Towards Multiscale Harmonic Analysis of Graphs and Document Corpora

Mauro Maggioni

Diffusion geometry group at Yale:

R.R. Coifman, P. Jones, A.D. Szlam, J. Bremer,

A. Singer, F. Warner, S.W. Zucker

Jan. 26th, 2006

Overview

1 - Functions *on* the data, parametrizations, embeddings, and features;

2 - Connections with classical signal processing;

3 - Fourier global analysis;

4 - Multiscale analysis;

5 - Examples.

Functions *on* the data

We are presented with a data set, modeled for example as a point cloud in $\mathbb{R}^n$, or a graph. Most (all?) the questions about this data sets and tasks we want to perform with it can be cast in terms of analysis of functions *on* the data set.

For example:

- Semi-supervised learning: some data points are labeled as belonging to certain classes, and would like to assign labels to non-labeled points. Well, given labels $\{L_j\}_{j=1,\ldots,J}$ on $\tilde{X} \subseteq X$, I can consider the J functions $\{\chi_{L_j}\}_{j=1,\ldots,J}$ on $\tilde{X}$ defined by $\chi_{L_j}(\tilde{x}) = 1$ if $\tilde{x} \in \tilde{X}$ was labeled as L_j, and 0 otherwise. The question is about "extending" or "interpolating" the functions χ_{L_j} at all the points in X.
- Given a preference function for a subset of the points $\tilde{X}$ (e.g. preference function on a set of movies/products/music/...), predict how much I am likely to prefer a point $X \setminus \tilde{X}$: again an "extension" or "interpolation" problem, for the preference function.

Parametrizations

A parametrization on a subset of the data provides a mapping from data points to $\mathbb{R}^d$, with good properties (e.g. small distortion). "It locally re-orders the point in an Euclidean way", in such a way that Euclidean constructions can be applied to the portion of the data being parametrized. It also reveals intrinsic dimensionality of the data.

Example of techniques: local principal component analysis, Jones' β-numbers for determining dimensionality; isomap, LLE, Laplacian eigenmaps, LTSA, and several others for nonlinear parametrization.

Parameters found with the above techniques may (or not) be interpreted as a physical/probabilistic/... variables which governing the structure of the problem. This is of course not guaranteed, since the algorithms above do not know what interpretability means.

Embeddings

Here I use this as a synonym of parametrization, maybe with an emphasis on the metric and distortion properties. For example the emphasis may be to map from a metric space into Euclidean space (maybe low-dimensional), with small distortion: this is the goal of classical multidimensional scaling (see M. Trosset's talk) and its generalizations and particularizations, such as approximate multiscale (Bourgain) and measured descent embeddings (Assaf Naor et al.), randomized projections (Lindenstrauss), hierarchical tree approximations (Bartal), etc...

Connections with classical signal processing

The above justifies the need to be able to perform analysis, regularization, denoising, interpolation and extrapolation of functions on the data. Of course we do this all the time for functions defined on one-dimensional (e.g. sounds) or low-dimensional (images, movies, hyper-spectral images...) functions. Many successful algorithms in low-dimension are often based on the availability of good basis functions: in particular the Fourier basis and wavelet bases have been proven to be excellent tools. They have also provided extremely useful in harmonic analysis, pure and applied, for example they lead to state-of-the-art solvers for PDEs certain very important classes of integral equations. [More about this on my talk in the Math Department on Monday!]. Main idea: write

$$f = \sum_k \alpha_k \psi_k$$

where ψ_k are the basis elements (this is just a change of coordinates!), then work on the new coordinates α_k. When is this change of basis useful? If tasks such as denoising, compression or others (e.g. evolving a PDE) become **much** easier in these coordinates.

One slide about Fourier analysis

(i) Fourier: approximate solutions of the heat equation on the interval $[0, \pi]$ (or rectangle) with sine and cosine functions: $\phi_k(x) = \sin(kx)$.

(ii) Fourier on Euclidean domains: instead of sines and cosines need the eigenfunctions of the Laplacian on the domain: ϕ_k :

$$\Delta\phi_k = \lambda_k\phi_k \, .$$

(iii) Fourier on manifolds: as above, with the natural Laplace-Beltrami operator on the manifold.

The good and the bad: FFT in some special cases; ϕ_k's are global approximants, and α_k are not as sparse as one may wish, hence denoising and compression are far from optimal.

Laplacian on manifolds

The Laplace-Beltrami operator Δ_{BL} can be defined naturally on a Riemannian manifold, and is a well-studied object in differential geometry and global analysis. The corresponding heat kernel $e^{-t\Delta}$ is the Green's function for the heat equation on the manifold, associated with Brownian motion "restricted" to the manifold. The spectral decomposition $\Delta\phi_i = \lambda_i\phi_i$ yields

$$H_t(x,y) := e^{-t\Delta}(x,y) = \sum_i \underbrace{e^{-t\lambda_i}}_{\mu_i^t} \phi_i(x)\phi_i(y)\,.$$

The eigenfunctions ϕ_i of the Laplacian generalize Fourier modes: Fourier analysis on manifolds, global analysis.

Surprisingly, they also allow to analyse the geometry of the manifold, and provide embeddings of the manifold ("eigenmaps" or "diffusion maps"): for $m = 1, 2, \ldots,$ $t > 0$ and $x \in \mathcal{M}$, define

$$\Phi_m^{(t)}(x) = (\mu_i^{\frac{t}{2}}\phi_i(x))_{i=1,\ldots,m} \in \mathbb{R}^n\,.$$

Laplacian on Graphs

Given a weighted graph (G, W, E), the combinatorial Laplacian is defined by $L = D - W$, where $(D)_{ii} = \sum_j W_{ij}$, and the normalized Laplacian is defined by

$$\mathcal{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}} .$$

These are self-adjoint positive-semi-definite operators, let λ_i and ϕ_i be the eigenvalues and eigenvectors. Fourier analysis on graphs. The heat kernel is of course defined by $H_t = e^{-t\mathcal{L}}$; the natural random walk is $D^{-1}W$.

If

- points are drawn from a manifold according to some (unknown) probability distribution [M. Belkin, P. Niyogi; RR Coifman, S. Lafon], or
- points are generated by a stochastic dynamical system driven by a Fokker-Planck equation [RR Coifman, Y. Kevrekidis, S. Lafon, B. Nadler]

the Laplace-Beltrami operator and, respectively the Fokker-Planck operator, can be approximated by a graph Laplacian on the discrete set of points with certain appropriately estimated weights.

Graph associated with data sets

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

Assume we know how to assign local similarities: map data set to weighted graph. Global distances are not to be trusted!

Data often given as points in high-dimension, but constraints (natural, physical...) force it to be *intrinsically low-dimensional*.

Model the data as a *weighted graph* (G, E, W): vertices represent data points (correspondence could be stochastic), edges connect similar data points, weights represent a similarity measure. Example: have an edge between web pages connected by a link; or between documents with very similar word frequencies. When points are in high-dimensional Euclidean space, weights may be a function of Euclidean distance, and/or the geometry of the points. How to define the similarity between very similar objects in each category is important but not always easy. That's the place where field-knowledge goes.

