The interplay of mathematics and statistics in manifold learning

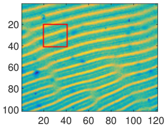
Tyrus Berry
Dept. of Mathematical Sciences, GMU

Feb. 15, 2019

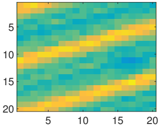
ANALYSIS OF POINT CLOUDS

- Data lie in $\mathbb{R}^m$ for large $m \Rightarrow$ Curse-of-dimensionality
- Often data are sampled from measures that are singular are nearly singular
- Geometric prior: Points lie near smooth manifold $\mathcal{M} \subset \mathbb{R}^m$
- Curse depends on the dimension $d < m$ of $\mathcal{M}$
- Goal: Learn/represent $\mathcal{M}$ with statistical error bounds

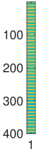
FINDING HIDDEN STRUCTURE IN DATA



$\Longrightarrow$

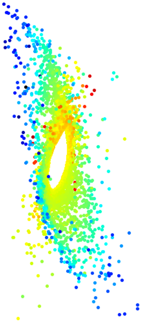


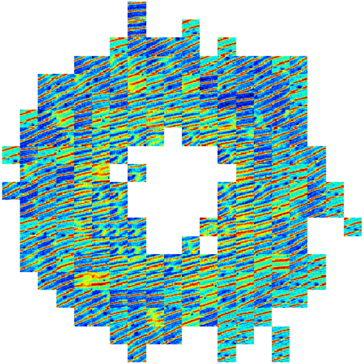
$\Longrightarrow$



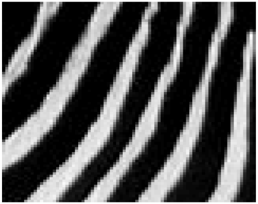
$\in \mathbb{R}^{400}$

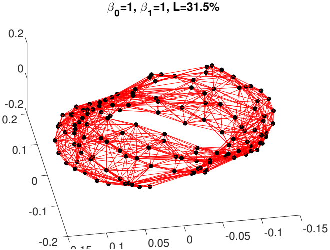
The sub-image geometry:

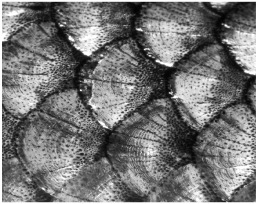


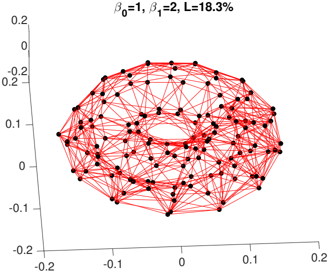


EXAMPLE: MANIFOLDS OF SUBIMAGES









DENSITY ESTIMATION ON $\mathbb{R}^m$

- Goal: Estimate density $p(x)$ from random variables $X_i \sim p$
- Kernel density estimation on $\mathbb{R}^m$ dates from the 1950's

$$p_{h,N}(x) \equiv \frac{1}{m_0 h^m N} \sum_{i=1}^{N} K\left(\frac{||x - X_i||}{h}\right) \qquad m_0 = \int_{\mathbb{R}^m} K(||z||)\, dz$$

- Theorem: $p_{h,N}(x)$ is a consistent estimator of $p(x)$ with
- Bias: $\mathbb{E}\left[p_{h,N}(x) - p(x)\right] = \mathcal{O}(h^2)$ and
- Variance: $\mathbb{E}\left[(p_{h,N}(x) - p(x))^2\right] = \mathcal{O}\left(\frac{h^{-m}}{N} p(x)\right)$.

