Operator analysis of geometric data structures

Wojciech Czaja

Reduced Order Modeling in General Relativity
Pasadena, June 6, 2013

Joint work with:

University of Maryland: J. J. Benedetto, A. Cloninger, J. A. Dobrosotskaya, T. Doster, K. W. Duke, M. Ehler, A. Halevy, B. Manning, T. McCullough, V. Rajapakse

National Cancer Institute: Y. Pommier, W. Reinhold, B. Zeeberg

Remote Sensing Laboratory: M. L. McLane

Norbert Wiener Center
for Harmonic Analysis and Applications

Outline

Norbert Wiener Center for Harmonic Analysis and Applications

Outline

Norbert Wiener Center
for Harmonic Analysis and Applications

Introduction

- There is an abundance of available data. This data is often large, high-dimensional, noisy, and complex, e.g., gravitational waves.
- Typical problems associated with such data are to cluster, classify, or segment it; and to detect anomalies or embedded targets.
- Our proposed approach to deal with these problems is by combining techniques from harmonic analysis and machine learning:
  - **Harmonic Analysis** is the branch of mathematics that studies the representation of functions and signals.
  - **Machine Learning** is the branch of computer science concerned with algorithms that allow machines to infer rules from data.

Data Organization and Manifold Learning

- There are many techniques for Data Organization and Manifold Learning, e.g., Principal Component Analysis (PCA), Locally Linear Embedding (LLE), Isomap, genetic algorithms, and neural networks.
- We are interested in a subfamily of these techniques known as *Kernel Eigenmap Methods*. These include Kernel PCA, LLE, Hessian LLE (HLLE), and Laplacian Eigenmaps.
- Kernel eigenmap methods require two steps. Given data space X of N vectors in $\mathbb{R}^D$.
  1. Construction of an $N \times N$ symmetric, positive semi-definite kernel, K, from these N data points in $\mathbb{R}^D$.
  2. Diagonalization of K, and then choosing $d \leq D$ *significant* eigenmaps of K. These become our new coordinates, and accomplish, e.g., better cluster separation, dimensionality reduction.

We are particularly interested in diffusion kernels K, which are defined by means of transition matrices.

Kernel Eigenmap Methods for Dimension Reduction - Kernel Construction

- Kernel eigenmap methods were introduced to address complexities not resolvable by linear methods.
- The idea behind *kernel methods* is to express correlations or similarities between vectors in the data space X in terms of a symmetric, positive semi-definite kernel function $K : X \times X \to \mathbb{R}$. Generally, there exists a Hilbert space $\mathbb{K}$ and a mapping $\Phi : X \to \mathbb{K}$ such that

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle.$$

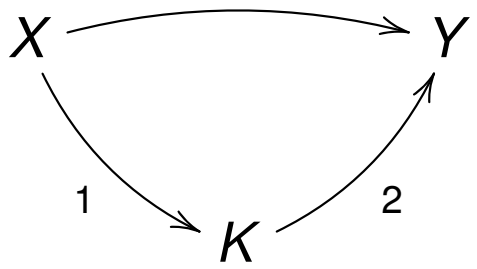
  Then, diagonalize by the spectral theorem and choose significant eigenmaps to obtain dimensionality reduction.
- Kernels can be constructed by many kernel eigenmap methods. These include Kernel PCA, LLE, HLLE, and Laplacian Eigenmaps.

Norbert Wiener Center for Harmonic Analysis and Applications

Kernel Eigenmap Methods for Dimension Reduction - Kernel Diagonalization

- The second step in kernel eigenmap methods is the diagonalization of the kernel.
- Let e_j, $j = 1, \ldots, N$, be the set of eigenvectors of the kernel matrix K, with eigenvalues λ_j.
- Order the eigenvalues monotonically.
- Choose the top $d << D$ significant eigenvectors to map the original data points $x_i \in \mathbb{R}^D$ to $(e_1(i), \ldots, e_d(i)) \in \mathbb{R}^d$, $i = 1, \ldots, N$.

Data Organization



There are other alternative interpretations for the steps of our diagram:

1. Constructions of kernels K may be independent from data and based on principles.
2. Redundant representations, such as frames, can be used to replace orthonormal eigendecompositions.

We need not select the target dimensionality to be lower than the dimension of the input. This leads, to data expansion, or data organization, rather then dimensionality reduction.