Weights from a local similarity kernel

The similarity between points of a set X can be summarized in a kernel $K(x,y)$ on $X \times X$. Usually we assume the following properties of K:

$$K(x,y) = K(y,x) \qquad \text{(symmetry)}$$
$$K(x,y) \geq 0 \qquad \text{(positivity} - \text{preserving)}$$
$$\langle v, Kv \rangle \geq 0 \qquad \text{(positive semi} - \text{definite)}$$

If $X \subseteq \mathbb{R}^n$, then choices for K include $e^{-(\frac{||x-y||}{\delta})^2}$, $\frac{\delta}{\delta+||x-y||}$, $\frac{\langle x,y \rangle}{||x||||y||}$.

If some "model" for X is available, the kernel can be designed to be consistent with that model.

In several applications, one starts by applying a map to X (projections onto lower-dimensional subspaces, nonlinear maps, etc...) before constructing the kernel.

Eigenfunctions of the Laplacian on data sets

We have already seen in other talks that eigenfunctions of the Laplacian can be used as:

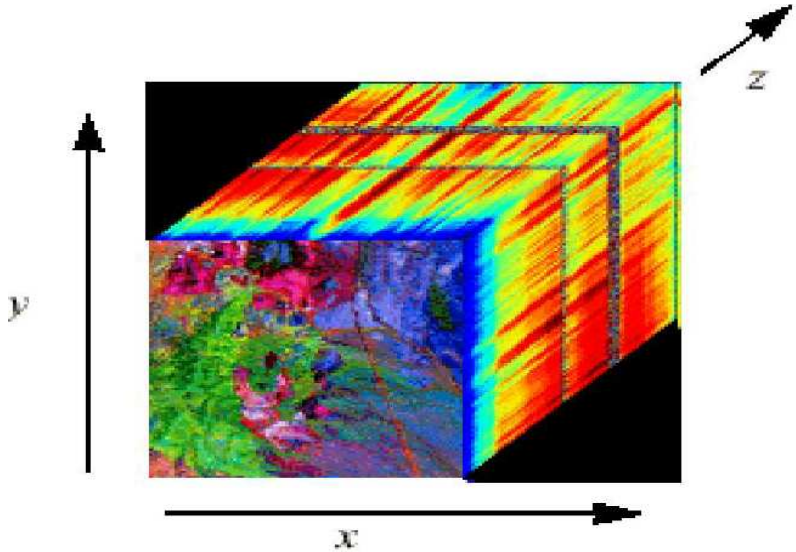
- Embedding, as a particular case of (kernelized) MDS (M. Trosset's talk)
- Local coordinate systems with good distortion properties (P. Jones' talk)
- Study large-time diffusion on the graph and diffusion distances between points on the graph (R. Coifman, S. Lafon, A. Tomkins's talks)

These are useful for further tasks in semi-supervised learning, classification and clustering, as shown by several people, working on several different data sets. For example: image segmentation (Shi-Malik), classifiers in the semi-supervised learning context [M. Belkin-P. Nyogi], fMRI data [F. Meyer, X. Shen], art data [W Goetzmann, PW Jones, MM, J Walden], hyperspectral Imaging in Pathology [MM, GL Davis, F Warner, F. Geshwind, A Coppi, R. DeVerse, RR Coifman], molecular dynamics simulations [RR. Coifman, G.Hummer, I. Kevrekidis, S. Lafon, MM, B. Nadler], text documents classification [RR. Coifman, S. Lafon, A. Lee; RR Coifman, MM].

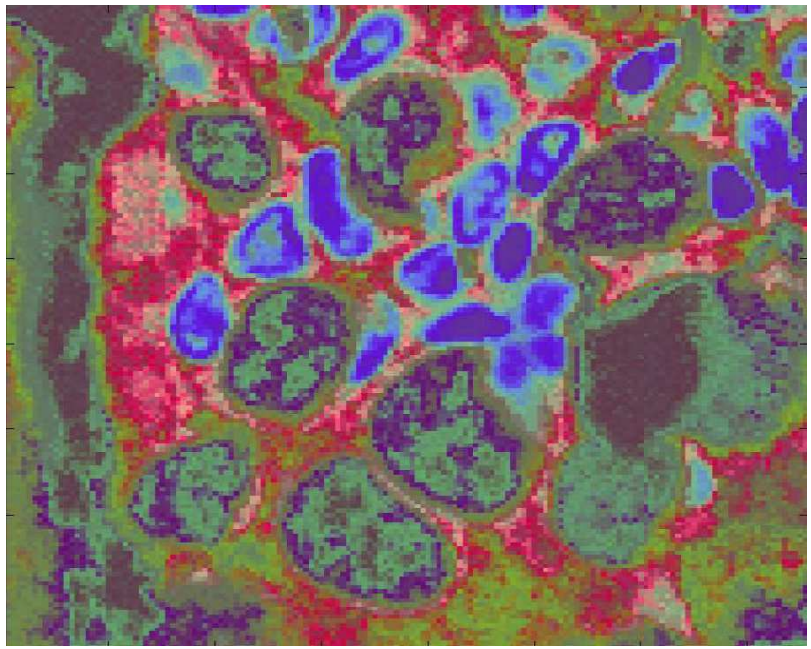
Application to Hyper-spectral Pathology

For each pixel in a hyper-spectral image we have a whole spectrum (a 128-dimensional vector for example). We view the ensemble of all spectra in a hyper-spectral image as a cloud in $\mathbb{R}^{128}$, induce a Laplacian on the point set and use the eigenfunctions for classification of spectra into different classes, which turn out to be biologically distinct and relevant.

On the left, we have mapped the values of the top 3 eigenfunctions to RGB.

