Harmonic and multiscale analysis of and on data sets in high dimensions

Mauro Maggioni

Mathematics and Computer Science
Duke University

French-Chinese-Singaporean joint workshop on Wavelets and Applications, N.U.S., 6/12/08

In collaboration with R.R. Coifman, P.W. Jones, Y-M. Jung, R. Schul, A.D. Szlam, and more...
Funding: ONR, NSF-DMS, Sloan Foundation
Plan

- Harmonic Analysis on graphs
- Eigenfunctions of the Laplacian
  - Playground
  - Why do we care?
  - A Theorem for manifolds
    - Riemann mapping
    - Eigenfunction mapping
    - Heat kernel mapping
- Things I will not talk about
  - Fourier analysis of functions on graphs
  - Diffusion wavelets
  - (Many) Applications
- Conclusion
Harmonic Analysis on graphs

Suppose we are given a weighted graph \((G, E, W)\), with \(n\) vertices and edges \(E\) with weights \(W_{ij}\) (similarities!). How can we start doing harmonic analysis?

Let \(D_{ii} = \sum_j W_{ij}\) and

\[
P = D^{-1}W, \quad T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}, \quad \mathcal{L} = I - T, \quad H = e^{-t\mathcal{L}}
\]

random walk \quad symm. “random walk” \quad Laplacian \quad Heat kernel

Spectral decomposition: \(\mathcal{L}\varphi_i = \lambda_i \varphi_i\). Fourier modes: \(\{\varphi_i\}\).
How do these eigenfunctions look like?

We know a lot when: $G=$regular lattice in $R^n$.

Otherwise...they may be quite complicated!

Studied by many people in many communities: physicists (resonances), computer scientists (graph layouts/cuts), mathematicians (global analysis on manifolds)...

Spectral graph layout: map $G \rightarrow R^d$ by

$$\Phi(x) = (\varphi_1(x), \ldots, \varphi_d(x)).$$

Can be used to map any graph to $R^d$. 
How do these eigenfunctions look like?

We know a lot when: $G=\text{regular lattice in } \mathbb{R}^n$.

Otherwise...they may be quite complicated!

Studied by many people in many communities: physicists (resonances), computer scientists (graph layouts/cuts), mathematicians (global analysis on manifolds)...

Spectral graph layout: map $G \rightarrow \mathbb{R}^d$ by

$$\Phi(x) = (\varphi_1(x), \ldots, \varphi_d(x)).$$

Can be used to map any graph to $\mathbb{R}^d$. 

Mauro Maggioni
Harmonic analysis of and on data sets
How do these eigenfunctions look like?

We know a lot when: $G=$regular lattice in $R^n$.

Otherwise...they may be quite complicated!

Studied by many people in many communities: physicists (resonances), computer scientists (graph layouts/cuts), mathematicians (global analysis on manifolds)...

Spectral graph layout: map $G \rightarrow R^d$ by

$$\Phi(x) = (\varphi_1(x), \ldots, \varphi_d(x)).$$

Can be used to map any graph to $R^d$. 
Eigenfunctions and spectral mappings

How do these eigenfunctions look like?

We know a lot when: \( G = \text{regular lattice in } \mathbb{R}^n. \)

Otherwise...they may be quite complicated!

Studied by many people in many communities: physicists (resonances), computer scientists (graph layouts/cuts), mathematicians (global analysis on manifolds)...

Spectral graph layout: map \( G \rightarrow \mathbb{R}^d \) by

\[
\Phi(x) = (\varphi_1(x), \ldots, \varphi_d(x)).
\]

Can be used to map any graph to \( \mathbb{R}^d \).
Eigenfunctions and spectral mappings

How do these eigenfunctions look like?

We know a lot when: $G=$regular lattice in $\mathbb{R}^n$.

Otherwise...they may be quite complicated!

Studied by many people in many communities: physicists (resonances), computer scientists (graph layouts/cuts), mathematicians (global analysis on manifolds)...

Spectral graph layout: map $G \rightarrow \mathbb{R}^d$ by

$$\Phi(x) = (\varphi_1(x), \ldots, \varphi_d(x)).$$

Can be used to map any graph to $\mathbb{R}^d$. 
www.math.duke.edu/~mauro

Look under "code" or "teaching".
A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, neurobiological signals, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

Common feature/assumption: data is given in a high dimensional space, however it has a much lower dimensional intrinsic geometry.