Operator Theory on Graphs

- Presented approach leads to analysis of operators on data-dependent structures, such as graphs or manifolds.
- Locally Linear Embedding, Diffusion Maps, Diffusion Wavelets, Laplacian Eigenmaps, Schroedinger Eigenmaps
- Mathematical core:
  - Pick a positive semidefinite bounded operator A as the infinitesimal generator of a semigroup of operators, e^{tA}, $t > 0$.
  - The semigroup can be identified with the Markov processes of diffusion or random walks, as is the case, e.g., with Diffusion Maps and Diffusion Wavelets
  - The infinitesimal generator and the semigroup share the common representation, e.g., eigenbasis

Norbert Wiener Center
for Harmonic Analysis and Applications

Example: Kernel PCA

Let $k : \mathbb{R}^D \to \mathbb{R}$ satisfy $k(x) = k(-x)$. Define

$$K(x_m, x_n) = \sum_{j=1}^{N} k(x_m - x_j)k(x_n - x_j)$$

A specific example of k is the Gaussian,

$$k(x) = e^{-c\|x\|^2} \text{ where } c > 0.$$

For this case, we then find a specific frame $\{\Phi_m\}_{m=1}^N$.

$$\Phi_m(x_n) = e^{-c(\|x_m\|^2+\|x_n\|^2)} \sum_{j=1}^{N} e^{2cx_j\cdot(x_m+x_n-x_j)},$$

so that $K(x_m, x_n) = \langle \Phi_m, \Phi_n \rangle$.

Norbert Wiener Center for Harmonic Analysis and Applications

Laplacian Eigenmaps - Theory

- M. Belkin and P. Niyogi, 2003
- Points close on the manifold should remain close in $\mathbb{R}^d$
- Let $f : \mathbb{R}^D \to \mathbb{R}$ represent the ideal embeding, then $|f(x) - f(y)| \leq \|\nabla f(x)\| \|x - y\| + o(\|x - y\|)$
- $\underset{\|f\|_{L^2(\mathcal{M})}=1}{\arg\min} \int_{\mathcal{M}} \|\nabla f(x)\|^2 = \underset{\|f\|_{L^2(\mathcal{M})}=1}{\arg\min} \int_{\mathcal{M}} \Delta_{\mathcal{M}}(f) f$
- Find eigenfunctions of the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$
- Use a discrete approximation of the Laplace-Beltrami operator
- Proven convergence (Belkin and Niyogi, 2003 – 2008)
- Introduced as an alternative to matched filtering techniques

Norbert Wiener Center
for Harmonic Analysis and Applications

Laplacian Eigenmaps - Implementation

1. Put an edge between nodes i and j if x_i and x_j are close. Precisely, given a parameter $k \in \mathbb{N}$, put an edge between nodes i and j if x_i is among the k nearest neighbors of x_j or vice versa.
2. Given a parameter $t > 0$, if nodes i and j are connected, set $W_{i,j} = e^{-\frac{\|x_i - x_j\|^2}{t}}$.
3. Set $D_{i,i} = \sum_j W_{i,j}$, and let $L = D - W$. Solve $Lf = \lambda Df$, under the constraint $y^\top Dy = Id$. Let $f_0, f_1, \ldots, f_d$ be $d+1$ eigenvector solutions corresponding to the first eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_d$. Discard f_0 and use the next d eigenvectors to embed in d-dimensional Euclidean space using the map $x_i \to (f_1(i), f_2(i), \ldots, f_d(i))$.