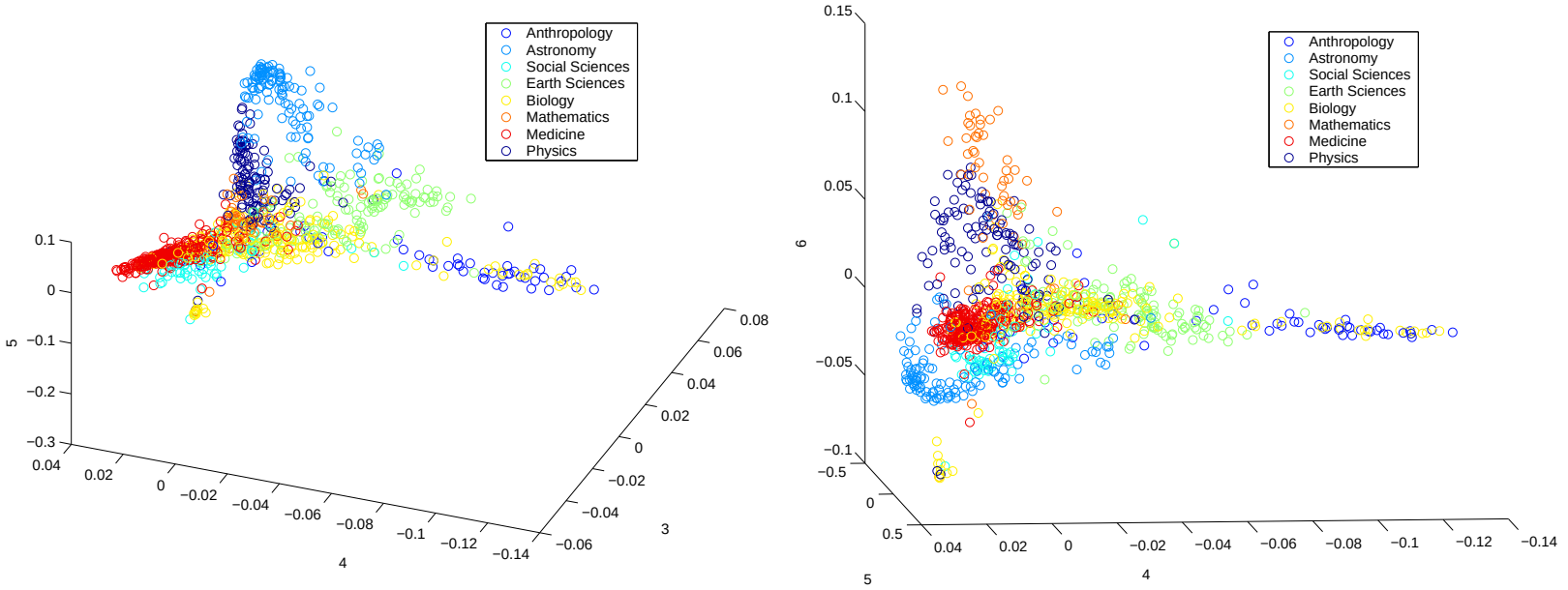


Hyperspectral data cube.
(DataFusionCorp.)



An example of a text document corpus

Consider about 1000 Science News articles, from 8 different categories. For each we compute about 10000 coordinates, the i-th coordinate of document d representing the frequency in document d of the i-th word in a fixed dictionary. The diffusion map gives the embedding below.

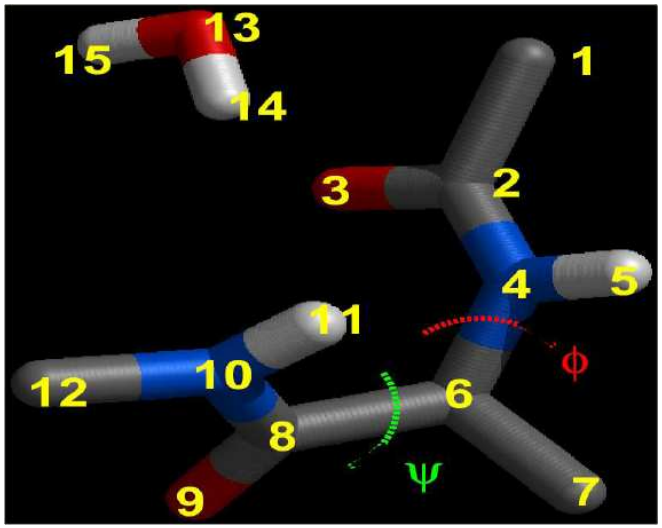


Embedding $\Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x))$: on the left coordinates $3, 4, 5$, and on the right coordinates $4, 5, 6$.

An example from Molecular Dynamics...

The dynamics of a small protein ini a bath of water molecules is approximated by a Fokker-Planck system of stochastic equations $\dot{x} = -\nabla U(x) + \dot{w}$.

The set of states of the protein is a noisy set of points in $\mathbb{R}^{36}$, since we have 3 coordinates for each of the 12 atoms. This set is a priori very complicated. However we expect for physical reasons that the constraints on the molecule to force this set to be essentially lower-dimensional. We can explore the space of states by running long simulations, for different initial conditions.

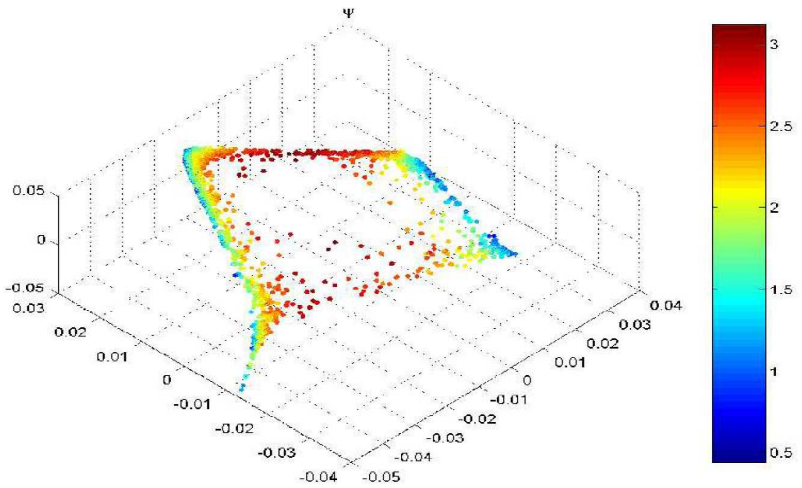


The alanine molecule

Example from Molecular Dynamics revisited

The dynamics of a small protein in a bath of water molecules is approximated by a Fokker-Planck system of stochastic equations $\dot{x} = -\nabla U(x) + \dot{w}$. Many millions of points in $\mathbb{R}^{36}$ can be generated by simulating of the stochastic ODE, U is needed only "on the fly" and only at the current positions (not everywhere in $\mathbb{R}^{36}$).

Then a graph
Laplacian on this set of points can be constructed, that approximated the Fokker-Planck operator, and the eigenfunctions of this approximation yield a low-dimensional description and parametrization of the set, as well as a subspace in which the long-term behavior of the system can faithfully projected.



Embedding of the set of states of the molecule.

Shortcomings of Fourier Analysis

It is good for analysis of globally uniformly smooth functions, bad for analyzing transient phenomena, singularities. This is mainly because the basis elements have global support: non-zero everywhere (compare with principal components!), so each of them is sensitive to even a local transient in the function.

In general we do not expect that the functions we are interested in are uniformly smooth. For example, document space (as a family of mappings) is an elephant (see D. Marchette's talk).

Remedies: try to *localize* the Fourier basis! This can be achieved in many ways: apply window functions (Gabor), local cosine bases (Coifman-Meyer). These bases tend to be localized at a given scale: how to pick this scale? What if the function has transients at different scales? Need basis elements at all scales, nicely fitting together: wavelets (Grossman-Morlet, Daubechies, Coifman, Meyer,...).

One slide about wavelets

Wavelets are concentrated both in time and frequency. Wavelets have two indices: $\phi_{j,k}$ is an "atom" concentrated in time at position k, width about 2^{-j}, and concentrated around frequency 2^j. They provide essentially the best possible building blocks for interesting and large classes of functions, much fewer α_k's for the representation of these functions in a wavelet basis are needed.