(i) physical constraints. For example the effective state-space of at least some proteins seems low-dimensional, at least when viewed at the time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example many dependencies among word frequencies in a document corpus force the distribution of word frequency to low-dimensional, compared to the dimensionality of the whole space.
Why do we care?

A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, neurobiological signals, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

Common feature/assumption: data is given in a high dimensional space, however it has a much lower dimensional intrinsic geometry.

(i) physical constraints. For example the effective state-space of at least some proteins seems low-dimensional, at least when viewed at the time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example many dependencies among word frequencies in a document corpus force the distribution of word frequency to low-dimensional, compared to the dimensionality of the whole space.
Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

Simplest example: $K_\sigma(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \sigma}$.

Model the data as a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := K(x_i, x_j)$, when positive. Then we have a weighted graph - which we started from - and natural operators of them, such as the Laplacian $\mathcal{L}$.

Note 1: $K$ depends on the type of data.

Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Assume the data \( X = \{ x_i \} \subset \mathbb{R}^n \). Assume we can assign local similarities via a kernel function \( K(x_i, x_j) \geq 0 \).

Simplest example: \( K_\sigma(x_i, x_j) = e^{-\|x_i - x_j\|^2/\sigma} \).

Model the data as a weighted graph \((G, E, W)\): vertices represent data points, edges connect \( x_i, x_j \) with weight \( W_{ij} := K(x_i, x_j) \), when positive. Then we have a weighted graph - which we started from - and natural operators of them, such as the Laplacian \( \mathcal{L} \).

Note 1: \( K \) depends on the type of data.

Note 2: \( K \) should be “local”, i.e. close to 0 for points not sufficiently close.
Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

Simplest example: $K_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.

Model the data as a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := K(x_i, x_j)$, when positive. Then we have a weighted graph - which we started from - and natural operators of them, such as the Laplacian $\mathcal{L}$.

Note 1: $K$ depends on the type of data.
Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Data base of about 60,000 $28 \times 28$ gray-scale pictures of handwritten digits, collected by USPS. Point cloud in $\mathbb{R}^{28^2}$. Goal: automatic recognition.

Set of 10,000 picture (28 by 28 pixels) of 10 handwritten digits. Color represents the label (digit) of each point.
1000 Science News articles, from 8 different categories. We compute about 10000 coordinates, $i$-th coordinate of document $d$ represents frequency in document $d$ of the $i$-th word in a fixed dictionary.
The dynamics of a small protein (12 atoms, $H$ atoms removed) in a bath of water molecules is approximated by a Langevin system of stochastic equations $\dot{x} = -\nabla U(x) + \dot{w}$. The set of states of the protein is a noisy ($\dot{w}$) set of points in $\mathbb{R}^{36}$.

Left: representation of an alanine dipeptide molecule. Right: embedding of the set of configurations.
In several instances the geometry of the data can help construct useful priors, for tasks such as classification, regression for prediction purposes.

This applies also to the case in which data is given as a graph. Some issues I am interested in:

- **geometric**: find intrinsic properties, such as local dimensionality, and local parameterizations.
- **approximation theory**: approximate functions on such data, respecting the geometry.
We start by analyzing the intrinsic geometry of the data, and then working on function approximation on the data.

- Find parametrizations for the data: manifold learning, dimensionality reduction. Ideally: number of parameters equal to, or comparable with, the intrinsic dimensionality of data (as opposed to the dimensionality of the ambient space), such a parametrization should at least approximately preserve distances and be stable under perturbations of the manifold.

- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.
We start by analyzing the intrinsic geometry of the data, and then working on function approximation on the data.

- Find parametrizations for the data: manifold learning, dimensionality reduction. Ideally: number of parameters equal to, or comparable with, the intrinsic dimensionality of data (as opposed to the dimensionality of the ambient space), such a parametrization should at least approximately preserve distances and be stable under perturbations of the manifold.

- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.
Spectral embedding seems to work quite well, very popular. Very few results proving its efficiency exist (e.g. Tutte’s theorem for planar graphs, 1963). In applications, more and more data can be acquired, and larger and larger graphs constructed: we assume the existence of a continuous underlying structure, in particular a manifold, from which the points are sampled and the graph constructed. Several results guarantee that natural operators on the graph approximate suitably (and tend to) those on the underlying continuous object.
Spectral embedding seems to work quite well, very popular. Very few results proving its efficiency exist (e.g. Tutte’s theorem for planar graphs, 1963).
In applications, more and more data can be acquired, and larger and larger graphs constructed: we assume the existence of a continuous underlying structure, in particular a manifold, from which the points are sampled and the graph constructed. Several results guarantee that natural operators on the graph approximate suitably (and tend to) those on the underlying continuous object.
Spectral embedding seems to work quite well, very popular. Very few results proving its efficiency exist (e.g. Tutte’s theorem for planar graphs, 1963). In applications, more and more data can be acquired, and larger and larger graphs constructed: we assume the existence of a continuous underlying structure, in particular a manifold, from which the points are sampled and the graph constructed. Several results guarantee that natural operators on the graph approximate suitably (and tend to) those on the underlying continuous object.
Let $\mathcal{M}$ be a Riemannian manifold. Coordinate chart $\mathcal{M} \supseteq \tilde{B} \to \tilde{B} \subseteq \mathbb{R}^d$, one-to-one, $F(x) = (f_1(x), f_2(x), \ldots, f_d(x))$.

Distortion of $F$ (on $B$): $\|F\|_{Lip} \times \|F^{-1}\|_{Lip}$, where

$$\|F\|_{Lip} = \sup_{x, y \in B, x \neq y} \frac{\|F(x) - F(y)\|}{d_M(x, y)}.$$

Prime example: coordinate chart on a simply connected planar domain $\mathcal{D}$, $|\mathcal{D}| = 1$, given by a Riemann mapping $F : \mathcal{D} \to \mathbb{D}$, normalized so that $F(z_0) = 0$.

For $z_0 \in \mathcal{D}$, let $r = \text{dist}(z_0, \partial \mathcal{D})$. Then

$$B(0, \kappa^{-1}) \subset F(B(z_0, \frac{r}{2})) \subset B(0, 1 - \kappa^{-1}),$$

with distortion less than $\kappa$. Think of $F$ on $B(z_0, \frac{r}{2})$ as a perturbation of the linear map $z \to F'(z_0)(z - z_0)$, and $|F'(z_0)| \sim \frac{1}{r}$. 
Let $\mathcal{M}$ be a Riemannian manifold. Coordinate chart $\mathcal{M} \supseteq B \to \tilde{B} \subseteq \mathbb{R}^d$, one-to-one, $F(x) = (f_1(x), f_2(x), \ldots, f_d(x))$. Distortion of $F$ (on $B$): $\|F\|_{\text{Lip}} \times \|F^{-1}\|_{\text{Lip}}$, where

$$
\|F\|_{\text{Lip}} = \sup_{x,y \in B, x \neq y} \frac{\|F(x) - F(y)\|}{d_M(x, y)}.
$$

Prime example: coordinate chart on a simply connected planar domain $\mathcal{D}$, $|\mathcal{D}| = 1$, given by a Riemann mapping $F : \mathcal{D} \to \mathbb{D}$, normalized so that $F(z_0) = 0$. For $z_0 \in \mathcal{D}$, let $r = \text{dist}(z_0, \partial \mathcal{D})$. Then

$$
B(0, \kappa^{-1}) \subset F(B(z_0, \frac{r}{2})) \subset B(0, 1 - \kappa^{-1}),
$$

with distortion less than $\kappa$. Think of $F$ on $B(z_0, \frac{r}{2})$ as a perturbation of the linear map $z \to F'(z_0)(z - z_0)$, and $|F'(z_0)| \sim \frac{1}{r}$. 