Swiss Roll

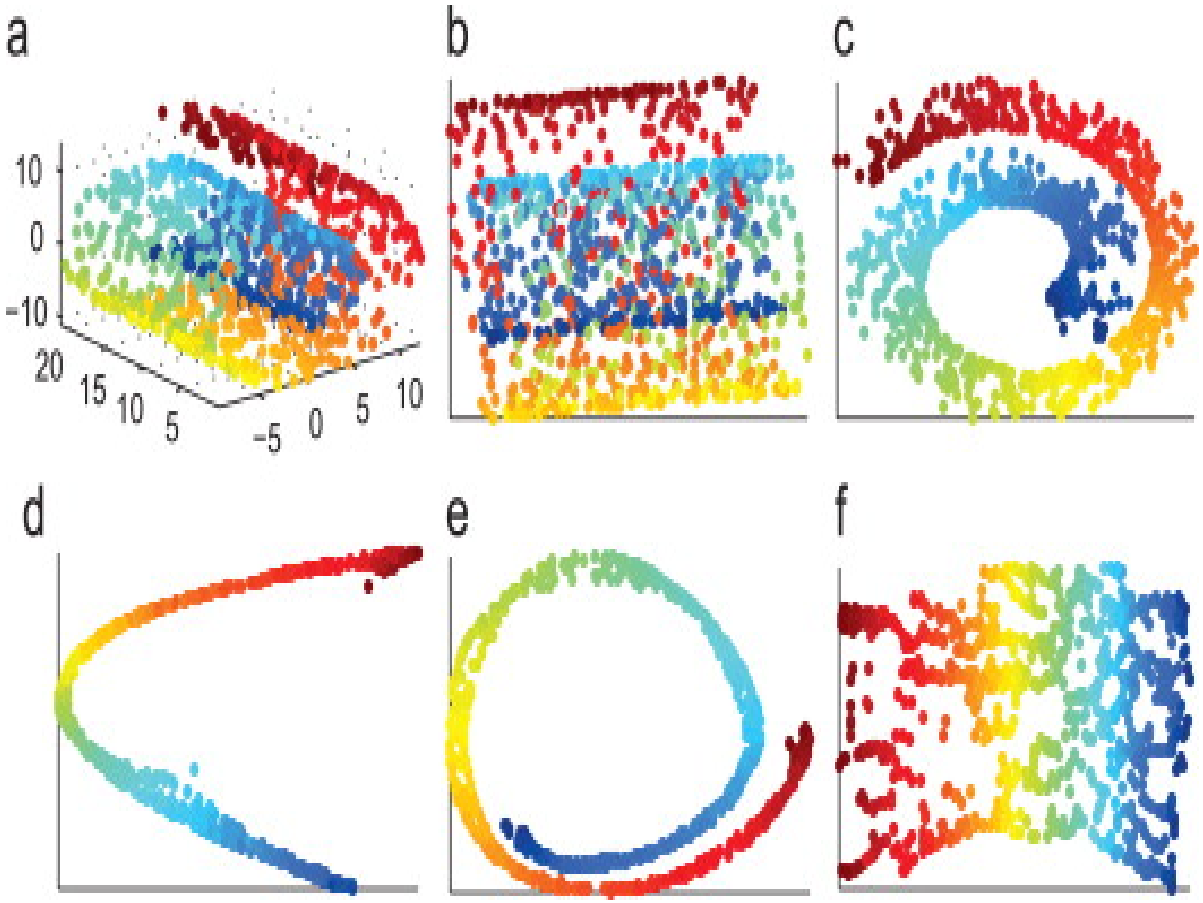


Figure : a) Original, b) PCA, c–f) LE, J. Shen et al., Neurocomputing, Volume 87, 2012

Norbert Wiener Center for Harmonic Analysis and Applications

Approximate Inversion of Laplacian Eigenmaps

- Laplacian Eigenmaps mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ is not invertible
- What if a new point $\psi \in \mathbb{R}^m$ is introduced into feature space?
- How do we approximately invert Φ?
  - Several papers (Sapiro, Schölkoph) attempt to find "approximate preimage" of ψ for simpler maps like kernel PCA
  - Approach: find the data point x that minimizes embedding error,

$$\min_{x \in \mathbb{R}^d} \|\Phi(x) - \psi\|_2$$

Laplacian Eigenmaps Inversion (with A. Cloninger)

1. Linearize Problem via Nyström extension to $\widehat{\Phi}(x) = V^* W$
2. Laplacian Eigenmaps construction guarantees sparsity of L, so incorporate Compressive Sensing LASSO problem

$$\widehat{W} = \arg\min \quad \|V^* L - \psi\|_2 + \tau \|L\|_1$$

3. Recover x via relation between $\widehat{L}$ and $\|x - x_i\|_2$ for the training points x_i that are nearest neighbors of x

From Laplacian to Schroedinger Eigenmaps

Consider the following minimization problem, $y \in \mathbb{R}^d$,

$$\min_{y^\top Dy = Id} \frac{1}{2} \sum_{i,j} \|y_i - y_j\|^2 W_{i,j} = \min_{y^\top Dy = E} tr(y^\top L y).$$

Its solution is given by the d minimal non-zero eigenvalue solutions of $Lf = \lambda Df$ under the constraint $y^\top Dy = Id$.

Similarly, for diagonal $\alpha \cdot V$, $\alpha > 0$, consider the problem

$$\min_{y^\top Dy = Id} \frac{1}{2} \sum_{i,j} \|y_i - y_j\|^2 W_{i,j} + \alpha \sum_i \|y_i\|^2 V_{i,i} = \min_{y^\top Dy = E} tr(y^\top (L + \alpha \cdot V) y), \tag{1}$$

which leads to solving equation $(L + \alpha V) f = \lambda D f$.

Schroedinger Eigenmaps

- Often we want to go from un-supervised to semi-supervised learning
- In SE, we replace L by $L + V$, where V is a nonnegative diagonal matrix (the potential)
- Schroedinger Eigenmaps (with Ehler, 2011) allow for the use of labeled data
- Enforce certain relations between the points
- Allow us to utilize expert input or templates in otherwise fully automated techniques such as LE.

Properties of Schroedinger Eigenmaps

Let the data graph be connected and let V be a symmetric positive semi-definite matrix.