Initially constructed on $\mathbb{R}$ (late 80's), then on $\mathbb{R}^n$, and constructions on meshed surfaces (graphics, PDEs).

We will talk about a recent general construction on graphs and manifolds, called diffusion wavelets [Coifman, MM].

Show example of approximation of discontinuous function!

Multiscale Analysis, I

We would like to construct multiscale bases, generalizing classical wavelets, on manifolds, graphs, point clouds.

The classical construction is based on geometric transformations (such as dilations, translations) of the space, transformed into actions (e.g. via representations) on functions. There are plenty of such transformations on $\mathbb{R}^n$, certain classes of Lie groups and homogeneous spaces (with automorphisms that resemble "anisotropic dilations"), and manifolds with large groups of transformations.

Here the space is in general highly *non-symmetric*, not invariant under "natural" geometric transformation, and moreover it is *"noisy"*.

Idea: use *diffusion and the heat kernel as dilations*, acting on functions on the space, to generate multiple scales.

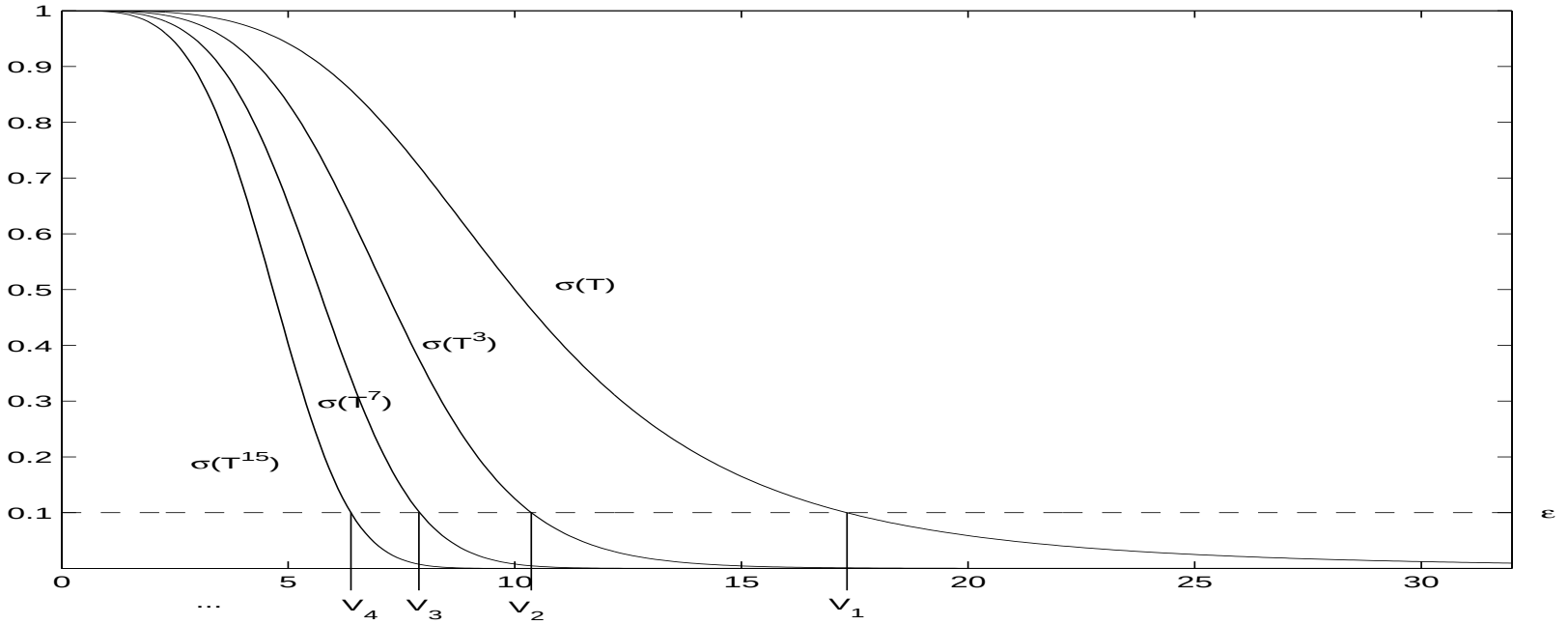
This is connected with the work on diffusion or Markov semigroups, and Littlewood-Paley theory of such semigroups (a la Stein).

We would like to have *constructive* methods for efficiently computing the multiscale decompositions and the wavelet bases.

Multiscale Analysis, II

Suppose for simplicity we have a weighted graph (G, E, W), with corresponding Laplacian $\mathcal{L}$. Let $T = I - \mathcal{L}$. T is self-adjoint, and assume that 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:



Multiscale Analysis, III

The idea is to let V_j be the range of T^{2^j-1}, so that clearly $V_{j+1} \subseteq V_j$. We would like a localized basis of *scaling functions* for V_j.

Consider the matrix T^{2^j-1} and factor it as

$$T^{2^j-1} \sim_\epsilon B_j \tilde{T}_j, \qquad B_j \text{ is } N \times N_j, \tilde{T}_j \text{ is } N_j \times N$$

Columns of B_j are basis vectors, and $\tilde{T}_j$ are the coefficients of the columns of T^{2^j-1} written in terms of B_j. For example:

- B_j could be a localized orthonormal basis (want a *sparse* QR factorization);
- B_j could be a subset of columns of T^{2^j-1}, whose existence is guaranteed by the Gram-Schmidt algorithm (or, better, Gu-Eisenstat). In this latter case the entries of B_j are nonnegative, and the basis vectors are nice local bump functions at scale j.

This is a *compression* step: the spectrum of T decays, so N_j decreases with j!! This is computationally intensive: T^{2^j-1} fills up quite rapidly!

Interpretation: at every scale pick a set of representative of random walkers. We expect them to be "meaningful" communities or related documents...a "topic".

