta$.

Find minimum error over a range of δ, for several different N.



Conclusion: Optimal δ for spectral convergence is smaller (the square of) optimal δ for pointwise convergence.

RELATIVE ERROR IN THE EIGENVALUES

- Let $\hat{\lambda}$ be the estimator for λ

$$\frac{\text{var}(\hat{\lambda})}{\lambda} = c_1 \delta^{-m-2} N^{-2} + c_2 \lambda N^{-1} + h.o.t.$$

- For λ sufficiently small we have high precision
- Setting the terms equal and assuming optimal δ for spectral estimation we find

$$\lambda < \lambda_{\text{max}} = \mathcal{O}\left(N^{\frac{d-2}{d+6}} \right)$$

- Confirms observed proportional error
- Small error up to limit λ_{max}

SPECTRAL CONVERGENCE REQUIRES DISCRETE SPECTRUM

- Can approximate Δ pointwise on $\mathbb{R}$ but not spectrally
- Finite volume implies discrete spectrum
- Conformally Invariant Diffusion Map (CIDM):

$$dV_{\mathcal{N}} = q\, dV_{\mathcal{M}}$$

- Volume of $\mathcal{N} = \int_{\mathcal{N}} dV_{\mathcal{N}} = \int_{\mathcal{M}} q\, dV_{\mathcal{M}} = 1$
- What geometries can we access? Those that have finite volume!

CONTINUOUS K-NEAREST NEIGHBORS (CKNN)

- Idea: Build an unweighted graph (TDA)
- Edge between the points x, y if $\frac{||x-y||}{\sqrt{\rho(x)\rho(y)}} < \delta$
- Define the unnormalized graph Laplacian $L_{\mathrm{un}} = D - K$

$$L_{\mathrm{un}}\vec{f} \to \mathcal{L}f \equiv q\rho^{m+2}[\Delta_M f - \nabla \log(q^2\rho^{m+2}) \cdot \nabla f]$$

- Must choose $\rho = q^{-1/m}$, to get a Laplacian $\mathcal{L}f = \Delta_{M,\tilde{g}} f$
- Again find the Conformally Invariant Diffusion Map (CIDM)!

CONTINUOUS K-NEAREST NEIGHBORS (CKNN)

- Let x_k denote the kth nearest neighbor of x

CkNN: Edge between the points x, y if $\frac{||x-y||}{\sqrt{||x-x_k|| \; ||y-y_k||}} < \delta$

- For appropriate k, $||x - x_k|| \propto q(x)^{-1/d}$ so CIDM!
- This is a variable bandwidth kernel with $K(t) = 1_{\{t<1\}}$ so

$$K\left(\frac{||x-y||}{\delta\sqrt{q(x)^{-1/d}q(y)^{-1/d}}}\right) = 1_{\left\{\frac{||x-y||}{\sqrt{||x-x_k|| \, ||y-y_k||}}<\delta\right\}}$$

CONTINUOUS K-NEAREST NEIGHBORS (CKNN)

Building unweighted graphs from data (TDA)

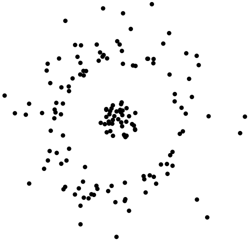
CkNN Graph: Edge $\{x, y\}$ added if $\frac{||x-y||}{\sqrt{||x-x_k||\;||y-y_k||}} < \delta$

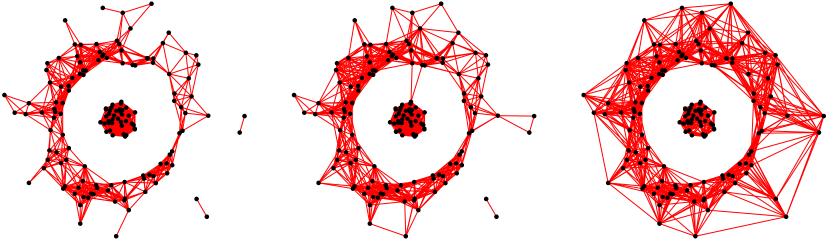
- Unit free normalization
- Distances are equalized inside the kernel (numerics)
- Consistency of CkNN clustering:
  - Conn. comp. of graph $\Leftrightarrow$ Kernel of L_{un}
  - Conn. comp. of $\mathcal{M}$ $\Leftrightarrow$ Kernel of $\Delta_{\tilde{g}}$ (Hodge theorem)

CkNN yields improved graph construction

2D Gaussian with annulus removed:

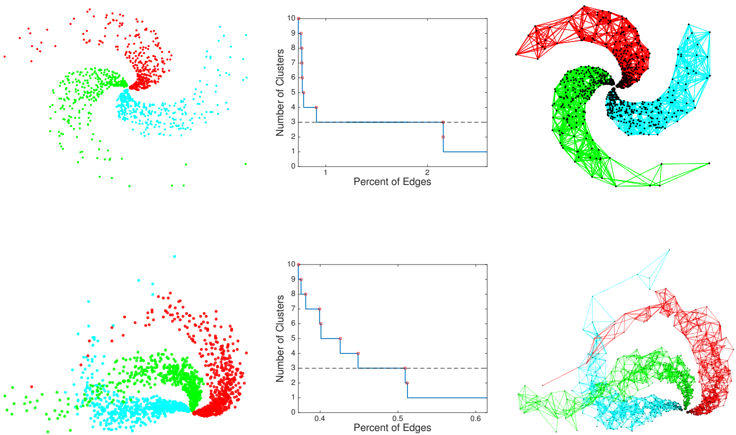
Persistent vs. consistent homology





Small bandwidth | Large bandwidth | CkNN

Improved clustering using CkNN



FUTURE DIRECTION: SMOOTHNESS PRIORS

- Manifold learning suffers from the curse-of-dimensionality
  - Bias-squared: $\mathcal{O}(\delta^4)$
  - Variance: $\mathcal{O}(N^{-1}\delta^{-2-d})$
  - Optimal bandwidth: $\delta = \mathcal{O}(N^{-1/(6+d)})$
  - Minimal Error: $\mathcal{O}(N^{-2/(6+d)})$
- Richardson Extrapolation: Combine multiple δ's
  - Reduces bias to $\mathcal{O}(\delta^{2k})$
  - Increases variance by a constant
  - Requires $\mathcal{M}$ to be C^k
- 'Solves' curse-of-dimensionality by assuming smoothness

- 5000 points
- 10-dim torus
- In $\mathbb{R}^{20}$