Theorem (with M. Ehler)

Let the data graph be connected, let V be a symmetric positive semi-definite, and let $n \leq dim(Null(V))$. Then the minimizer of (1) satisfies:

$$\|y^{(\alpha)}\|_V^2 = trace^2({y^{(\alpha)}}^T V y^{(\alpha)}) \leq C\frac{1}{\alpha}.$$

In particular, if $V = diag(v_1, \ldots, v_N)$, then

$$v_i \|y_i^{(\alpha)}\|^2 \leq \sum_{i=1}^{N} v_i \|y_i^{(\alpha)}\|^2 \leq C_1 \frac{1}{\alpha}, \qquad \text{for all } i = 1, \ldots, N.$$

Pointwise Convergence of SE

Given n data points $x_1, x_2, \ldots, x_n$ sampled independently from a uniform distribution on a smooth, compact, d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$, define the operator $\hat{L}_{t,n} : C(\mathcal{M}) \to C(\mathcal{M})$ by

$$\hat{L}_{t,n}(f)(x) = \frac{1}{(4\pi t)^{d/2} t} \left(\frac{1}{n} \sum_j f(x) e^{-\frac{\|x - x_j\|^2}{4t}} - \frac{1}{n} \sum_j f(x_j) e^{-\frac{\|x - x_j\|^2}{4t}} \right).$$

Let $v \in C(\mathcal{M})$ be a potential. For $x \in \mathcal{M}$, let $y_n(x) = \underset{x_1, x_2, \ldots, x_n}{\arg\min} \|x - x_i\|$ and define $V_n : C(\mathcal{M}) \to C(\mathcal{M})$ by $V_n f(x) = v(y_n(x)) f(x)$.

Theorem (Pointwise Convergence, with A. Halevy)

Let $\alpha > 0$, and set $t_n = (\frac{1}{n})^{\frac{1}{d+2+\alpha}}$. For $f \in C^\infty(\mathcal{M})$,

$$\lim_{n \to \infty} \hat{L}_{t_n,n} f(x) + V_n f(x) = C \Delta_{\mathcal{M}} f(x) + v(x) f(x) \text{ in probability.}$$

Spectral Convergence of SE - Theorem

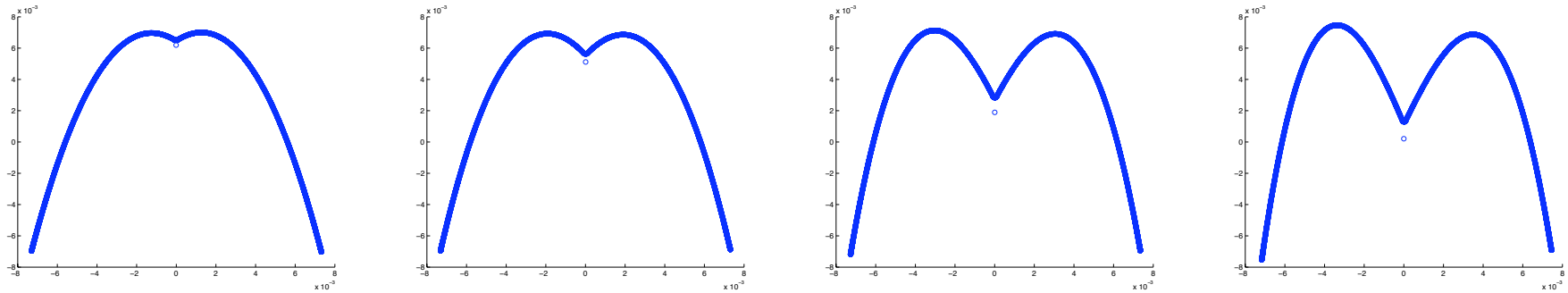
Let $\hat{L}_{t,n}$ be the unnormalized discrete Laplacian.

Theorem (Spectral Convergence of SE, with A. Halevy)

Let $\lambda^i_{t,n}$ and $e^i_{t,n}$ be the ith eigenvalue and corresponding eigenfunction of $\hat{L}_{t,n} + V_n$. Let λ_i and e_i be the ith eigenvalue and corresponding eigenfunction of $\Delta_{\mathcal{M}} + V$. Then there exists a sequence $t_n \to 0$ such that, in probability,

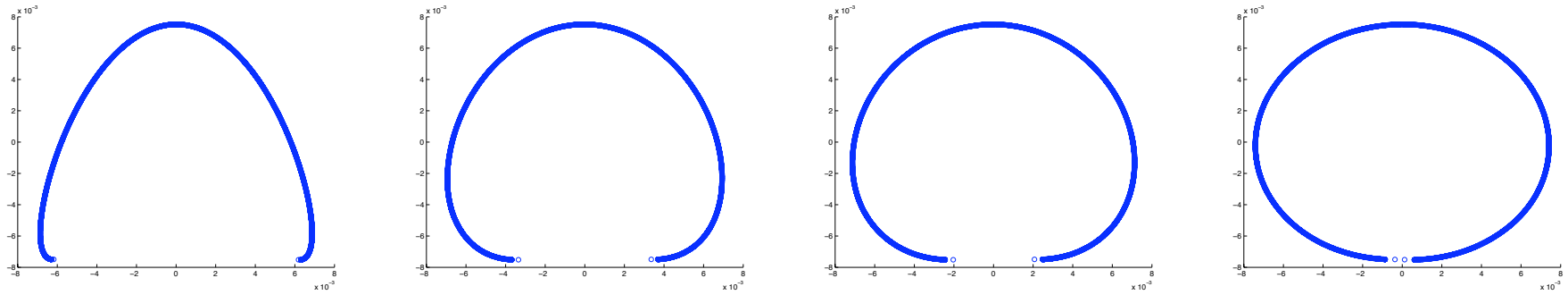
$$\lim_{n\to\infty} \lambda^i_{t_n,n} = \lambda_i \quad \textit{and} \quad \lim_{n\to\infty} \|e^i_{t_n,n} - e_i\| = 0.$$