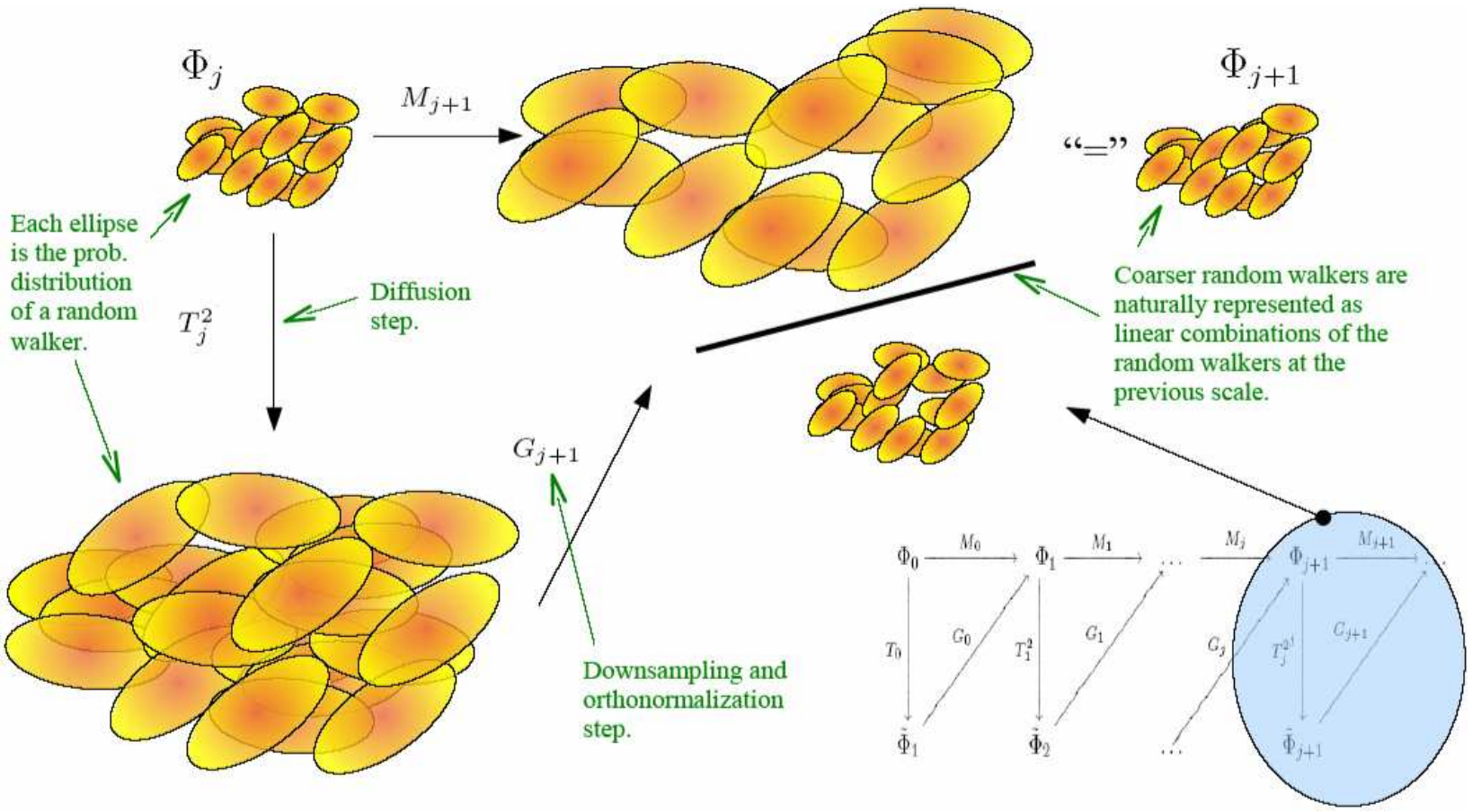
One step further: a fully multiscale construction.

Factor $T = B_1 \tilde{T}_1$ as above, then write $\tilde{T}_1$ with respect to the basis B_1 also in the range, obtaining a *square* matrix T_1:

$$T_1 = \tilde{T}_1 B_1^{\dagger} .$$

This means that T_1 now writes the result of diffusing a column of B_1 in terms of the basis B_1, instead that the original basis. Then to go the next scale, look at T_1^2, and factor this to obtain B_2 and $\tilde{T}_2$, write T_2 with respect to B_2 in the range, square T_2, and so on! We represent the diffusion at each scale (i.e. T^{2^j}) in terms of the representative random walkers B_j at that scale, both in the domain and the range.

Construction of (orthonormal) Diffusion Wavelets



Summary of the Algorithm

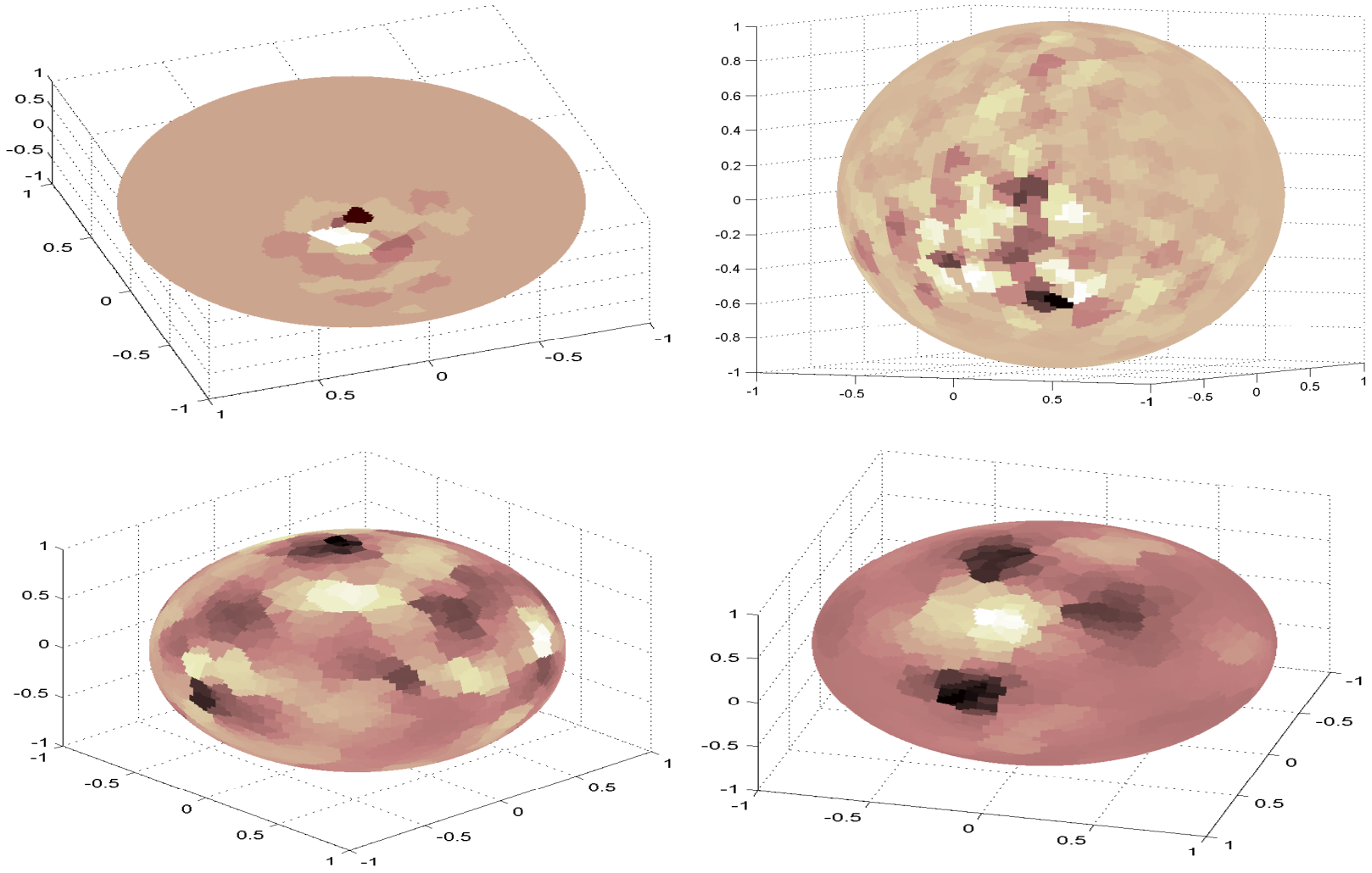
Input: A diffusion operator represented on some orthonormal basis (e.g.: δ-functions), and a precision ϵ. If you start with a document-word matrix, construct a document graph and let the diffusion operator be the random walk on that graph.