DENSITY ESTIMATION ON $\mathbb{R}^m$

Finding the bias:

$$\mathbb{E}[p_{h,N}(x)] = \frac{1}{m_0 h^m N}\mathbb{E}\left[\sum_{i=1}^{N} K\left(\frac{||x-X_i||}{h}\right)\right]$$

$$= \frac{1}{m_0 h^m}\int_{\mathbb{R}^m} K\left(\frac{||x-y||}{h}\right) p(y)\,dy$$

$$(\text{decay of } K) = \frac{1}{m_0 h^m}\int_{||x-y||<h^\alpha} K\left(\frac{||x-y||}{h}\right) p(y)\,dy$$

$$\left(z=\frac{y-x}{h}\right) = \frac{1}{m_0}\int_{||z||<h^{\alpha-1}} K(||z||)\, p(x+hz)\,dz$$

$$(\text{Taylor}) = \frac{1}{m_0}\int_{||z||<h^{\alpha-1}} K(||z||)\left(p(x)+Dp(x)hz+\mathcal{O}(h^2||z||^2)\right) dz$$

$$(\text{symmetry}) = \frac{1}{m_0}\int_{||z||<h^{\alpha-1}} K(||z||)\left(p(x)+\mathcal{O}(h^2||z||^2)\right) dz$$

$$(\alpha<1) = p(x)\frac{1}{m_0}\int_{\mathbb{R}^m} K(||z||)\,dz + \mathcal{O}(h^2) = p(x)+\mathcal{O}(h^2)$$

RANDOM VARIABLES ON MANIFOLDS

- A smooth embedded manifold $\mathcal{M} \subset \mathbb{R}^m$ inherits:
  - A Riemannian metric $g_x(v, w)$ (defines geometry)
  - A volume form $dV(x) = \sqrt{\det(g_x)}$
  - $\text{vol}(\mathcal{M}) = \int_{x\in\mathcal{M}} 1\, dV(x)$
- Data X_i are sampled from p supported on $\mathcal{M} \subset \mathbb{R}^m$
- Expectation integrates over a singular density:

$$\mathbb{E}_p[K(||x - X_i||)] = \int_{\mathbb{R}^m} K(||x-y||)dp(y) = \int_{y\in\mathcal{M}} K(||x-y||)q(y)\, dV(y)$$

- On manifolds we will estimate $q(x)$ where $q\, dV = dp$

DENSITY ESTIMATION ON MANIFOLDS $\mathcal{M} \subset \mathbb{R}^m$

- For data X_i sampled from $q\, dV$ on $\mathcal{M}$
- Use a kernel density estimator (KDE) of form

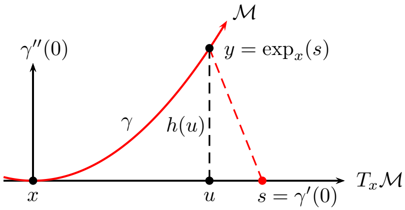
$$q_{h,N}(x) \equiv \frac{1}{N m_0 h^d} \sum_{i=1}^{N} K\left(\frac{||x - X_i||_{\mathbb{R}^n}}{h}\right), \qquad m_0 = \int_{\mathbb{R}^d} K(||z||)\, dz$$

- The expectation of the estimator is:

$$\mathbb{E}[q_{h,N}(x)] = \frac{1}{m_0 h^d N} \mathbb{E}\left[\sum_{i=1}^{N} K\left(\frac{||x - X_i||}{h}\right)\right] = \frac{1}{m_0 h^d} \int_{y \in \mathcal{M}} K\left(\frac{||x - y||}{h}\right) q(y)\, dV$$

INTEGRATING ON MANIFOLDS

The exponential map changes variables to tangent space:



- Distances preserved: $||y - x||^2 = ||s||^2 + \mathcal{O}(s_i^4)$
- Natural volume element: $dV(y) = ds$
- Taylor: $q(\exp_x(s)) = q(x) + \nabla q(x) \cdot s + \mathcal{O}(s_i^2)$

