Towards a mathematical foundation for machine learning (and forecasting)

Tyrus Berry

George Mason University

Aug. 26, 2021

Supported by NSF-DMS 1854204 and 2006808

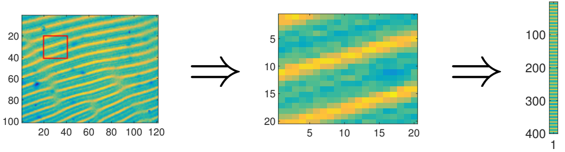
AVOIDING THE CURSE OF DIMENSIONALITY

Learning $f \in \mathcal{C}^s(\mathbb{R}^n, \mathbb{R})$ from N data points $\Rightarrow$ Error $\propto N^{-s/n}$

Coping mechanisms:

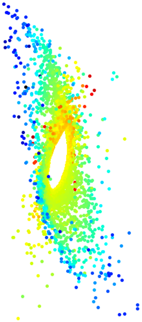
- **Smooth it away:** Assume f is very smooth, ie. $s \propto n$
- **Independence:** Assume $Y = f(X)$ is conditionally independent of X given $Z = g(X) \in \mathbb{R}^m$ with $m \ll n$.
- **Redundancy:** Assume $h(X) = 0$ for some $h \in \mathcal{C}^{m+1}(\mathbb{R}^n, \mathbb{R}^{n-m})$.
  - $\Rightarrow$ Data lies on/near a manifold $\mathcal{M} \subset \mathbb{R}^n$

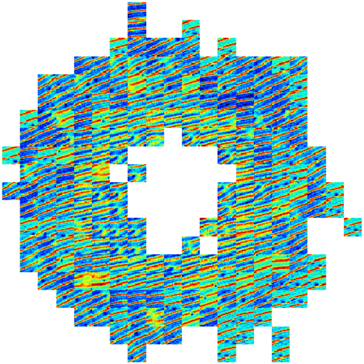
Finding Hidden Structure in Data



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

The sub-image geometry:





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
  - $\Delta f = \text{div} \circ \nabla = \frac{1}{\sqrt{|g|}} \partial_i g^{ij} \sqrt{|g|} \partial_j f$
  - $g(\nabla f, \nabla h) = \frac{1}{2}(f\Delta h + h\Delta f - \Delta(fh))$

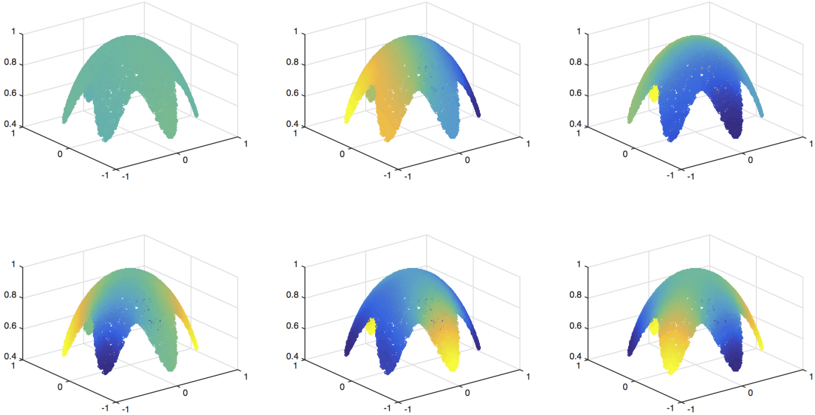
WHAT IS MANIFOLD LEARNING?