Mauro Maggioni
Harmonic analysis of and on data sets
We look for analogue of the above, on Riemannian manifolds of finite volume: Back to $\mathcal{D}$, $z_0$ and $r$, $F$ as above. Classical formula known to Riemann:

$$F(z) = \exp\left\{-G(z, z_0) - iG^*(z, z_0)\right\},$$

$G(\cdot, z_0)$ the Green’s function for $\mathcal{D}$, and $G^*$ the (multivalued) conjugate. But $G(z, z_0) = \int_0^\infty K(z, z_0, t)dt$, $K$ the (Dirichlet) heat kernel for $\mathcal{D}$. Since

$$K(z, z_0, t) = \sum_{j=1}^\infty \varphi_j(z)\varphi_j(z_0)e^{\lambda_j t}$$

where $\{\varphi_j\}$ are global Dirichlet eigenfunctions. Since $|F'(z)| = |\nabla G(z, z_0)|e^{-G(z, z_0)} \sim \frac{1}{r}$ on $B(z_0, \frac{r}{2})$, one may guess that there are eigenfunctions $\varphi_j$ such that

$$|\nabla \varphi_j| \gtrsim \frac{1}{r}$$

on $B(z_0, \kappa^{-1}r)$, for some $\kappa > 1$, independent of $\mathcal{D}$. (A short calculation with Weyl’s estimates makes this reasonable.)
A Theorem, philosophy

We look for analogue of the above, on Riemannian manifolds of finite volume: Back to \( \mathcal{D}, z_0 \) and \( r, F \) as above. Classical formula known to Riemann:

\[
F(z) = \exp\left\{ -G(z, z_0) - iG^*(z, z_0) \right\},
\]

where \( G(\cdot, z_0) \) the Green’s function for \( \mathcal{D} \), and \( G^* \) the (multivalued) conjugate. But \( G(z, z_0) = \int_0^\infty K(z, z_0, t)dt \), \( K \) the (Dirichlet) heat kernel for \( \mathcal{D} \). Since

\[
K(z, z_0, t) = \sum_{j=1}^{\infty} \varphi_j(z)\varphi_j(z_0)e^{\lambda_j t}
\]

where \( \{\varphi_j\} \) are global Dirichlet eigenfunctions. Since

\[
|F'(z)| = |
\nabla G(z, z_0)|e^{-G(z, z_0)} \sim \frac{1}{r}
\]
on \( B(z_0, \frac{r}{2}) \), one may guess that there are eigenfunctions \( \varphi_j \) such that

\[
|\nabla \varphi_j| \gtrsim \frac{1}{r}
\]
on \( B(z_0, \kappa^{-1}r) \), for some \( \kappa > 1 \), independent of \( \mathcal{D} \). (A short calculation with Weyl’s estimates makes this reasonable.)
We look for analogue of the above, on Riemannian manifolds of finite volume: Back to $\mathcal{D}$, $z_0$ and $r$, $F$ as above. Classical formula known to Riemann:

$$F(z) = \exp\left\{ -G(z, z_0) - iG^*(z, z_0) \right\},$$

$G(\cdot, z_0)$ the Green's function for $\mathcal{D}$, and $G^*$ the (multivalued) conjugate. But $G(z, z_0) = \int_0^\infty K(z, z_0, t)dt$, $K$ the (Dirichlet) heat kernel for $\mathcal{D}$. Since

$$K(z, z_0, t) = \sum_{j=1}^\infty \varphi_j(z)\varphi_j(z_0)e^{\lambda_j t}$$

where $\{\varphi_j\}$ are global Dirichlet eigenfunctions. Since

$$|F'(z)| = |\nabla G(z, z_0)|e^{-G(z, z_0)} \sim \frac{1}{r}$$
on $B(z_0, \frac{r}{2})$, one may guess that there are eigenfunctions $\varphi_j$ such that

$$|\nabla \varphi_j| \gtrsim \frac{1}{r}$$
on $B(z_0, \kappa^{-1}r)$, for some $\kappa > 1$, independent of $\mathcal{D}$. (A short calculation with Weyl's estimates makes this reasonable.)
Charts and local parametrizations

\[ \mathbb{R}^n \]

\[ \dim \mathcal{M} = d \]

\[ \Phi(w) - w \approx d_{\mathcal{M}}(x_1, x_2) \]

Mauro Maggioni

Harmonic analysis of and on data sets
We prove that there is a \textbf{locally} defined $F$ that has these properties, and that this choice of $F$ will come from \textbf{globally} defined Laplacian eigenfunctions. On a metric embedded ball $B \subset \mathcal{M}$ we will choose \textbf{global} Laplacian eigenfunctions $\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_d}$ and constants $\gamma_1, \gamma_2, \ldots, \gamma_d \leq \kappa$ (for a universal constant $\kappa$) and define

$$\Phi := (\gamma_1 \varphi_{i_1}, \gamma_2 \varphi_{i_2}, \ldots, \gamma_d \varphi_{i_d}).$$

