Harmonic Analysis on data sets in high-dimensional space

Mauro Maggioni

Mathematics and Computer Science
Duke University


In collaboration with R.R. Coifman, P.W. Jones, R. Schul, A.D. Szlam
Funding: NSF-DMS, ONR.
Plan

- Setting and Motivation
- Diffusion on Graphs
- Eigenfunction embedding
- Multiscale construction
- Examples and applications
- Conclusion
Structured data in high-dimensional spaces

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 a large time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example the set of distributions of word frequencies in a document corpus is low-dimensional, since there are lots of dependencies between the probabilities of word appearances.
Structured data in high-dimensional spaces

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 a large time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example the set of distributions of word frequencies in a document corpus is low-dimensional, since there are lots of dependencies between the probabilities of word appearances.
It has been shown, at least empirically, that in such situations the geometry of the data can help construct useful priors, for tasks such as classification, regression for prediction purposes.

Problems:

- **geometric**: find intrinsic properties, such as local dimensionality, and local parameterizations.
- **approximation theory**: approximate functions on such data, respecting the geometry.
Handwritten Digits

Data base of about 60,000 $28 \times 28$ gray-scale pictures of handwritten digits, collected by USPS. Point cloud in $R^{28^2}$. 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.
A simple example from Molecular Dynamics

[Joint with C. Clementi]

The dynamics of a small protein (22 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 \( \mathbb{R}^{66} \).

Left and center: \( \phi \) and \( \psi \) are two backbone angles, color is given by two of our parameters obtained from the geometric analysis of the set of configurations. Right: embedding of the set of configurations.
Goals

This is a regime for analysis quite different from that discussed in most talks. We think it is useful to tackle it by analyzing both the intrinsic geometry of the data, and then working on function approximation on the data (and then repeat!).

- 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 be at least approximately an isometry with respect to the manifold distance, and finally it should be stable under perturbations of the manifold. In the examples above: variations in the handwritten digits, topics in the documents, angles in molecule...

- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.
Goals

This is a regime for analysis quite different from that discussed in most talks. We think it is useful to tackle it by analyzing both the intrinsic geometry of the data, and then working on function approximation on the data (and then repeat!).

- 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 be at least approximately an isometry with respect to the manifold distance, and finally it should be stable under perturbations of the manifold. In the examples above: variations in the handwritten digits, topics in the documents, angles in molecule...

- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.
Random walks and heat kernels on the data

Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

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. Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W,$$  \hspace{1cm} \text{random walk}

$$T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}},$$  \hspace{1cm} \text{symm. “random walk”}

$$H = e^{-t(I-T)}$$  \hspace{1cm} \text{Heat kernel}

Note 1: $K$ typically depends on the type of data.
Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Random walks and heat kernels on the data

Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

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. Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W,$$  \hspace{1cm}  $$T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}},$$  \hspace{1cm}  $$H = e^{-t(I-T)}$$

random walk \hspace{5cm} symm. “random walk” \hspace{5cm} Heat kernel

Note 1: $K$ typically 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 \( K(x_i, x_j) \geq 0 \).

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

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

random walk \hspace{2cm} symm. “random walk” \hspace{2cm} Heat kernel

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

Note 2: \( K \) should be “local”, i.e. close to 0 for points not sufficiently close.
When \( n \) points are randomly sampled from a Riemannian manifold \( \mathcal{M} \), uniformly w.r.t. volume, then the behavior of the above operators, as \( n \to +\infty \), is quite well understood. In particular, \( T \) approximates the heat kernel on \( \mathcal{M} \), and \( \mathcal{L} = I - T \), the normalized Laplacian, approximates (up to rescaling), the Laplace-Beltrami operator on \( \mathcal{M} \).

These approximations should be taken with a grain of salt: typically the number of points is not large enough to guarantee that the discrete operators above are close to their continuous counterparts.
One can try to use the operator $T$, or its eigenfunctions, which is intrinsic to the data, to construct parametrizations of the data. This is indeed possible; in fact, we [P.W. Jones, R. Schul, MM] showed one can obtain even better parametrizations by using $T$ itself, or heat kernels.