- **Manifold learning ⇔ 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

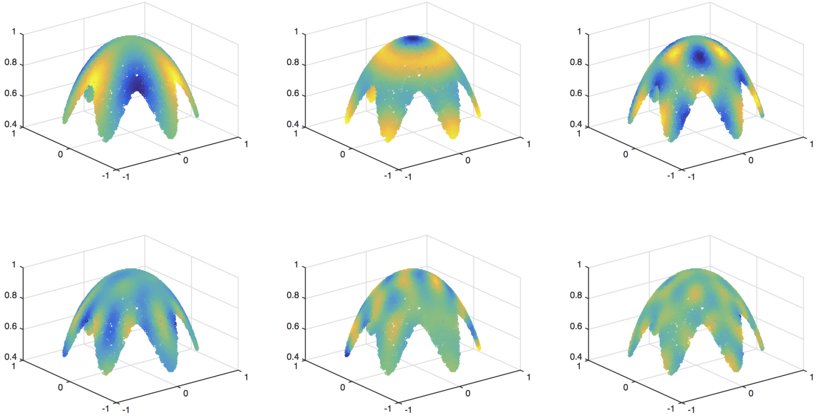
$$\lambda_i = \min_{\substack{f \perp \varphi_k \\ k=1,\ldots,i-1}} \left\{ \frac{\int_{\mathcal{M}} ||\nabla f||^2 \, dV}{\int_{\mathcal{M}} |f|^2 \, dV} \right\}$$

- 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 $\mathcal{M}$

Harmonic Analysis on Manifolds

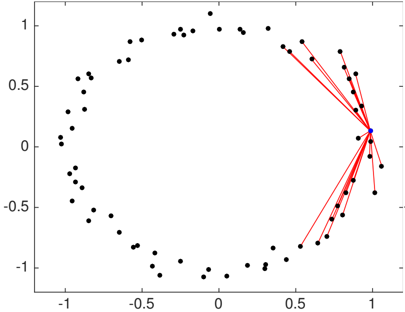


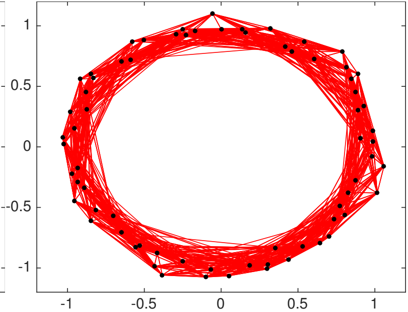
Harmonic Analysis on Manifolds



SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

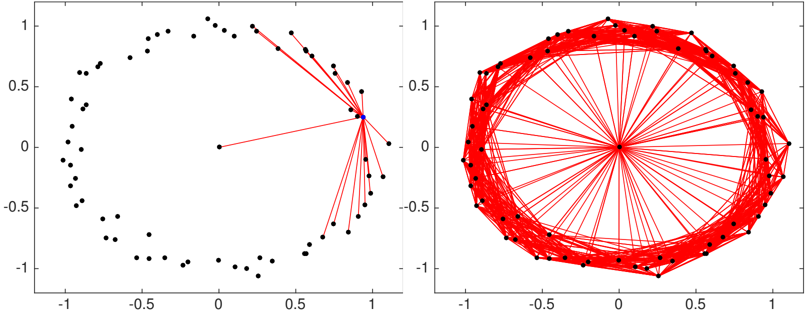
- Assume data lies on (or at least near) a manifold
- Approximate manifold with graph ⇒ Connect nearby points





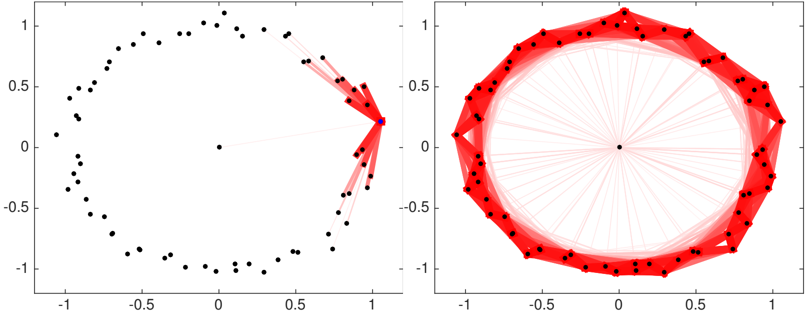
SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

- **Problem:** Noise and outliers can lead to *bridging*



SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

- To prevent bridging we weight the edges
- Edges are given weights $K_\delta(x,y) = e^{-\frac{||x-y||^2}{4\delta^2}}$