SE as Semisupervised Method



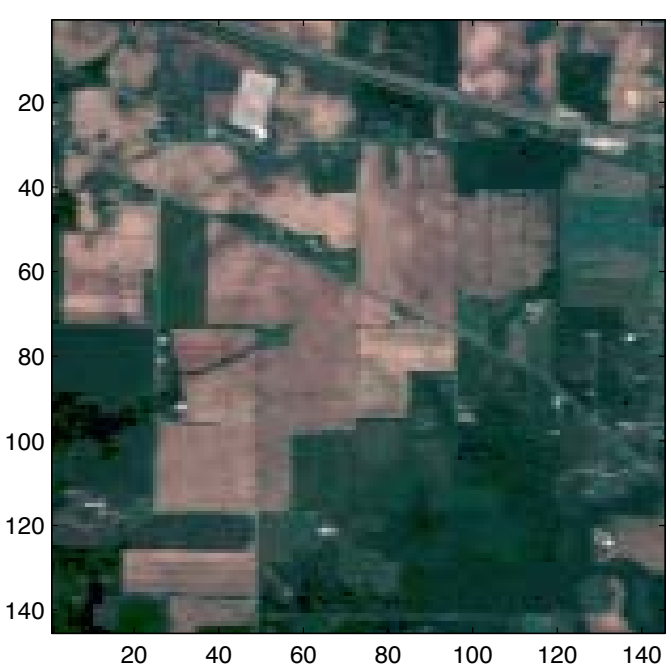
The Schroedinger Eigenmaps with diagonal potential $V = diag(0, \ldots, 0, 1, 0, \ldots, 0)$ only acting in one point y_{i_0} in the middle of the arc for $\alpha = 0.05, 0.1, 0.5, 5$. This point is pushed to zero.

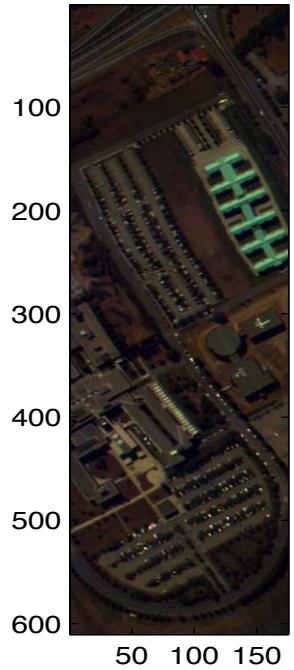
SE as Semisupervised Method



By applying the potential to the end points of the arc for $\alpha = 0.01, 0.05, 0.1, 1$, we are able to control the dimension reduction such that we obtain an almost perfect circle.