This choice of $\Phi$, depending heavily on $z_0$ and $r$, is globally defined, and on $B(z_0, \kappa^{-1}r)$ enjoys the same properties as the Riemann map. In other words, $\Phi$ maps $B(z_0, \kappa^{-1}r)$ to, roughly, a ball of unit size, with low distortion.
A Theorem, picture

Eigenfunction \( \phi_j \) with \( \lambda_j \sim R_\mathbb{R}^{-2} \)

"oscillates half-time" in given direction

\[ \partial_p \phi \sim \frac{1}{R_\mathbb{R}} \text{ on } B_2(R_\mathbb{R}) \]
Parametrizations through eigenfunctions

Figure: Top left: a non-simply connected domain in $\mathbb{R}^2$, and the point $z$ with its neighborhood to be mapped. Top right: the image of the neighborhood under the map. Bottom: Two good eigenfunctions.
A smooth, $d$-dimensional compact manifold, possibly with boundary. Metric tensor $g \in C^\alpha$. Fix $z_0 \in \mathcal{M}$, let $(U, F)$ be a coordinate chart such that $z_0 \in U$ and normalized so that $g^{il}(F(z_0)) = \delta^{il}$. Assume that for any $x \in U$, $\xi, \nu \in \mathbb{R}^d$,

$$
c_{\min}(g)\|\xi\|_{\mathbb{R}^d}^2 \leq \sum_{i,j=1}^{d} g^{ij}(F(x))\xi_i\xi_j, \quad \sum_{i,j=1}^{d} g^{ij}(F(x))\xi_i\nu_j \leq c_{\max}(g)\|\xi\|_{\mathbb{R}^d}\|\nu\|_{\mathbb{R}^d}.
$$

(2)

Let $r_U(z_0) = \sup\{r > 0 : B_r(F(z_0)) \subseteq F(U)\}$.
Let \((\mathcal{M}, g), z \in \mathcal{M}\) be a \(d\) dimensional manifold, \(|\mathcal{M}| = 1\), \((U, F)\) as above. Then \(\exists \kappa > 1\), depending on \(d\), \(c_{\text{min}}, c_{\text{max}}, \|g\|_{\alpha \wedge 1}, \alpha \wedge 1\), and \(C_{\text{count}}\), such that: if \(\rho \leq r_U(z)\), \(\exists i_1, \ldots, i_d\) such that, if
\[
\gamma_l = \left( \int_{B(z, \kappa^{-1} \rho)} \varphi_{i_l}^2 \right)^{-\frac{1}{2}}, \quad l = 1, \ldots, d,
\]

(a) the map \(\Phi : B(z, \kappa^{-1} \rho) \rightarrow \mathbb{R}^d\)
\[
x \mapsto (\gamma_1 \varphi_{i_1}(x), \ldots, \gamma_d \varphi_{i_d}(x))
\]
satisfies for any \(x_1, x_2 \in B(z, \kappa^{-1} \rho)\)
\[
\frac{\kappa^{-1}}{\rho} d_\mathcal{M}(x_1, x_2) \leq \|\Phi(x_1) - \Phi(x_2)\| \leq \frac{\kappa}{\rho} d_\mathcal{M}(x_1, x_2). \quad (3)
\]

(b) \(\kappa^{-1} \rho^{-2} \leq \lambda_{i_1}, \ldots, \lambda_{i_d} \leq \kappa \rho^{-2}\).

(c) \(\gamma_l\) satisfy \(\gamma_1, \ldots, \gamma_d \leq \kappa\).
Recall that $F(z) = \exp\{ -G(z, z_0) - iG^*(z, z_0) \}$, lead to

$$G(z, z_0) = \int_0^\infty K(z, z_0, t) dt,$$

$K$ the (Dirichlet) heat kernel for $\mathcal{D}$, and from

$$K(z, z_0, t) = \sum_{j=1}^{\infty} \varphi_j(z) \varphi_j(z_0) e^{\lambda_j t}$$

we extracted eigenfunctions with large gradient, as suggested by $|F'(z)| = |\nabla G(z, z_0)| e^{-G(z, z_0)} \sim \frac{1}{r}$ on $B(z_0, \frac{r}{2})$.