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

DISCRETE ANALOGS OF CONTINUOUS OBJECTS

Continuous	Discrete
$L^2(\mathcal{M}, q)$	$\mathbb{R}^N$
Functions, $f : \mathcal{M} \to \mathbb{R}$	Vectors, $\vec{f}_i = f(x_i)$
'Basis', δ_x	Basis, $\vec{e}_i = \delta_{x_i}$
Laplace-Beltrami, Δ	Normalized Graph Laplacian, $\mathbf{L}$
Eigenfunctions, $\Delta\varphi_j = \lambda_j\varphi_j$	Eigenvectors, $\mathbf{L}\vec{\varphi}_j = \lambda_j\vec{\varphi}_j$
Inner product, $\langle f, h\rangle_{L^2}$	Dot Product, $\frac{1}{N}\vec{f} \cdot \vec{h}$

$$\frac{1}{N}\vec{f} \cdot \vec{h} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)h(x_i) \to_{N\to\infty} \int_{\mathcal{M}} f(x)h(x)\, dV(x)$$

RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

- Arbitrary sampling (Coifman & Lafon, ‘Diffusion maps’, 2006)
- Other kernel functions (Berry & Sauer, 2015)
- Non-compact manifolds (Berry & Harlim, 2015)
- Boundary (R. Vaughn Thesis 2020)

$$\vec{h}^\top L\vec{f} \to \int \nabla h \cdot \nabla f \, dV$$

- Spectral convergence (von Luxburg et al. 2008, Trillos et al. 2020, Berry & Sauer 2019)

Conformally Invariant Diffusion Maps (CIDM)

- Data samples $\{x_i\}_{i=1}^N \subset \mathcal{M} \subset \mathbb{R}^n$ of volume $p_{\mathrm{eq}}\, dV$
- Continuous k-Nearest Neighbors (CkNN) dissimilarity:

$$d(x_i, x_j) \equiv \frac{||x_i - x_j||}{\sqrt{||x_i - x_{kNN(i)}||\,||x_j - x_{kNN(j)}||}}$$

- Variable bandwidth kernel, $K_{ij} = \exp\left(\frac{-d(x_i,x_j)^2}{\delta^2}\right)$
- Degree matrix $D_{ii} = \sum_j K_{ij}$ (diagonal)
- Graph Laplacian, $L = \frac{D-K}{\delta^{d+2}}$
- **Theorem:** $L\vec{f} = \Delta_{\hat{g}} f + \mathcal{O}\left(\delta^2, N^{-1/2}\delta^{-1-d/2}\right),\ \hat{g} = p_{\mathrm{eq}}^{2/d} g$
- **Solve:** $(I - D^{-1/2}KD^{-1/2})\vec{v} = \lambda\vec{v}$, set $\vec{\varphi} = D^{-1/2}\vec{v}$

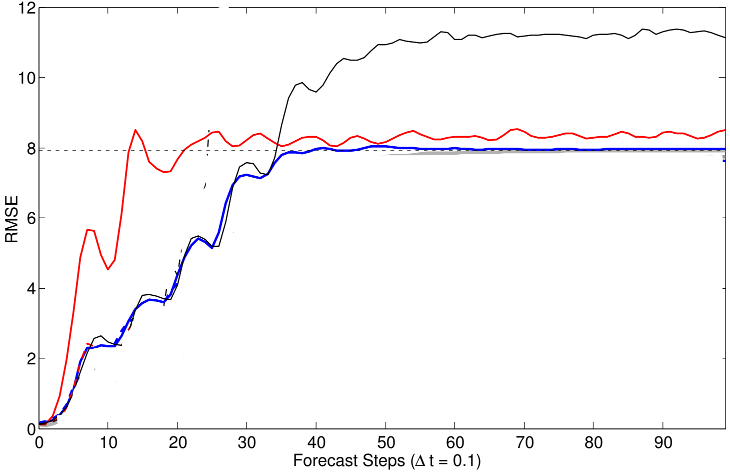
BEYOND MANIFOLD LEARNING

- Data never really lies on a manifold (due to noise)
- A manifold is a measure zero set
- Data is never sampled from a measure zero set
- **Solution 1:** Spectral robustness for bounded noise (Coifman and Lafon), but lose convergence
- **Solution 2:** Manifold + Noise, requires semi-geodesic coordinates, need new algorithms to regain convergence
- **Solution 3:** Generalize beyond manifolds
  - Metric measure spaces
  - Gromov-Hausdorff limits of manifolds