Hyperspectral data

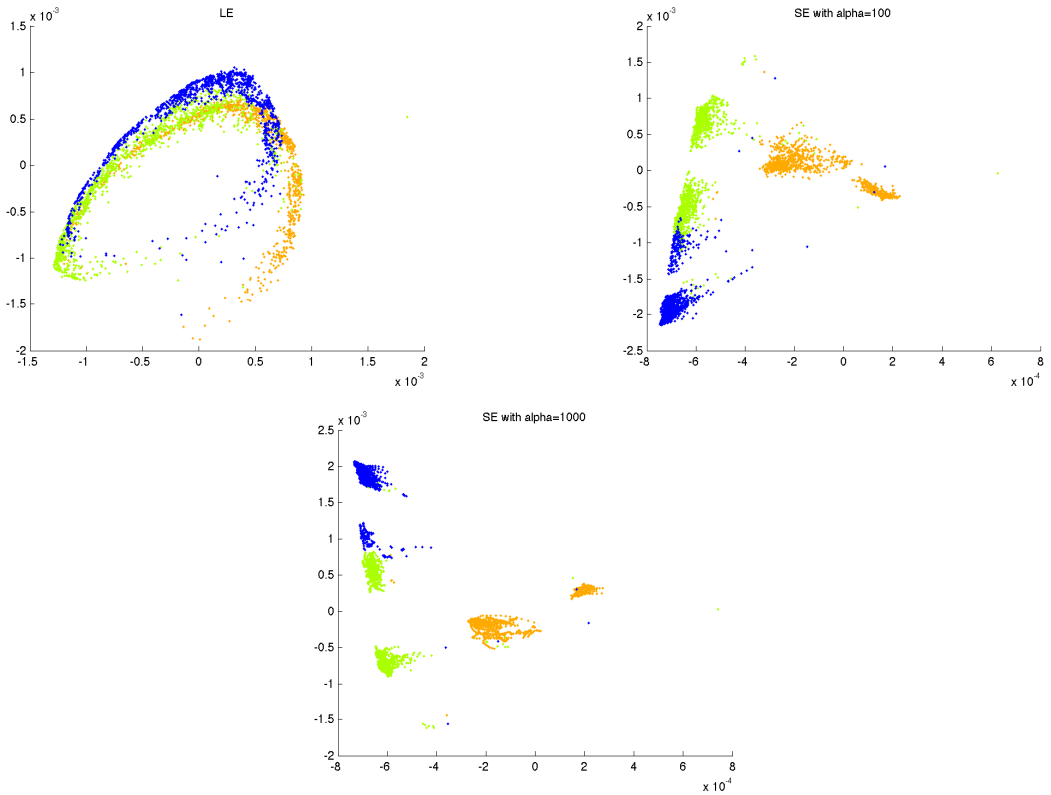




(left) The Indian Pines image is a 145×145 pixel image with 224 spectral bands. It was acquired using an AVIRIS spectrometer. (right) The Pavia University image is a 610×340 pixel image that contains 115 spectral bands. It was acquired using a ROSIS sensor.

Norbert Wiener Center for Harmonic Analysis and Applications

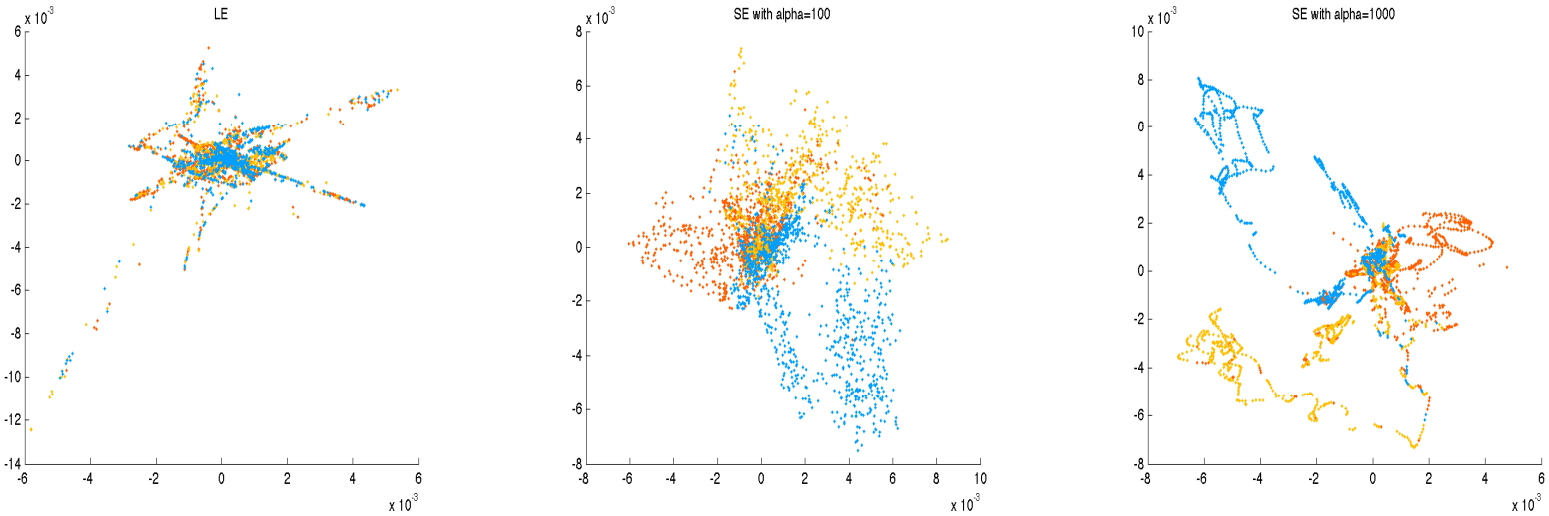
Impact of SE on Cluster analysis



Pavia University: Dimensions 4 and 5 of the LE and SE embeddings for classes 2 (meadows), 3 (gravel), and 7 (bitumen)

Norbert Wiener Center for Harmonic Analysis and Applications

Impact of SE on Cluster analysis



Indian Pines: Dimensions 17 and 22 of the LE and SE embeddings for classes 2 (corn 1), 3 (corn 2), and 10 (soybeen 1)

Norbert Wiener Center
for Harmonic Analysis and Applications

Outline

Norbert Wiener Center for Harmonic Analysis and Applications

Computational Bottleneck

1. If N is the ambient dimension, and n is the number of points, time complexity of constructing an adjacency graph is $O(DN^2)$
2. What can we do about D?
3. What can we do about the exponent 2?
4. What can we do about N?
5. What can we do about the computational complexity of eigendecomposition?