When the data is nonlinear, these embedding are more powerful, and have stronger guarantees, and wider applicability, when $\mathcal{M}$ is nonlinear, of both standard linear embeddings (PCA, random projections,...) and nonlinear embeddings (ISOMAP, LLE, Hessian eigenmap, etc...).
Theorem (Heat Triangulation Theorem - with P.W. Jones, R. Schul)

Let \((\mathcal{M}, g)\) be a Riemannian manifold, with \(g\) at least \(C^\alpha, \alpha > 0\), and \(z \in \mathcal{M}\). Let \(R_z\) be the radius of the largest ball on \(\mathcal{M}\), centered at \(z\), which is bi-Lipschitz equivalent to a Euclidean ball. Let \(p_1, \ldots, p_d\) be \(d\) linearly independent directions. There are constants \(c_1, \ldots, c_5 > 0\), depending on \(d\), \(c_{\text{min}}, c_{\text{max}}\), \(\|g\|_{\alpha \wedge 1}, \alpha \wedge 1\), and the smallest and largest eigenvalues of the Gramian matrix \((\langle p_i, p_j \rangle)_{i=1, \ldots, d}\), such that the following holds. Let \(y_i\) be so that \(y_i - z\) is in the direction \(p_i\), with \(c_4 R_z \leq d_{\mathcal{M}}(y_i, z) \leq c_5 R_z\) for each \(i = 1, \ldots, d\) and let \(t_z = c_6 R_z^2\). The map

\[
\Phi : B_{c_1 R_z}(z) \rightarrow \mathbb{R}^d
\]

\[
x \mapsto (R_z^d K_{t_z}(x, y_1)), \ldots, R_z^d K_{t_z}(x, y_d))
\]

satisfies, for any \(x_1, x_2 \in B_{c_1 R_z}(z)\),

\[
\frac{c_2}{R_z} d_{\mathcal{M}}(x_1, x_2) \leq \|\Phi(x_1) - \Phi(x_2)\| \leq \frac{c_3}{R_z} d_{\mathcal{M}}(x_1, x_2).
\]
Currently working at the discrete analogue, and implementation, of the above. Obstacles we are overcoming:

- Intrinsic dimension $d$ unknown. Tools to overcome: dimension estimation through multiscale local PCA and/or through multiscale heat kernel time decay.
- $R_z$ is unknown. Tools to overcome: forget about finding $R_z$! Greedily find the largest ball around $z$ for which the heat kernel triangulation works. Theorem guarantees it will work at least on a large ball.
- Computational cost. Want linear in $n$, this is not trivial if heat kernels at “medium time” (say, $\sqrt{n}$) are needed. Tools: multiscale analysis of the heat kernel.
- Discrete data sets have a geometry which is more complicated than $C^{1+\alpha}$ manifolds: dimensionality changes from point to point, strange things can happen to eigenfunctions and heat kernels, etc...
An example from Molecular Dynamics - revisited

Small protein (22 atoms, H atoms removed) in water: each configuration is a 66-dimensional vector. Want to study the geometry of the effective state space visited by the protein during long simulations. Define similarity between two configurations $x_i, x_j$, based on Euclidean distance modulo rotations+translations. Assuming a Langevin equation $\dot{x} = -\nabla U(x) + \dot{w}$, one can construct a weight on the edges of the graph in such a way that the random walk $P$ on the graph is a discrete approximation to forward (or backward) Fokker-Planck operator describing the propagation of probability distributions under the Langevin SDE.
Analysis on the set

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. Conjecture: most functions of interest are not 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 (in a few slides, 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;
- ...

Mauro Maggioni
Harmonic Analysis on data sets in high-dimensional space
Suppose for simplicity we have a weighted graph \((G, E, W)\), with corresponding Laplacian \(\mathcal{L}\) and random walk \(P\). Let us renormalize, if necessary, \(P\) so it has norm 1 as an operator on \(L^2\): let \(T\) be this operator. Assume for simplicity that \(T\) is self-adjoint, and high powers of \(T\) are low-rank: \(T\) is a diffusion, so range of \(T^t\) is spanned by smooth functions of increasingly (in \(t\)) smaller gradient.

A “typical” spectrum for the powers of \(T\) would look like this:
A Multi-Resolution Analysis for $L^2(\mathbb{R})$ is a sequence of subspaces $\{V_j\}_{j \in \mathbb{Z}}$, $V_{j-1} \subseteq V_j$, with $\cap V_j = \{0\}$, $\cup V_j = L^2(\mathbb{R})$, with an orthonormal basis $\{\varphi_{j,k}\}_{k \in \mathbb{Z}} := \{2^{j/2} \varphi(2^j \cdot -k)\}_{k \in \mathbb{Z}}$ for $V_j$. Then there exist $\psi$ such that $\{\psi_{j,k}\}_{k \in \mathbb{Z}} := \{2^{j/2} \psi(2^j \cdot -k)\}_{k \in \mathbb{Z}}$ spans $W_j$, the orthogonal complement of $V_{j-1}$ in $V_j$. $\hat{\varphi}_{j,k}$ is essentially supported in $\{|\xi| \leq 2^j\}$, and $\hat{\psi}_{j,k}$ is essentially supported in the L.P.-annulus $2^{j-1} \leq |\xi| \leq 2^j$. Because $V_{j-1} \subseteq V_j$, $\varphi_{j-1,0} = \sum_{k'} \alpha_{k'} \varphi_{j,k'}$: refinement eqn.s, FWT.