Types of Forecasting: Deterministic

- Deterministic Forecasting, $x_{k+1} = F(x_k)$
- Regression problem: Learn F from data
- Iterative Methods: $x_{k+n} = \tilde{F}^n(x_k)$ where $\tilde{F} \approx F$
- Direct Methods: $x_{k+n} = \tilde{F}_n(x_k)$ where $\tilde{F}_n \approx F^n$

Direct vs. Iterative vs Probabilistic



Types of Forecasting: UQ

- Deterministic Forecasting, $x_{k+1} = F(x_k), x_0 \sim p_0$
- Uncertainty Quantification, $p_{k+1} = \mathcal{F}(p_k) = p_k \circ F$
- *Can be* considered a regression problem
- *Option 1:* Learn F, then apply UQ (MC, PC, etc.)
- *Option 2:* Learn $\mathcal{F}$ directly in a basis

$$A_{ij} = \langle \phi_i, \mathcal{F}\phi_j \rangle = \langle \phi_i, \phi_j \circ F \rangle \approx \frac{1}{N} \sum_{k=1}^{N} \phi_i(x_k)\phi_j(x_{k+1})$$

TYPES OF FORECASTING: STOCHASTIC

- Stochastic Forecasting, $x_{k+1} = F(x_k, \omega_k)$
- *Not* a regression problem
- Don't just want $\bar{F}(\cdot) = \mathbb{E}_\omega[F(\cdot,\omega)]$
- We want the forward operator

$$p_{k+1} = \mathcal{F}(p_k) = \int p_k \circ F(\cdot,\omega)\, d\pi(\omega)$$

- Note: $\int p_k \circ F(\cdot,\omega)\, d\pi(\omega) \neq p_k \circ \int F(\cdot,\omega)\, d\pi(\omega)$

Stochastic forecasting = Operator Estimation

- Represent $\mathcal{F}$ in a basis

$$A_{ij} = \langle \phi_i, \mathcal{F}\phi_j \rangle = \langle \phi_i, \phi_j \circ F \rangle \approx \frac{1}{N}\sum_{k=1}^{N} \phi_i(x_k)\phi_j(x_{k+1})$$

- Error Sources: Bias, variance, and truncation
- Which basis?
  - Respect the measure $\Rightarrow$ Eliminate bias
  - Leverage smoothness $\Rightarrow$ Minimize variance
  - Capture global structure $\Rightarrow$ Minimize truncation

FORECASTING THE FOKKER-PLANK PDE

- Dynamical system: $dx = a(x)\,dt + b(x)\,dW_t$
- Uncertain initial state $x(0)$ with density $p(x,0)$
- Density solves Fokker-Planck PDE, $p_t = \mathcal{L}^* p$ where

$$\mathcal{L}^* p = -\nabla \circ (pa) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left(p \sum_k b_{ik} b_{jk}\right)$$

- Semigroup solution, $p(x,t) = e^{t\mathcal{L}^*} p(x,0)$

THE SHIFT MAP (STOCHASTIC KOOPMAN)

- Given data samples $x_i = x(t_i)$ with $\tau = t_{i+1} - t_i$
- Define the *shift map* of a function by $Sf(x_i) = f(x_{i+1})$
- Using the Itô lemma we can show:

$$Sf(x_i) = f(x_{i+1}) = e^{\tau\mathcal{L}}f(x_i) + \int_{t_i}^{t_{i+1}} \nabla f^\top b \, dW_s + \int_{t_i}^{t_{i+1}} Bf \, ds$$

- Notice: $\mathbb{E}[S(f)] = e^{\tau\mathcal{L}}f$
- Need to minimize the stochastic integrand $\nabla f^\top b$