In fact, this suggest that the heat kernel itself could be used to generate good coordinate charts. Instead of $d$ eigenfunctions we are able to pick $d$ heat kernels $\{K_t(x, y_i)\}_{i=1,\ldots,d}$, and obtain a coordinate chart with similar (in fact, stronger!) guarantees.
Heat triangulation, a picture

\[ K_{\sim R_2^d}(x_i, w) = \text{heat } w \text{ receives from } x_i \text{ at time } R_2^d \sim d(x_i, w) \cdot R_2^d \]

for \( d \) reasonably chosen points \( x_1, \ldots, x_d \).

The heat kernel computes distances by averaging along all paths, weighted by their probability of happening (Wiener measure for Brownian motion), with paths of length \( \sim d(x_i, w) \) having the highest probability.
Heat triangulation, another picture

Note: this can be interpreted as a “kernel map” that linearizes the data to the “largest extent possible” under a distortion constraint.
Theorem (Heat Triangulation Theorem)

Let $(\mathcal{M}, g), z \in \mathcal{M}$ and $(U, F)$ be as above, with the exception we now make no assumptions on the finiteness of the volume of $\mathcal{M}$ and the existence of $C_{count}$. Let $\rho \leq r_U(z)$. Let $p_1, \ldots, p_d$ be $d$ linearly independent directions. There are constants $c > 0$ and $c', \kappa > 1$, depending on $d$, $c_{\text{min}}, c_{\text{max}}, \rho^{\alpha \wedge 1} \| g \|_{\alpha \wedge 1}$, $\alpha \wedge 1$, and the smallest and largest eigenvalues of the Gramian matrix $(\langle p_i, p_j \rangle)_{i,j=1,\ldots,d}$, such that the following holds. Let $y_i$ be so that $y_i - z$ is in the direction $p_i$, with $c\rho \leq d_{\mathcal{M}}(y_i, z) \leq 2c\rho$ for each $i = 1, \ldots, d$ and let $t = \kappa^{-1}\rho^2$. The map

$$x \mapsto (\rho^d K_t(x, y_1)), \ldots, \rho^d K_t(x, y_d)) \quad (4)$$

satisfies, for any $x_1, x_2 \in B(z, \kappa^{-1}\rho)$,

$$\frac{\kappa^{-1}}{c'\rho} d_{\mathcal{M}}(x_1, x_2) \leq \|\Phi(x_1) - \Phi(x_2)\| \leq \frac{\kappa c'}{\rho} d_{\mathcal{M}}(x_1, x_2). \quad (5)$$
Equipped with good systems of coordinates on large pieces of the set, one can start doing analysis and approximation intrinsically on the set.

- **Fourier analysis on data**: use eigenfunctions for function approximation [Belkin, Niyogi, Coifman, Lafon]. Ok for globally uniformly smooth functions - are functions of interest in this class?

- **Diffusion wavelets**: can construct multiscale analysis of wavelet-like functions on the set, adapted to the geometry of diffusion, at different time scales [joint with R. Coifman].

- The *diffusion semigroup* itself on the data can be used as a smoothing kernel. We recently obtained very promising results in image denoising and semisupervised learning [joint with A.D. Szlam and R. Coifman].
Applications

- Hierarchical organization of data and of Markov chains (e.g. documents, regions of state space of dynamical systems, etc...);
- Distributed agent control, Markov decision processes (e.g.: compression of state space and space of relevant value functions);
- Machine Learning (e.g. nonlinear feature selection, semisupervised learning through diffusion, multiscale graphical models);
- Approximation, learning and denoising of functions on graphs (e.g.: machine learning, regression, etc...);
- Sensor networks: compression of measurements collected from the network (e.g. wavelet compression on scattered sensors);
- Multiscale modeling of dynamical systems (e.g.: nonlinear and multiscale PODs);
- Compressing data and functions on the data;
- Data representation, visualization, interaction;
- ...