Numerical acceleration

1. Data Compression via Incoherent Random Projections
2. Fast Approximate k Nearest Neighbors algorithms
3. Quantization Landmarking
4. Randomized low-rank SVD decompositions

1. Setting for data compression

- Dataset $\{x_1, x_2, \ldots, x_N\}$ in $\mathbb{R}^D$, sampled from a compact K-dimensional Riemannian manifold
- Assume $\|x_i - x_j\| \leq A$ for all i, j and some $A > 0$
- Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$ be the first K nonzero eigenvalues computed by LE, assumed simple, with $r = \min_{i,j} |\lambda_i - \lambda_j|$, and let f_j be a normalized eigenvector corresponding to λ_j
- Use a random orthogonal projector Φ to map the points to $\mathbb{R}^M$. Let $\hat{f}_j$ be the jth eigenvector computed by LE for the projected data set

1. Laplacian Eigenmaps with random projections

Theorem (with A. Halevy)

Fix $0 < \alpha < 1$ *and* $0 < \rho < 1$. *If*

$$M \geq \frac{4 - 2\ln(1/\rho)}{\epsilon^2/200 + \epsilon^3/3000} K \ln(CKD/\epsilon), \text{ where } \epsilon = \frac{r\alpha}{4AN(N-1)},$$

then, with probability at least $1 - \rho$,

$$\|f_j - \hat{f}_j\| < \alpha.$$

The constant C depends on properties of the manifold. Precisely, $C = \frac{1900RV}{\tau^{1/3}}$, *where R, V and* $1/\tau$ *are the geodesic covering regularity, volume, and condition number, respectively.*

Norbert Wiener Center for Harmonic Analysis and Applications

1. Application: Classification of Hyperspectral Data

(a) Urban Dataset

Table : Comparison of performance on Urban

Method	Time (min)	Accuracy (percent)
LE	15.26	79.05
LERP	11.78	78.44

1. Application: Classification of Hyperspectral Data

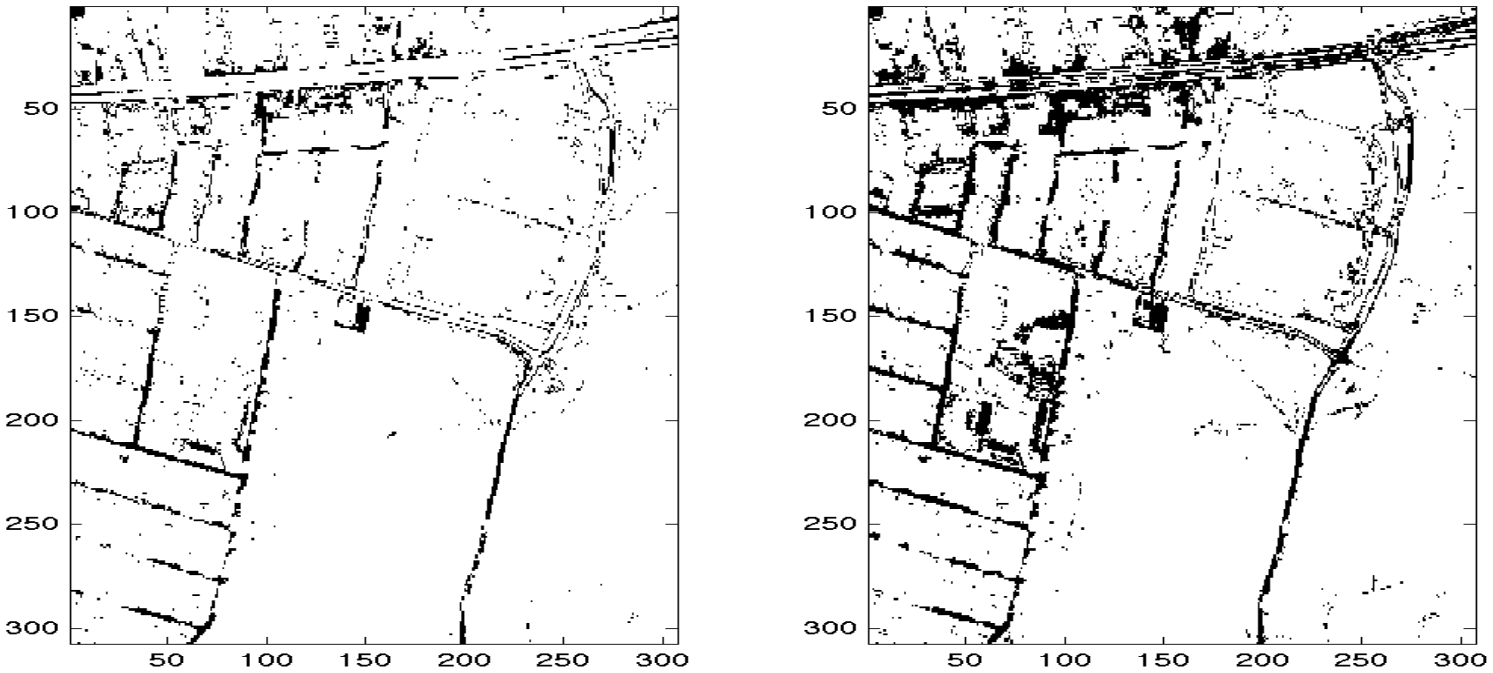
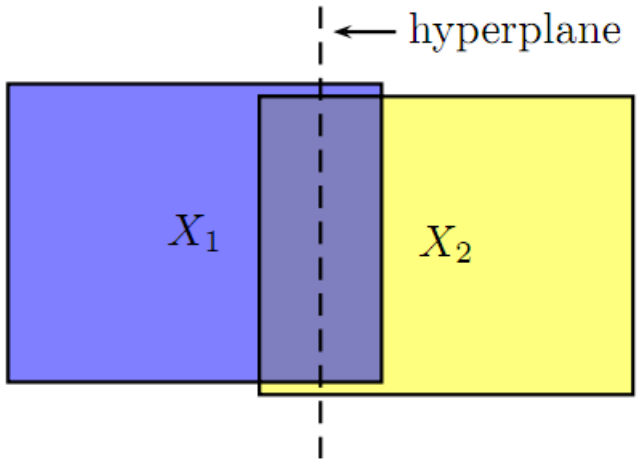