DENSITY ESTIMATION ON MANIFOLD $\mathcal{M} \subset \mathbb{R}^m$

Let $\exp_x : T_x\mathcal{M} \to U \subset \mathcal{M}$

$$\mathbb{E}[q_{h,N}(x)] = \frac{1}{m_0 h^d} \int_{||x-y||<h^\alpha} K\left(\frac{||x-y||}{h}\right) q(y)\, dV(y)$$

$$(y = \exp_x(hs)) = \frac{1}{m_0 h^d} \int_{||s||<h^\alpha} K\left(\sqrt{||s||^2 + \mathcal{O}(h^2 s_i^4)}\right) q(\exp_x(s))\, ds$$

$$(\text{Taylor}) = \frac{1}{m_0} \int_{||s||<h^{\alpha-1}} \left(K(||s||) + \mathcal{O}(h^2 s_i^4) K'(||s||)/||s||\right) \cdot \left(q(x) + \nabla q(x) hs + \mathcal{O}(h^2 ||s||^2)\right) ds$$

$$(\text{symmetry}) = \frac{1}{m_0} \int_{||z||<h^{\alpha-1}} K(||s||)\, q(x) + \mathcal{O}(h^2 ||s||^2)\, ds$$

$$(\alpha < 1) = q(x) \frac{1}{m_0} \int_{\mathbb{R}^d} K(||s||)\, ds + \mathcal{O}(h^2) = q(x) + \mathcal{O}(h^2)$$

DENSITY ESTIMATION ON MANIFOLDS $\mathcal{M} \subset \mathbb{R}^m$

Using the KDE:

$$q_{h,N}(x) \equiv \frac{1}{N m_0 h^d} \sum_{i=1}^{N} K\left(\frac{||x - X_i||_{\mathbb{R}^m}}{h}\right), \qquad m_0 = \int_{\mathbb{R}^d} K\left(||z||\right) dz$$

For a d-dimensional manifold without boundary:

Theorem. $q_{h,N}(x)$ is a consistent estimator of $q(x)$ with

- Bias: $\mathbb{E}\left[q_{h,N}(x) - q(x)\right] = \mathcal{O}(h^2)$
- Variance: $\mathbb{E}\left[(q_{h,N}(x) - q(x))^2\right] = \mathcal{O}\left(\frac{h^{-d}}{N} q(x)\right)$

Pelletier (2005), Hein (2006), Ozakin-Gray (2009), Kim-Park (2013)

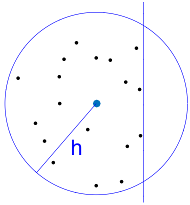
Boundary KDE

We saw that KDE extends to manifolds:

$$q_{h,N}(x) \equiv \frac{1}{N m_0 h^d} \sum_{i=1}^{N} K\left(\frac{||x - X_i||}{h}\right)$$

What about manifolds with boundary?

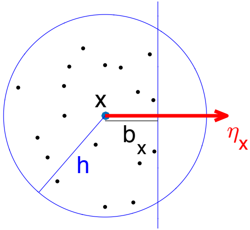
- KDE is not consistent at the boundary - normalization is wrong.
- Previous work only in $\mathbb{R}, \mathbb{R}^2$ and assumes known boundary
- Unknown boundary ???



PROBLEM NEAR THE BOUNDARY

Suppose we can locate the boundary

and b_x is distance from x to boundary



Integration is truncated in direction of boundary, η_x

$$m_0^\partial(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K\left(\sqrt{||z_\perp||^2 + z_{||}^2}\right) dz_{||} dz_\perp$$

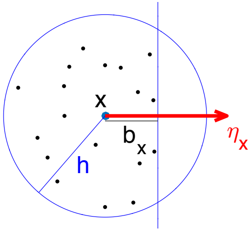
If K is Gaussian kernel then:

$$m_0^\partial(x) = (1 + \mathrm{erf}(b_x/h))/2$$

LOCATING THE BOUNDARY

Boundary Direction Estimator (BDE)

$$\mu(x) \equiv \frac{1}{Nh^{m+1}} \sum_{i=1}^{N} K\left(\frac{||x - X_i||}{h}\right)(X_i - x)$$