Harmonic analysis of and on data sets
Let $T = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ as above be the $L^2$-normalized symmetric "random walk". The eigenvalues of $T$ and its powers "typically" look like this:
We now consider a simple example of a Markov chain on a graph with 8 states.

$$T = \begin{pmatrix} 0.80 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.20 & 0.79 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.01 & 0.49 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.499 & 0.001 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.001 & 0.499 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.49 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.49 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 \end{pmatrix}$$

From the matrix it is clear that the states are grouped into four pairs \( \{\nu_1, \nu_2\} \), \( \{\nu_3, \nu_4\} \), \( \{\nu_5, \nu_6\} \), and \( \{\nu_7, \nu_8\} \), with weak interactions between the the pairs.
Some powers of the Markov chain $T$, $8 \times 8$, of decreasing effective rank.

Compressed representations $T_6 := T^{2^6}$ ($4 \times 4$), $T_{13} := T^{2^{13}}$ ($2 \times 2$), and corresponding soft clusters.
We construct multiscale analyses associated with a diffusion-like process $T$ on a space $X$, be it a manifold, a graph, or a point cloud. This gives:

(i) A coarsening of $X$ at different “geometric” scales, in a chain $X \rightarrow X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_j \cdots$;

(ii) A coarsening (or compression) of the process $T$ at all time scales $t_j = 2^j$, $\{T_j = [T^{2^j}]_{\Phi_j}\}_j$, each acting on the corresponding $X_j$;

(iii) A set of wavelet-like basis functions for analysis of functions (observables) on the manifold/graph/point cloud/set of states of the system.

All the above come with guarantees: the coarsened system $X_j$ and coarsened process $T_j$ have “$\epsilon$-close” random walks as $T^{2^j}$ on $X$. This comes at the cost of a very careful coarsening: up to $O(|X|^2)$ operations ($< O(|X|^3)$!), and only $O(|X|)$ in certain special classes of problems.
Construction of Diffusion Wavelets

Diagram for downsampling, orthogonalization and operator compression. (All triangles are $\epsilon$—commutative by construction)
Properties of Diffusion Wavelets

- Multiscale analysis and wavelet transform
- Compact support and estimates on support sizes (not as good as one really would like!);
- Vanishing moments (w.r.t. low-frequency eigenfunctions);
- Bounds on the sizes of the approximation spaces (depend on the spectrum of $T$, which in turn depends on geometry);
- Approximation and stability guarantees of the construction (tested in practice).

One can also construct diffusion wavelet packets, and therefore quickly-searchable libraries of waveforms.
Diffusion Wavelets on Dumbell manifold

Mauro Maggioni

Harmonic analysis of and on data sets
Example: Multiscale text document organization

Scaling functions at different scales represented on the set embedded in $\mathbb{R}^3$ via $(\xi_3(x), \xi_4(x), \xi_5(x))$. $\phi_{3,4}$ is about Mathematics, but in particular applications to networks, encryption and number theory; $\phi_{3,10}$ is about Astronomy, but in particular papers in X-ray cosmology, black holes, galaxies; $\phi_{3,15}$ is about Earth Sciences, but in particular earthquakes; $\phi_{3,5}$ is about Biology and Anthropology, but in particular about dinosaurs; $\phi_{3,2}$ is about Science and talent awards, inventions and science competitions.
Semi-supervised Learning on Graphs

Given:
- \( X \): all the data points
- \( (\tilde{X}, \{\chi_i(x)\}_{x \in \tilde{X}, i=1,...,I}) \): a small subset of \( X \), with labels: \( \chi_i(x) = 1 \) if \( x \) is in class \( i \), 0 otherwise.

Objective:
- guess \( \chi_i(x) \) for \( x \in X \setminus \tilde{X} \).

Motivation:
- data can be cheaply acquired (\( X \) large), but it is expensive to label (\( \tilde{X} \) small). If data has useful geometry, then it is a good idea to use \( X \) to learn the geometry, and then perform regression by using dictionaries on the data, adapted to its geometry.
Algorithm:

- use the geometry of $X$ to design a smoothing kernel (e.g. heat kernel), and apply such smoothing to the $\chi_i$’s, to obtain $\tilde{\chi}_i$, soft class assignments on all of $X$. This is already pretty good.