Output: Multiscale orthonormal or biorthogonal scaling function bases Φ_j and wavelet bases Ψ_j, encoded through the corresponding multiscale filters M_j, **as well as T^{2^j} represented (compressed) on Φ_j.**

Pictorially, each basis function in Φ_j is a "bump" at scale j, and T^{2^j} represented on this basis is the diffusion among these bumps, at time scale 2^j. One can also think of each Φ_j as a point, and T^{2^j} as a diffusion on these points, and so this construction gives multiscale representations of the graph itself.

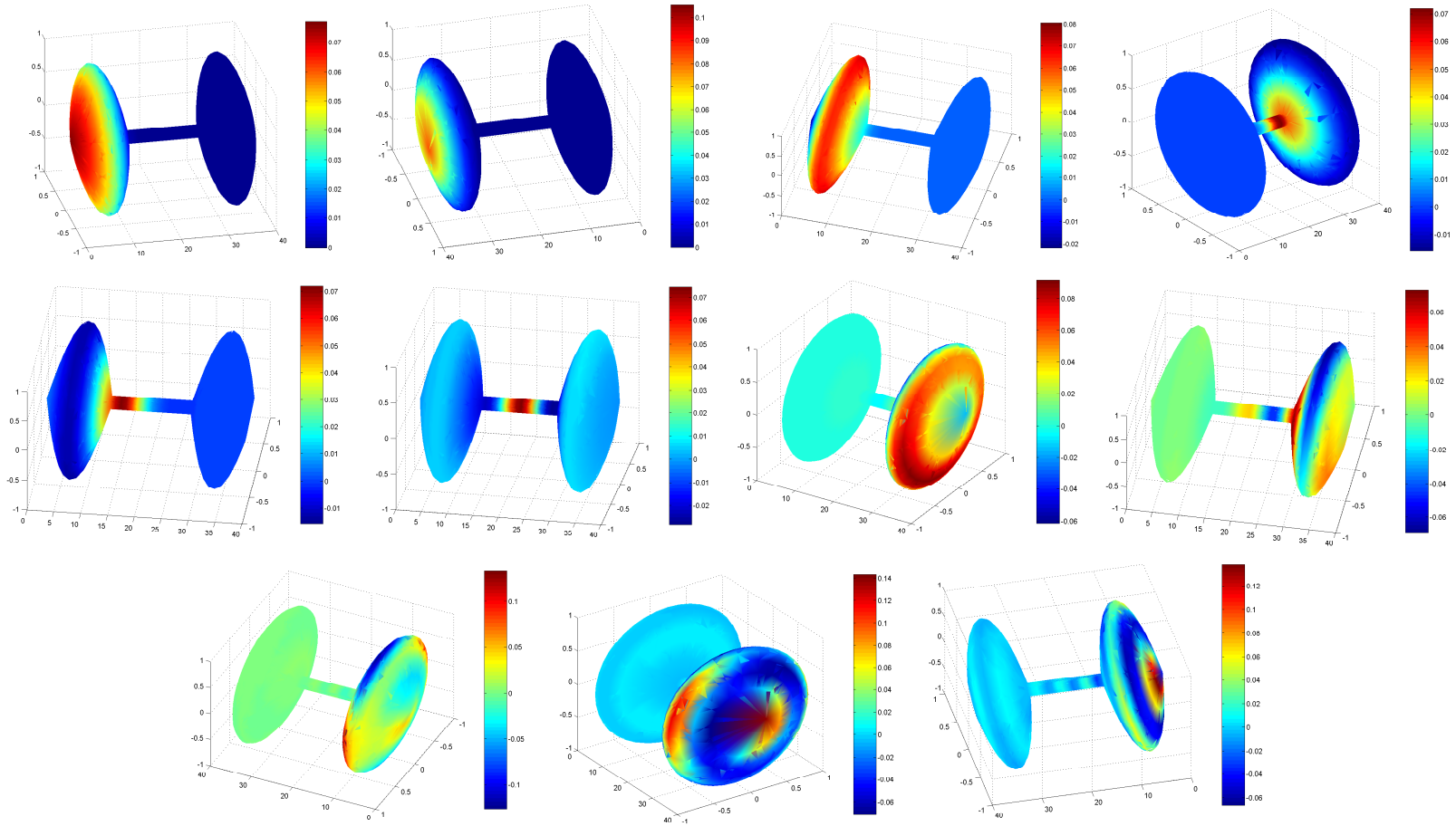
This allows to construct multiscale features on the set, obtain descriptions of the set at different scales. It also allows for a fast wavelet transform, best basis algorithms, signal processing on graphs and manifolds, efficient application of T^{2^j}, and direct inversion of the Laplacian.