We prove:

$$\mathbb{E}[\mu(x)] = -\vec{\eta}_x m_1^{\partial}(x) q(x) + \mathcal{O}(h)$$

If K is Gaussian then:

$$m_1^{\partial}(x) = \frac{\exp(-b_x^2/h^2)}{2\sqrt{\pi}}$$

FINDING b_x

First, compute the standard KDE

$$\mathbb{E}[q_{h,N}(x)] = m_0^{\partial}(x)q(x) + \mathcal{O}(h)$$

Next, compute the Boundary Direction Estimator (BDE)

$$\mathbb{E}[\mu(x)] = -\vec{\eta}_x m_1^{\partial}(x)q(x) + \mathcal{O}(h)$$

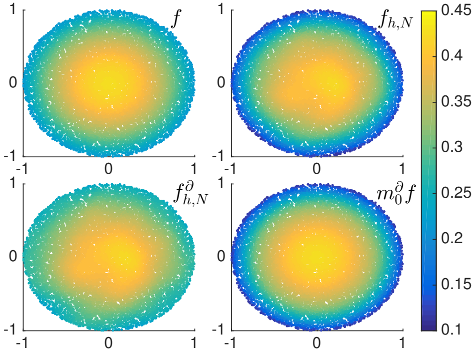
Divide to cancel the unknown $q(x)$:

$$\mathbb{E}\left[\frac{q_{h,N}(x)}{||\mu(x)||}\right] = \frac{\sqrt{\pi}(1+\mathrm{erf}\,(b_x/h))}{\exp(-b_x{}^2/h^2)} + \mathcal{O}(h)$$

Use Newton's method to solve for b_x

Now we can compute the correct normalization!

BOUNDARY KDE ON DISC



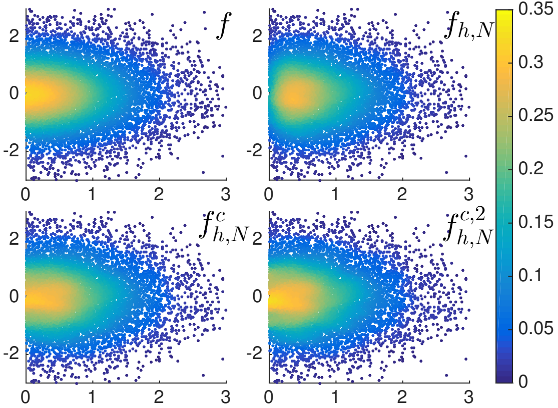
$f(x) =$ example distribution

$f_{h,N}(x) =$ standard KDE

$f^{\partial}_{h,N}(x) =$ corrected KDE using $m_0^{\partial}(x)$

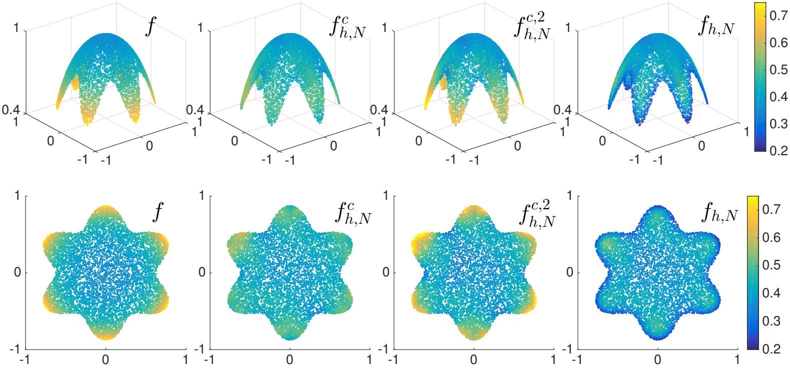
BOUNDARY KDE ON HALF GAUSSIAN

We also derive higher order estimator $f_{h,N}^{c,2}$ (extrapolation).



Density estimation on manifolds

Another example



MANIFOLD LEARNING

- **Goal**: Represent all the information about a manifold
- Riemannian metric, g, contains all geometric information
- Laplace-Beltrami operator, Δ, is equivalent to g
- Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami
- **Caveat:** Cannot easily answer all questions about manifold

What is Manifold Learning?