Representing the Shift Map

- Choose a basis $\{\varphi_j\}$ orthonormal with respect to $\langle \cdot, \cdot \rangle_{p_{\text{eq}}}$
- The coefficients $c_l(t) = \langle p(x,t), \varphi_l \rangle$ have evolution:

$$\begin{aligned} c_l(t+\tau) &= \langle p(x, t+\tau), \varphi_l \rangle \\ &= \left\langle e^{\tau \mathcal{L}^*} p(x,t), \varphi_l \right\rangle = \left\langle p(x,t), e^{\tau \mathcal{L}} \varphi_l \right\rangle \\ &= \sum_j c_j(t) \left\langle \varphi_j, e^{\tau \mathcal{L}} \varphi_l \right\rangle_{p_{\text{eq}}} = \sum_j A_{lj} c_j(t) \end{aligned}$$

- So $\vec{c}(t+\tau) = A\vec{c}(t)$
- Where $A_{lj} = \left\langle \varphi_j, e^{\tau \mathcal{L}} \varphi_l \right\rangle_{p_{\text{eq}}} \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(x_i) \varphi_l(x_{i+1})$

Forecasting with the Shift Map

$$
\begin{array}{ccc}
p(x,t) & \xrightarrow{\text{Diffusion Forecast}} & p(x,t+\tau) \\
\Big\downarrow \langle p, \varphi_j \rangle & & \Big\uparrow \sum_j c_j \varphi_j p_{\mathrm{eq}} \\
\vec{c}(t) & \xrightarrow{A_{lj} \equiv \mathbb{E}[\langle \varphi_j, S\varphi_l \rangle_{p_{\mathrm{eq}}}]} & \vec{c}(t+\tau) = A\vec{c}(t).
\end{array}
$$

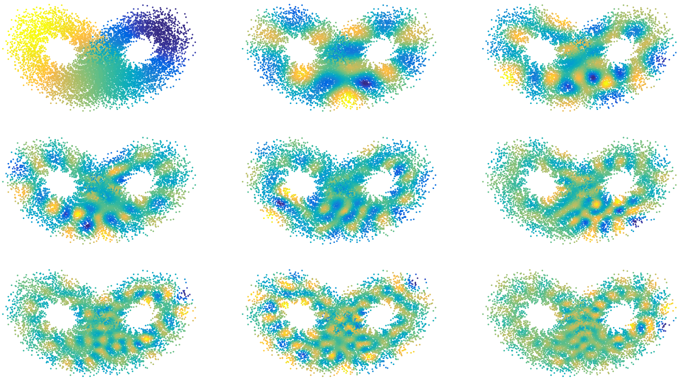
- Estimate A_{lj} with $\hat{A}_{lj} = \frac{1}{N}\sum_{i=1}^{N} \varphi_j(x_i)\varphi_l(x_{i+1})$
- $\mathbb{E}[\hat{A}_{lj}] = A_{lj}$ with error $\mathcal{O}(||\nabla \varphi_l||_{p_{\mathrm{eq}}}\sqrt{\tau/N})$

Choosing a basis

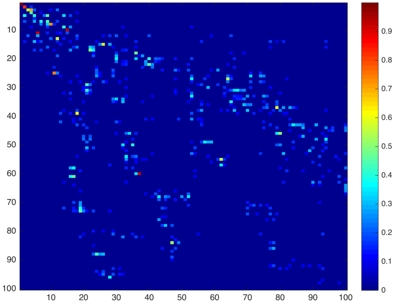
- Need to minimize the error term $\mathcal{O}(||\nabla\varphi_l||_{p_{\text{eq}}}\sqrt{\tau/N})$
- The eigenfunctions $\Delta_{\hat{g}}\varphi_j = \lambda_j\varphi_j$ minimize $||\nabla\varphi_j||_{p_{\text{eq}}} = \lambda_j$
- Find φ_j with Manifold Learning: CIDM