Diffusion Wavelets on the Sphere

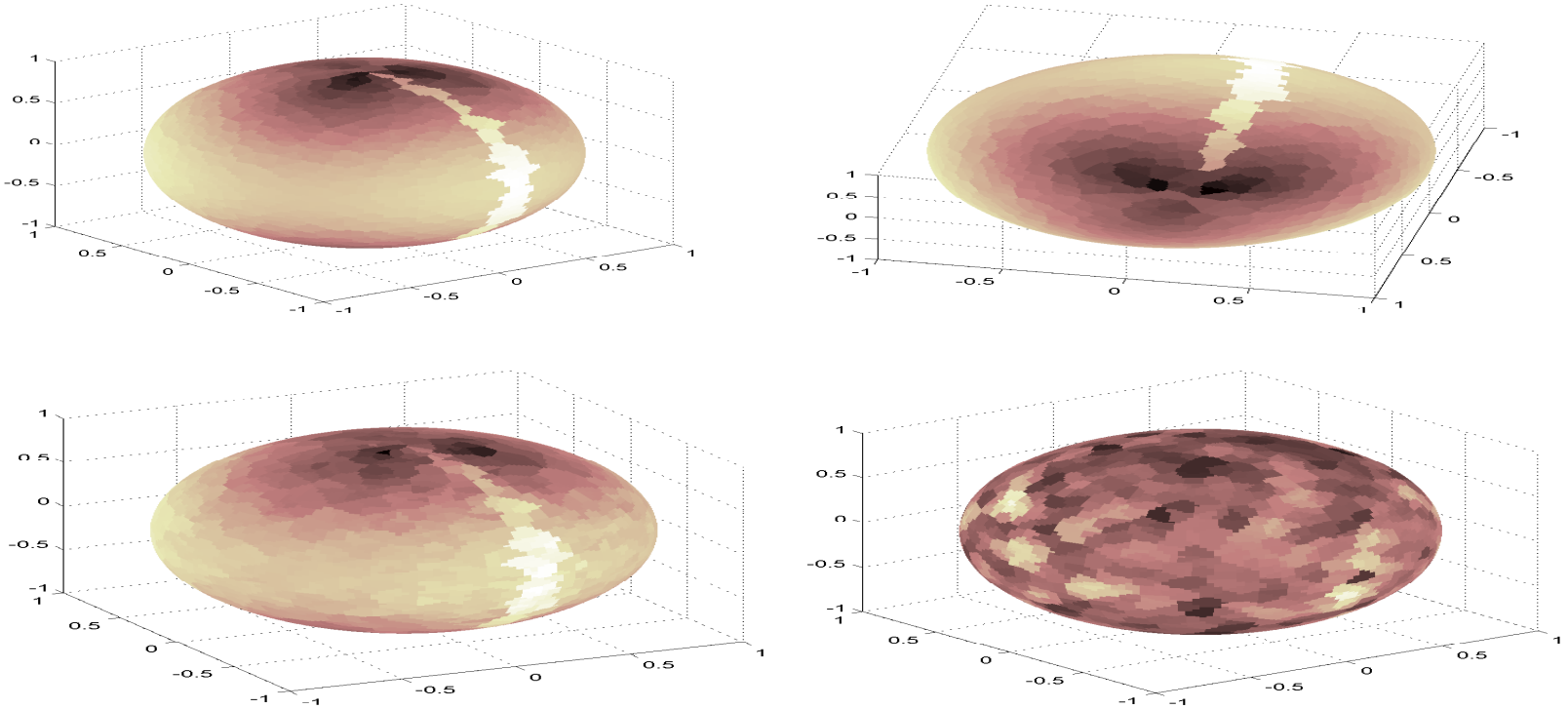


Some diffusion wavelets and wavelet packets on the sphere, sampled randomly uniformly at 2000 points.

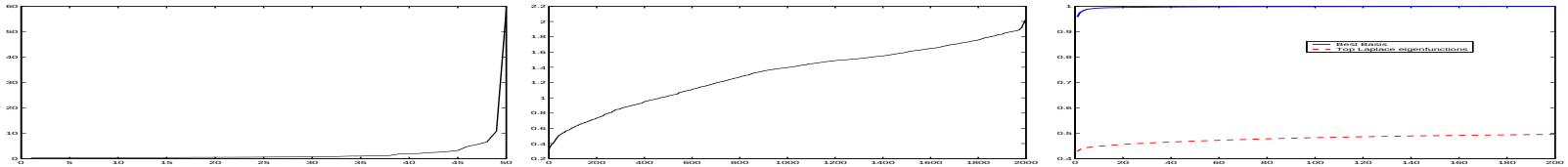
Diffusion Wavelets on Dumbell manifold



Signal Processing on Manifolds

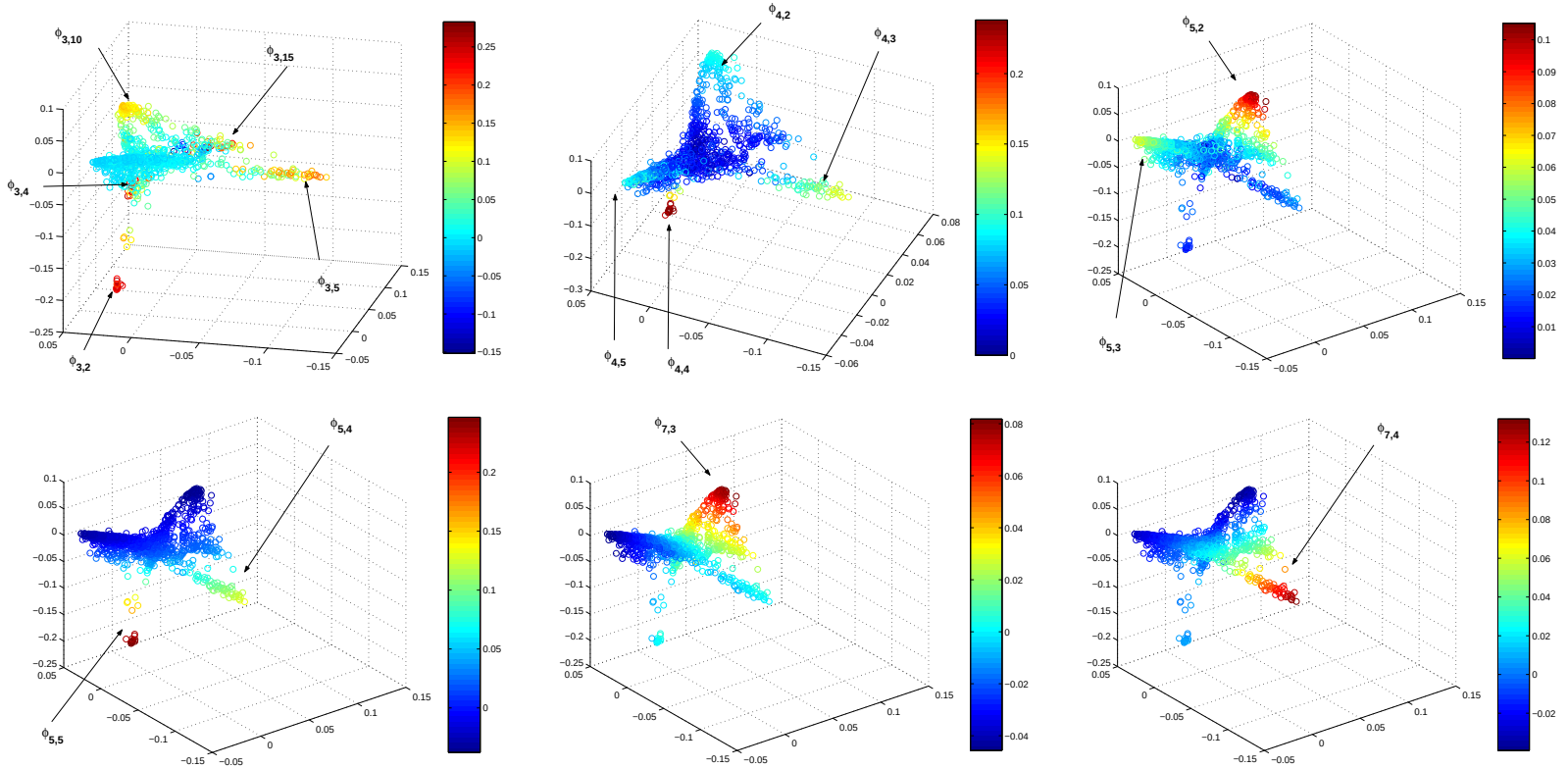


Left: reconstruction of the function F with top 50 best basis packets. Right: reconstruction with top 200 eigenfunctions of the Beltrami Laplacian operator.



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.

Multiscale construction on a document corpus



Scaling functions at different scales represented on the set embedded in 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.