- **Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami**
- **Hodge theorem:**
  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 $\mathcal{M}$

MANIFOLD LEARNING VIA OPERATOR ESTIMATION

- Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami
- To find Δ we dig deeper into the Taylor expansion:

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

- Integrate over a symmetric region and odd terms drop out

$$\int_{||s||<h^{\gamma-1}} K(||s||) f(\exp_x(hs))\, ds = m_0 f(x) + \frac{h^2}{2} m_2 \Delta f(x) + \mathcal{O}(h^4)$$

- Where $m_2 = \int_{\mathbb{R}^d} K(||s||) s_i^2 \, ds$

MANIFOLD LEARNING VIA OPERATOR ESTIMATION

- Assume uniform sampling $q(x) = \frac{1}{\text{vol}(\mathcal{M})}$
- Apply kernel operator $Kf(x) = \frac{\text{vol}(\mathcal{M})}{m_0 h^d N} \sum_{i=1}^{N} K(x, X_i) f(X_i)$

$$\mathbb{E}[Kf(x)] = \frac{\text{vol}(\mathcal{M})}{m_0} \int_{||s||<h^{\alpha-1}} \left(K\left(||s||\right) + \mathcal{O}(h^2 s_i^4) K'(||s||)/||s|| \right) \cdot \left(f(x) + \nabla f(x) hs + \frac{h^2}{2} s^\top H(\tilde{f})(0)s \right) ds$$

$$= f(x) + mh^2 (f(x)\omega(x) + \Delta f(x)) + \mathcal{O}(h^4)$$

- Where $\omega(x) = \int \mathcal{O}(s_i^4) K'(||s||)/||s|| \, ds$ and $m \equiv \frac{m_2}{2m_0}$

MANIFOLD LEARNING VIA OPERATOR ESTIMATION

- The $\omega(x)$ term is a problem:

$$\mathbb{E}[Kf(x)] = f(x) + mh^2(f(x)\omega(x) + \Delta f(x)) + \mathcal{O}(h^4)$$

- Set $D(x) = K1(x)$ so $\mathbb{E}[D(x)] = 1 + mh^2\omega(x) + \mathcal{O}(h^4)$
- Normalize:

$$\mathbb{E}\left[D^{-1}Kf\right] = f(x) + \frac{h^2 m_2}{2m_0}\Delta f(x) + \mathcal{O}(h^4)$$

- Define: $\frac{1}{mh^2}\left(D^{-1}K - I\right) f = \Delta f + \mathcal{O}(h^2)$

MATRICES AS INTEGRAL OPERATORS

- Functions are represented as vectors $\vec{f}_i = f(x_i)$
- A kernel matrix $K_{ij} = K(x_i, x_j)$ represents an operator

$$\frac{1}{N}\left(K\vec{f}\right)_i = \frac{1}{N}\sum_j K(x_i, x_j)f(x_j) \to \int_{\mathcal{M}} K(x_i, y)f(y)q(y)\,dV(y)$$

- Diagonal matrix: $D_{ii} = N^{-1}\sum_j K_{ij} = N^{-1}K\vec{1}$
- Discrete Laplacian matrix: $L = \frac{1}{mh^2}\left(D^{-1}K - I\right)$
- Then $\left(L\vec{f}\right)_i = \Delta f(x_i) + \mathcal{O}(h^2)$
- This says that L is a pointwise consistent estimator of Δ

Belkin and Niyogi (2005)

Quick Review

- Kernel matrices estimate operators
- Δ appears in the bias term of KDE
- $L = \frac{1}{mh^2}\left(D^{-1}K - I\right)$ is pointwise consistent
- Notice $D^{-1}K \Rightarrow$ No need to know dimension/volume
- So far we have assumed uniform sampling
- We have not considered the variance of this estimator