- The key to success is to repeat: incorporate the $\tilde{\chi}_i$’s into the geometry graph, and design a new smoothing kernel $\tilde{K}$ that takes into account the new geometry. Use $\tilde{K}$ to smooth the initial label, to obtain final classification.

Experiments on standard data sets show this technique is very competitive.
<table>
<thead>
<tr>
<th></th>
<th>FAKS</th>
<th>FAHC</th>
<th>FAEF</th>
<th>Best of other methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>digit1</td>
<td>2.0</td>
<td>2.1</td>
<td>1.9</td>
<td>2.5 (LapEig)</td>
</tr>
<tr>
<td>USPS</td>
<td>4.0</td>
<td>3.9</td>
<td>3.3</td>
<td>4.7 (LapRLS, Disc. Reg.)</td>
</tr>
<tr>
<td>BCI</td>
<td>45.5</td>
<td>45.3</td>
<td>47.8</td>
<td>31.4 (LapRLS)</td>
</tr>
<tr>
<td>g241c</td>
<td>19.8</td>
<td>21.5</td>
<td>18.0</td>
<td>22.0 (NoSub)</td>
</tr>
<tr>
<td>COIL</td>
<td>12.0</td>
<td>11.1</td>
<td>15.1</td>
<td>9.6 (Disc. Reg.)</td>
</tr>
<tr>
<td>gc241n</td>
<td>11.0</td>
<td>12.0</td>
<td>9.2</td>
<td>5.0 (ClusterKernel)</td>
</tr>
<tr>
<td>text</td>
<td>22.3</td>
<td>22.3</td>
<td>22.8</td>
<td>23.6 (LapSVM)</td>
</tr>
</tbody>
</table>

In the first column we chose, for each data set, the best performing method with model selection, among all those discussed in Chapelle’s book. In each of the remaining columns we report the performance of each of our methods with model selection, but with the best settings of parameters for constructing the nearest neighbor graph, among those considered in other tables. The aim of this rather unfair comparison is to highlight the potential of the methods on the different data sets. The training set is $1/15$ of the whole set.
Current related work

- Estimating intrinsic dimensionality of data, in a multiscale fashion (borrowing strength across scales) [with Y.-M. Jung];
- Implementation of the heat triangulation theorem, for “free” data navigation, visualization, and human interaction [with R. Brady and E. Monson];
- Multiscale regression on data and variables, to borrow strength from related (but not necessarily linearly correlated) variables and gain sparsity in the multiscale representation [with S. Mukherjee, S. Lunagomez and J. Guinney].
Acknowledgements

- R.R. Coifman, [Diffusion geometry; Diffusion wavelets; Uniformization via eigenfunctions; Multiscale Data Analysis], P.W. Jones (Yale Math), S.W. Zucker (Yale CS) [Diffusion geometry];
- P.W. Jones (Yale Math), R. Schul (UCLA) [Uniformization via eigenfunctions; nonhomogenous Brownian motion];
- S. Mahadevan (U.Mass CS) [Markov decision processes];
- A.D. Szlam (UCLA) [Diffusion wavelet packets, top-bottom multiscale analysis, linear and nonlinear image denoising, classification algorithms based on diffusion];
- G.L. Davis (Yale Pathology), R.R. Coifman, F.J. Warner (Yale Math), F.B. Geshwind, A. Coppi, R. DeVerse (Plain Sight Systems) [Hyperspectral Pathology];
- H. Mhaskar (Cal State, LA) [polynomial frames of diffusion wavelets];
- J.C. Bremer (Yale) [Diffusion wavelet packets, biorthogonal diffusion wavelets];
- M. Mahoney, P. Drineas (Yahoo Research) [Randomized algorithms for hyper-spectral imaging]
- J. Guinney, S. Lunagomez, J. Mattingly, S. Mukherjee, Q. Wu (Duke Math,Stat,ISDS) [stochastic systems and learning]; A. Lin, E. Monson (Duke Phys.) [Neuron-glia cell modeling]; D. Brady, R. Willett (Duke EE) [Compressed sensing and imaging]

Funding: ONR, NSF, Sloan Foundation

Thank you!

www.math.duke.edu/~mauro
Internet Multi-Resolution Analysis: Foundations, Applications and Practice


For more information:

www.ipam.ucla.edu/programs/mra2008/