We would like to generalize this construction to graphs. The frequency domain is the spectrum of $e^{-\mathcal{L}}$. Let $V_j := \langle \{\phi_i : \lambda_i^{2^j} \geq \epsilon\} \rangle$. Would like o.n. basis of well-localized functions for $V_j$, and to derive refinement equations and downsampling rules in this context.
Construction of Diffusion Wavelets

Mauro Maggioni

Harmonic Analysis on data sets in high-dimensional space
Multiscale Analysis

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 $t_j = 2^j$, $\{T_j\}$, 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, in the sense that the coarsened system $X_j$ and corresponding coarsened process $T_j$ behave exactly as $T^{2^j}$ on $X$. The guarantee come of course at the cost of a very careful coarsening procedure. In general it can take up to $O(|X|^2)$ operations, and only $O(|X|)$ in certain classes of problems (e.g. diffusion on nice manifolds).
From left to right: function $F$; reconstruction of the function $F$ with top 50 best basis packets; reconstruction with top 200 eigenfunctions of the Beltrami Laplacian operator.

Left to right: 50 top coefficients of $F$ in its best diffusion wavelet basis, distribution coefficients $F$ in the delta basis, first 200 coefficients of $F$ in the best basis and in the basis of eigenfunctions.
Diffusion Wavelets on Dumbell manifold

Harmonic Analysis on data sets in high-dimensional space
Other constructions:
- Biorthogonal diffusion wavelets, in which scaling functions are probability densities (useful for multiscale Markov chains)
- Top-bottom constructions: recursive subdivision
- Both...

Applications:
- Document organization and classification
- Markov Decision Processes
- Nonlinear Analysis of Images
- Semi-supervised learning through diffusion processes on data
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.
Some example of scaling functions on the documents, with some of the documents in their support, and some of the words most frequent in the documents.
Semi-supervised Learning on Graphs

[Joint with A.D.Szlam]

Given:
- \( X \): all the data points
- \((\tilde{X}, \{\chi_i(x)\}_{x \in \tilde{X}, i=1,\ldots,I})\): 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.
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.

<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>
Nonlinear image denoising

An image is sometimes modeled as a function on $[0, 1]^2$ (or $[0, 1]^3$ if hyperspectral). Well understood that spatial relationships are important, but not the only thing: there are interesting features (e.g. edges, textures), at different scales. Naive idea: apply heat propagation on $[0, 1]^2$ to the image...:( Better idea: do not use simple linear smoothing, but anisotropic/nonlinear smoothing, in order to preserve important structures (mainly edges). Process is image-dependent! We propose: do not work on $[0, 1]^2$, but in a space of features of the image. Map
\[ \Psi(x, y) \in Q \rightarrow (x, y, (l \ast g_1)(x, y), \ldots, (l \ast g_m)(x, y)) \subset \mathbb{R}^{m+2}, \]
and denoise $l$ as a function on $\Psi(Q)$, with the heat kernel on $\Psi(Q)$.

[Joint with A.D.Szlam]
Nonlinear image denoising, I

Left to right: 1) a clean image, with range from 0 to 255. 2) A noisy image obtained by adding Gaussian noise $\mathcal{N}(0, 1)$. 3) TV denoising kindly provided by Guy Gilboa. 4) Denoising using a diffusion built on the graph of $5 \times 5$ patches, with a constrained search.
1) Lena with Gaussian noise added. 2) Denoising using a 7x7 patch graph. 3) Denoising using hard thresholding of curvelet coefficients. The image is a sum over 9 denoisings with different grid shifts. 4) Denoising with a diffusion built from the 9 curvelet denoisings.
Left to right: 1) Barbara corrupted by Gaussian noise $40N(0, 1)$ from 0 to 255. 2) Denoising using a diffusion built on the graph of $7 \times 7$ patches, with a constrained search.
Challenge: what is the geometry of patches in an image?

Set of patches from Barbara projected onto low-dimensions by using PCA. The color is equal to the pixel intensity at the center of the patch.
Figure: Poisson corrupted image (left) denoised using the patch graph (right).
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. Mattingly, S. Mukherjee and 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: NSF, ONR.

Thank you!

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


For more information:

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