DIFFUSION MAPS: ALLOWING ARBITRARY SAMPLING

- For $X_i \sim q$
- Define $K_{ij} = K\left(\frac{||x_i - x_j||}{h}\right)$ and $D_i = \sum_j K_{ij}$
- Right normalization: $\hat{K}_{ij} = K_{ij} D_j^{-1}$ and $\hat{D}_i = \sum_j \hat{K}_{ij}$
- Left normalization: $\tilde{K}_{ij} = \hat{D}_i^{-1} \hat{K}_{ij}$ and finally $L = \frac{\tilde{K} - I}{mh^2}$
- **Theorem:** L is a consistent pointwise estimator of Δ
- **Bias:** $\mathbb{E}[(L\vec{f})_i - \Delta f(x_i)] = \mathcal{O}(h^2)$
- **Variance:** $\mathbb{E}[((L\vec{f})_i - \Delta f(x_i))^2] = \mathcal{O}\left(\frac{||\nabla f(x_i)||^2 q(x_i)^{3-4d}}{N^{1/2} h^{2+d}}\right)$

Coifman and Lafon (2006), Singer (2006), Berry and Harlim (2015)

HUGE PROBLEM!

- **Variance:** $\mathbb{E}[((L\vec{f})_i - \Delta f(x_i))^2] = \mathcal{O}\left(\frac{||\nabla f(x_i)||^2 q(x_i)^{3-4d}}{Nh^{2+d}}\right)$
- Notice the exponent $3 - 4d < 0$ in $q(x_i)^{3-4d}$
- If q is not bounded away from zero the variance blows up!
- **Solution:** Variable bandwidth (balloon estimator)
- For KDE variable bandwidth has little advantage:

"Nevertheless, for densities of the usual shapes, balloon estimators are mostly hot air."

-David W. Scott

- For Operator estimation variable bandwidth is important

Berry and Harlim (2015)

Variable Bandwidth Kernels

Repeating the Diffusion Maps construction with the kernel:

$$K_{h,\beta}(x,y) = h\left(\frac{||x-y||}{hq(x)^{\beta/2}q(y)^{\beta/2}}\right)$$

We find:

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

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

Berry and Harlim (2015)

SUMMARY OF OPERATOR ESTIMATION

- Kernel matrices estimate operators
- Diffusion maps normalized Laplacian $L = \frac{1}{mh^2}\left(\hat{D}^{-1}\hat{K} - I\right)$ is pointwise consistent for any sampling density
- Variance is unbounded as density decreases
- More data $\Rightarrow$ More points with low density $\Rightarrow$ Higher error!
- Variable bandwidth gives control over variance

CONTINUOUS K-NEAREST NEIGHBORS

- Let x_k denote the kth nearest neighbor of x
- Construct graph by:

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

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

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

CkNN Convergence result

- For the CkNN kernel K define:
- The unnormalized Laplacian matrix $L_{\mathrm{un}} = D - K$ so

$$\frac{2}{m_2 h^{d+2}} L_{\mathrm{un}} \to \Delta_{\tilde{g}} f = q^{-2/d}(\Delta f - (d-2)\nabla \log(q^{1/d}) \cdot \nabla f)$$

- $\tilde{g} \equiv q^{2/d} g$ is a conformal change of metric on $\mathcal{M}$
- Volume form: $d\tilde{V} = \sqrt{|\tilde{g}|} = \sqrt{|q^{2/d} g|} = q\sqrt{|g|} = q\, dV$
- Thus, L_{un} converges to the Laplace-Beltrami operator on $\mathcal{M}$ with respect to a different metric

IS $\beta = -1/d$ A NATURAL CHOICE?