Figure : Urban class 2 (secondary road): left - LE, right - LERP

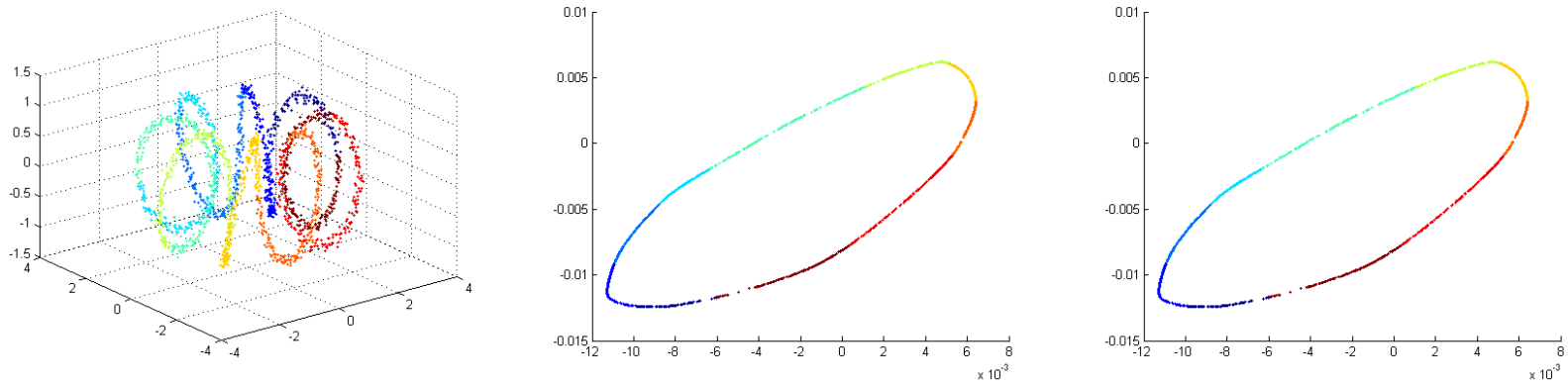
Norbert Wiener Center
for Harmonic Analysis and Applications

2. Fast Approximate k Nearest Neighbors

- There are many approximate nearest neighbor algorithms, e.g., Locality-sensitive Hashing (P. Indyk), Best Bin First (D. Lowe), or Clustered Point Sets Search (D. Mount). We present the Divide and Conquer method of Chen, Fang, and Saad
- Divide the set of points into two overlapping subsets using spectral bisection based on the Lanczos algorithm
- Once the size of a subset is less than a threshold r, compute using brute-force.
- If a data point belongs to more than one of the subsets, its nearest neighbors are selected from the neighbors found in each of the subsets.



2. Numerical Experiments: Synthetic Data



(a) Helix (b) Exact (c) Approximate

Figure : Mapping a one-dimensional helix embedded in $\mathbb{R}^3$. In the above example the exponent used is approx. 1.16 (depends on the size of overlap).

4. Robust Principle Component Analysis

- Consider PCA of data, with a fraction of the entries grossly corrupted due to, e.g., sensor malfunction on some measurements or random pixels occluded by irrelevant data.
- Candès introduced a version of PCA that eliminates such gross corruption via compressive sensing.
- Algorithm relies on using Singular Value Decompositions (SVD) which is computationally too expensive.
- Independently, Rokhlin introduced a randomized, approximate SVD algorithm that works well when matrix is low rank.

Speed up of Robust PCA (with A. Cloninger and G. Warnell)

Under certain assumptions on corrupted entries, Rokhlin's randomized SVD algorithm is used to speed up Candès PCA by **several orders of magnitude without loss of precision**.