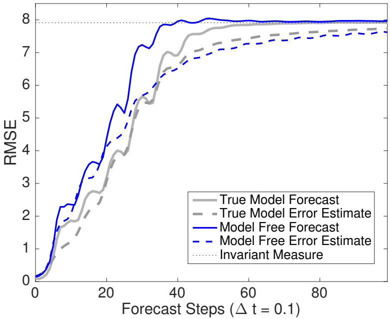
MANIFOLD LEARNING ⇒ CUSTOM 'FOURIER' BASIS

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



Shift map ⇒ Markov matrix



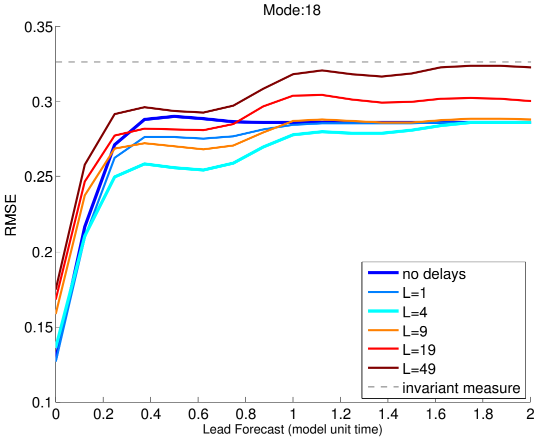


Diffusion Forecast Example

(Loading Video...)

Attractor Reconstruction

- Evolution of $y = h(x)$ is not closed
- Adding some delays helps, but adding too many hurts



RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

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

Local Kernels

- A *local kernel* has exponential decay:

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

- **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)

RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

- Arbitrary sampling (Coifman & Lafon, ‘Diffusion maps’, ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- 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 lower bounds on $\text{inj}(x)$ and $\inf_{y\in\mathcal{M}} R(x, y)$ independent of x

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 $\text{inj}(x_i) > \delta > 0$

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

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

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

Consistency is not enough!

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

Berry and Harlim (ACHA, 2015)

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):

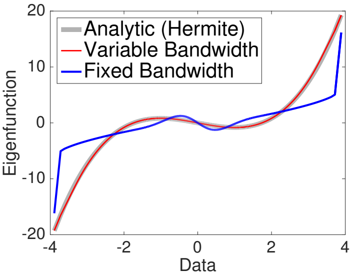
$$\mathbf{L}_{\delta,\alpha,\beta}\vec{f} = \Delta f + c_1 \nabla f \cdot \nabla \log q + \mathcal{O}\left(\delta^2, \frac{q^{-c_2}}{\sqrt{N}h^{1+d/2}}\right)$$

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

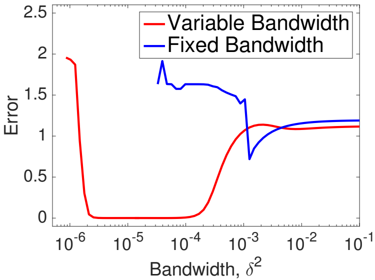
Example: Variable Bandwidth Kernel

Gaussian data: Brownian motion in quadratic potential

Eigenfunctions (Hermite)



Error vs. Bandwidth



Berry and Harlim (ACHA, 2015)

Summary of Manifold Learning

- Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami
- Can estimate Laplace-Beltrami with a graph Laplacian
- For a non-compact manifold:
  - Manifold must be tangible
  - Requires a variable bandwidth kernel
- Other contributions:
  - Access any desired geometry (local kernels)
  - Manifolds with boundary
  - Spectral convergence

CONTINUOUS K-NEAREST NEIGHBORS (CKNN)

Building unweighted graphs from data (TDA)

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

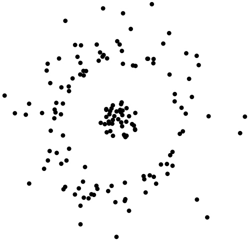
- $x_k = k$-th nearest neighbor of x
- Unnormalized graph Laplacian: $\mathbf{L}_{\mathrm{un}} = \mathbf{D} - \mathbf{K}$
- **Corollary:** $\mathbf{L}_{\mathrm{un}} \vec{f} \to \overrightarrow{\Delta_{\tilde{g}} f}$ where $(\tilde{g} = q^{2/d} g, d\tilde{V} = q\, dV)$
- **New result:** Spectral convergence $\mathbf{L}_{\mathrm{un}} \to \Delta_{\tilde{g}}$
- 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)