Multiscale construction on a document corpus, II

Doclet	Document Titles	Words
$\varphi_{2,3}$	Acid rain and agricultural pollution Nitrogen's Increasing Impact in agriculture	nitrogen,plant, ecologist,carbon, global
$\varphi_{3,3}$	Racing the Waves Seismologists catch quakes Tsunami! At Lake Tahoe? How a middling quake made a giant tsunami Waves of Death Seabed slide blamed for deadly tsunami Earthquakes: The deadly side of geometry	earthquake,wave, fault,quake, tsunami
$\varphi_{3,5}$	Hunting Prehistoric Hurricanes Extreme weather: Massive hurricanes Clearing the Air About Turbulence New map defines nation's twister risk Southern twisters Oklahoma Tornado Sets Wind Record	tornado,storm, wind,tornadoe, speed

Multiscale construction on a document corpus, III

Wordlets:
[university,science,study,team,group,found,psychologist,colleague,sn,finding],

[bitter, glutamate, beetle, tongue, nih, vegetable, taste, corn, pest, seed],

[monarch, tag, caterpillar, bt, butterfly, honeybee, pollen, salmon, bee, wing],

[bead, magnet, mosquito, mosquitoe, intestinal, text, alloy, sampling, coating, sphere],

[smoker, nicotine, gum, saliva, dioxin, cavity, cigarette, neandertal, dental, coral],

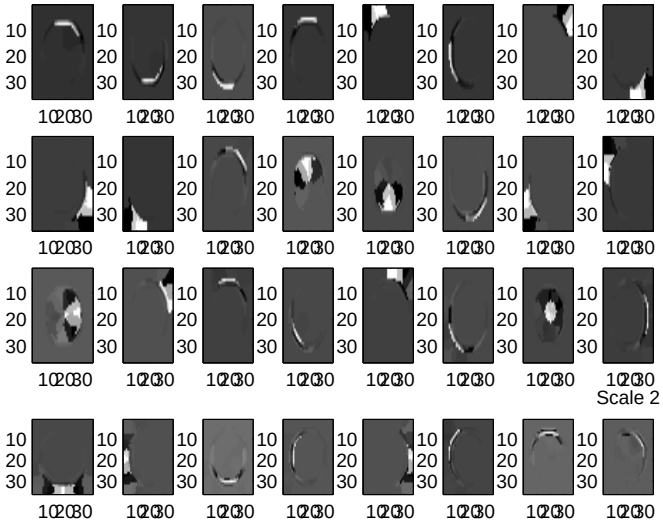
[utah, powder, fungus, smith, bitter, flower, trap, asian, win, caterpillar],

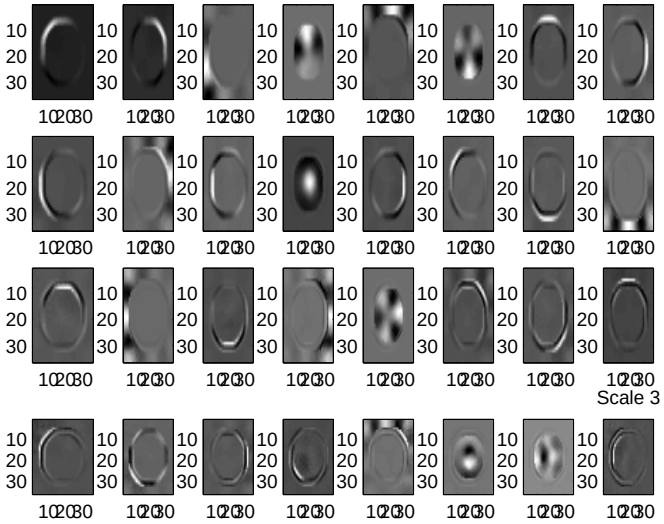
[beer, wine, king, powder, antioxidant, alcohol, vitamin, drink, cholesterol, taste],

[intestinal, vaccine, immune, antibiotic, disease, salmonella, infection, pig, infectious, pathogen],

[solvent, polymer, pcb, chemist, ozone, gray, dioxide, iq, salmon, epstein]

Nonlinear Analysis of Images





Scaling functions on the graph of patches extracted from an image of a white full circle on black background, with noise.

Potential Theory, Efficient Direct Solvers

The Laplacian $\mathcal{L} = I - T$ has an inverse (on $\ker(\mathcal{L})^\perp$) whose kernel is the Green's function, that if known would allow the solution of the Dirichlet or Neumann problem (depending on the boundary conditions imposed on the problem on $\mathcal{L}$). If $||T|| < 1$, one can write the Neumann series

$$(I-T)^{-1} f = \sum_{k=1}^{\infty} T^k f = \prod_{k=0}^{\infty} (I + T^{2^k}) f \,.$$

Since we have compressed all the dyadic powers T^{2^k}, we have also computed the Green's operator in compressed form, in the sense that the product above can be applied *directly* to any function f (or, rather, its diffusion wavelet transform). Hence this is a direct solver, and potentially offers great advantages, especially for computations with high precision, over iterative solvers.

Final observations and conclusions

Multiscale diffusion geometry gives the ability to obtain multiscale functions on the data, multiscale representations of the graph and of a diffusion on the graph. These functions can be used for local parametrizations, embeddings and features.

Learning tasks can be viewed as function approximation + regularization constraints on functions on the data, and multiscale analysis is necessary here as much as it is in even 1D.

We are still at the beginning of exploring the possibilities offered by these ideas.

Collaborators

- R.R. Coifman, P.W. Jones (Yale Math) [Diffusion geometry; Diffusion wavelets; Uniformization via eigenfunctions], S.W. Zucker (Yale CS) [Diffusion geometry];
- G.L. Davis (Pathology), F.J. Warner (Yale Math), F.B. Geshwind , A. Coppi, R. DeVerse (Plain Sight Systems) [Hyperspectral Pathology];
- S. Mahadevan (U.Mass CS) [Markov decision processes];
- A.D. Szlam (Yale) [Diffusion wavelet packets, top-bottom multiscale analysis, linear and nonlinear image denoising, classification algorithms based on diffusion];
- J.C. Bremer (Yale) [Diffusion wavelet packets, biorthogonal diffusion wavelets];
- R. Schul (UCLA) [Uniformization via eigenfunctions; nonhomogenous Brownian motion];
- W. Goetzmann (Yale, Harvard Business School), J. Walden (Berkley Business School), P.W. Jones (Yale Math) [Applications to finance]
- M. Mahoney (Yahoo Research), F. Meyer (UC Boulder), X. Shen (UC Boulder) [Randomized algorithms for hyper-spectral and fMRI imaging]
- H. Mashkar (LA State) [polynomial frames of diffusion wavelets, characterization of function spaces];
- Y. Kevrekidis (Princeton Eng.), S. Lafon (Google), B. Nadler (Weizman) [stochastic dynamics];

This talk, papers, Matlab code (currently working on a Matlab toolbox) available at

`www.math.yale.edu/~mmm82`

Thank you!