1. $L_{\rm un}$ yeilds a Laplace-Beltrami operator only if $\beta = -1/d$
2. The volume element is the sampling measure $d\tilde{V} = q\, dV$
3. The dot product is consistent with the volume form:

$$\frac{1}{N}\left[\vec{f}\cdot\vec{f}\right] = \frac{1}{N}\left[\sum_{i=1}^{N} f(x_i)^2\right] = \int f(x)^2 q(x)\, dV = \langle f, f\rangle_{d\tilde{V}}$$

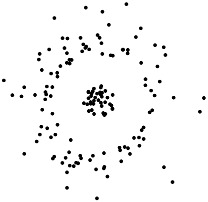
   This consistency is crucial since the discrete spectrum of $L_{\rm un}$ are the minimizers of the functional

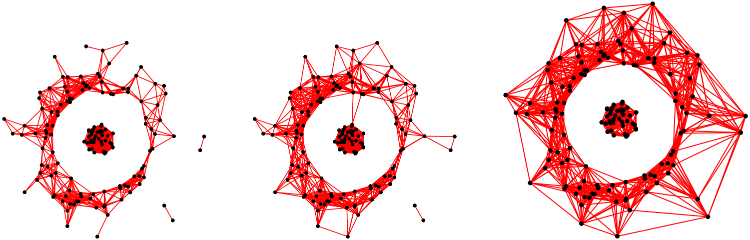
$$\Lambda(f) = \frac{\vec{f}^\top c^{-1} L_{\rm un}\vec{f}}{\vec{f}^\top \vec{f}} \to \frac{\left\langle f, \Delta_{\tilde{g}} f\right\rangle_{d\tilde{V}}}{\langle f, f\rangle_{d\tilde{V}}}$$

4. By using CkNN we never need to estimate d!

BETTER GRAPHS THROUGH STATISTICS

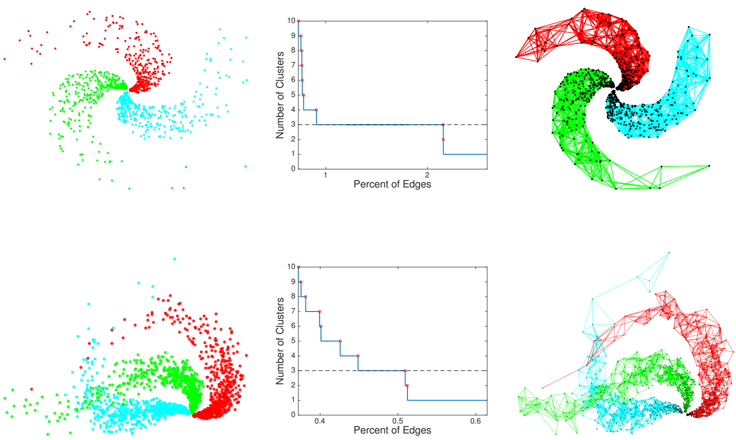
2D Gaussian with a annulus removed:





Small bandwidth Large bandwidth CkNN

IMPROVED CLUSTERING USING CKNN



SUMMARY

Statistics play a crucial role in manifold learning

- Kernel density estimation can be done on Riemannian manifolds with boundary
- Kernel matrices estimate geometric operators which contain all information about a manifold
- Statistics tell us when the algorithms will actually work!
- Balloon estimators from KDE improve operator estimation
- Results: Better graph constructions and better clustering

NEXT STEPS...

- Curse-of-intrinsic-dimensionality is still a problem
- We need better priors!
  - Smoothness $\Rightarrow$ Higher-order kernels
  - Symmetry $\Rightarrow$ Group/Lie structure
  - High curvature embeddings $\Rightarrow$ Projections?
  - Multi-scale structure $\Rightarrow$ Hierarchical kernels/Deep learning
- Accessing the rest of the geometric information
  - Laplace de-Rham operator on differential forms
  - Beyond manifolds: fractals, nearly singular measures

RELEVANT ARTICLES

T. Berry, J. Harlim, Variable Bandwidth Diffusion Kernels. JACHA, (2015).

T. Berry, T. Sauer, Density estimation on manifolds with boundary, Comp. Stat. and Data Analysis 107, 1-17 (2017).

T. Berry, T. Sauer, Consistent manifold representation for topological data analysis, preprint.