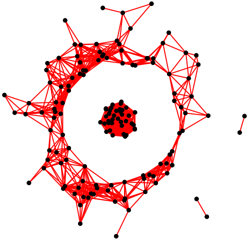
(Berry & Harlim (ACHA, 2015); Berry & Sauer (in review)

CkNN YIELDS IMPROVED GRAPH CONSTRUCTION

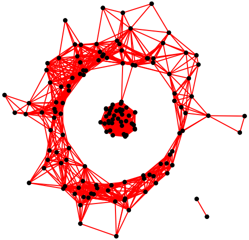
2D Gaussian with annulus removed:

Persistent vs. consistent homology

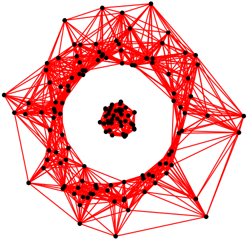




Small bandwidth

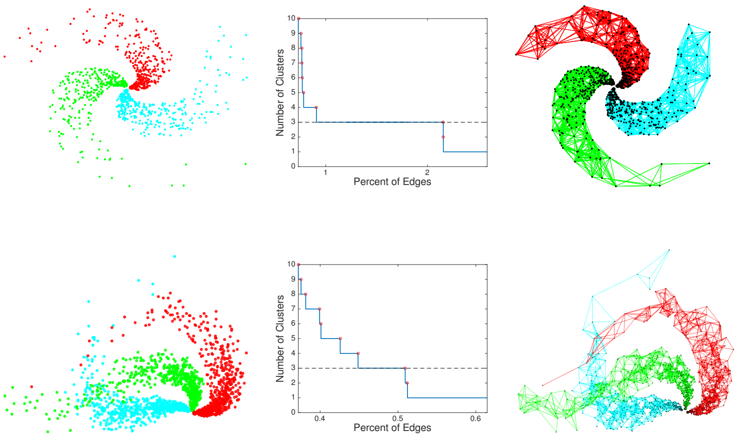


Large bandwidth



CkNN

IMPROVED CLUSTERING USING CKNN



Code and papers available at:

http://math.gmu.edu/˜berry/

Manifold Learning Papers Discussed

- B. and Giannakis, *Spectral Exterior Calclulus.*
- R. Vaughn *Diffusion Maps for Manifolds with Boundary.*
- B. and Sauer, *Consistent Manifold Representation for Topological Data Analysis.*
- Coifman and Lafon, *Diffusion maps.*
- B. and Harlim, *Variable Bandwidth Diffusion Kernels.*
- B. and Sauer, *Local Kernels and Geometric Structure of Data.*

References

[1] V. Vapnik, The nature of statistical learning theory. Springer (2000).

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press (2018).

[3] A. Pinkus, N-widths in Approximation Theory. Vol. 7. Springer Science & Business Media, (2012).

[4] R. A. DeVore and G. G. Lorentz. Constructive approximation. Vol. 303. Springer Science & Business Media, (1993).

[5] R. A. DeVore, Nonlinear approximation. Acta numerica 7, 51-150 (1998).

[6] D. W. Scott and S. R. Sain. Multidimensional density estimation. Handbook of statistics 24, 229-261 (2005).

[7] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory 39, no. 3, 930-945 (1993).

[8] A. R. Barron, Approximation and estimation bounds for artificial neural networks. Machine learning 14, no. 1, 115-133 (1994).

[9] H. J. Bungartz and M. Griebel, Sparse grids. Acta numerica, 13, 147-269 (2004).

[10] K. Li, Sliced Inverse Regression for Dimension Reduction. Journal of the American Statistical Association, 86(414), (1991).

[11] Y. Li, L. Zhu, Asymptotics for Sliced Average Variance Estimation. The Annals of Statistics, 35(1), 41-69 (2007).