Diffusion Geometries, Global and Multiscale

Mauro Maggioni

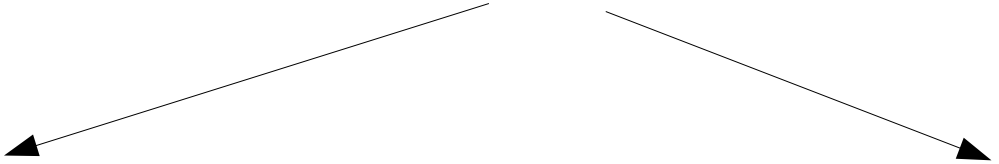
Joint work with
R.R. Coifman, S. Lafon,
J.C. Bremer Jr., A.D. Szlam,
P.W. Jones, R.Schul

Papers, talks, other materials available at: www.math.yale.edu/~mmm82

Data and functions on the data

In learning, clustering and analysis of data, one is interested in finding interesting structures in the data, and be able to predict properties of future data. Example of data: large body of "documents", such as web pages.

Functions on the data



Features: these are functions of the "coordinates" in which the data is given, that enhance or reveal some property of interest of the data; somtimes thought as mapping the data into a more convenient space. Can be both *supervised* or *unsupervised*. E.g.: word frequencies, or some finer measure of relevance of each word in each document.

Functions "interesting by definition".
In general complicated, and "to be learned". *Supervised.*
E.g.: most visited web pages (or: most interesting to you, most relevant to a given query etc...).

Need: analysis of functions on the data.

Data and functions on the data (cont'd)

Many difficulties, in general very few points compared to the dimensionality of the space in which the data lies.
Typical assumptions:
- the data is “low-dimensional” in nature.
- the data has some smoothness (+noise).
- the geometry of the data is relevant towards understanding functions on the data.

Many (many) methods that try to take advantage of the above assumptions:

- Manifold parametrization, (non)linear dimensionality reduction, “kernel methods”.
- Learning, complexity control.
- Approximation theory, but on manifolds, high-dimensional spaces.

In this talk: discuss tools for Fourier analysis and multiscale wavelet analysis of functions on manifolds, varifolds, graphs, “datasets”.

Basic Ingredients

Model the data as a weighted graph (**G,E,W**):

- the vertices represent data points
- the edges connect similar data points
- the weights represent a similarity measure.

Example: have an edge between web pages connected by a link; and/or between documents with very similar word frequencies.

In important cases, vertices are points in high-dimensional Euclidean space, weights may be a function of Euclidean distance, and/or the geometry of the points.

High dimensional data: examples

- Documents, web searching
- Customer databases
- Hyperspectral imagery (satallite, biomedical, etc...)
- Social networks
- Gene arrays, proteomics data
- Art transactions data
- Traffic (automobilistic, network) statistics
- ....

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. However we ask only to define similarity only for very similar objects.

A local "similarity" operator on the set

Let $X = \{x_1, \ldots, x_n\}$ a data set. The similarity between points of X is summarized in a kernel function $k(x, y)$ s.t.:

- symmetric: $k(x, y) = k(y, x)$,
- positivity-preserving: $k(x, y) \geq 0$,
- positive semi-definite:

$$\sum_{x,y \in X} \alpha(x)\overline{\alpha}(y)k(x, y) \geq 0$$

Examples of similarity kernels

If X lies in $\mathbb{R}^n$, examples of similarity include:

- $k(x,y) = e^{-(\frac{||x-y||}{\delta})^2}$,
- $k(x,y) = \frac{\langle x,y \rangle}{||x||||y||}$,
- $k(x,y) = \frac{1}{\epsilon + ||x-y||}$,
- $\cdots$

Example of feature distances:

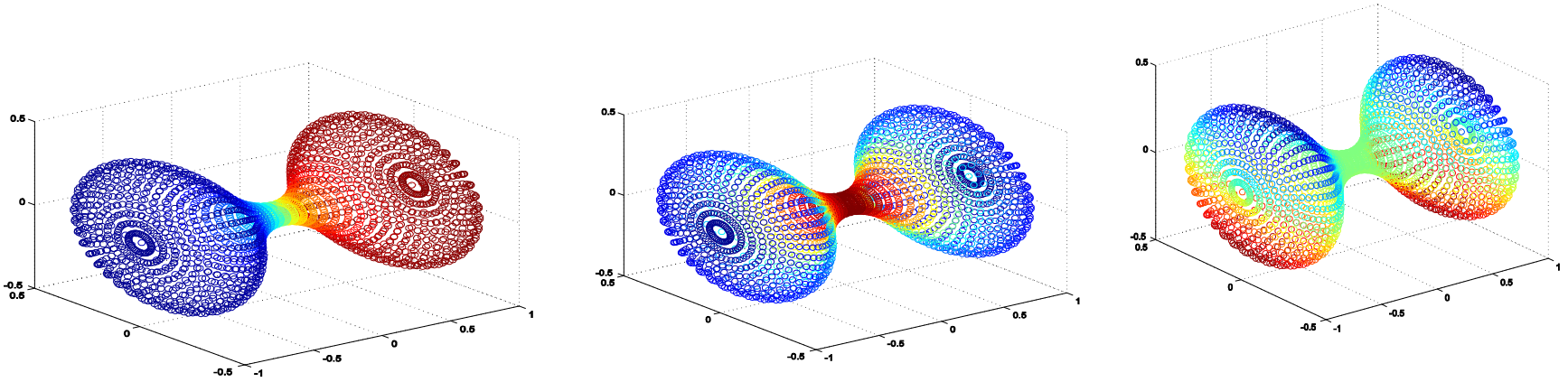
$$\tilde{k}(x,y) = k(f(x), f(y)),$$

where $k(x,y)$ is one of the above, and f a (non-linear) function to $\mathbb{R}^n$.

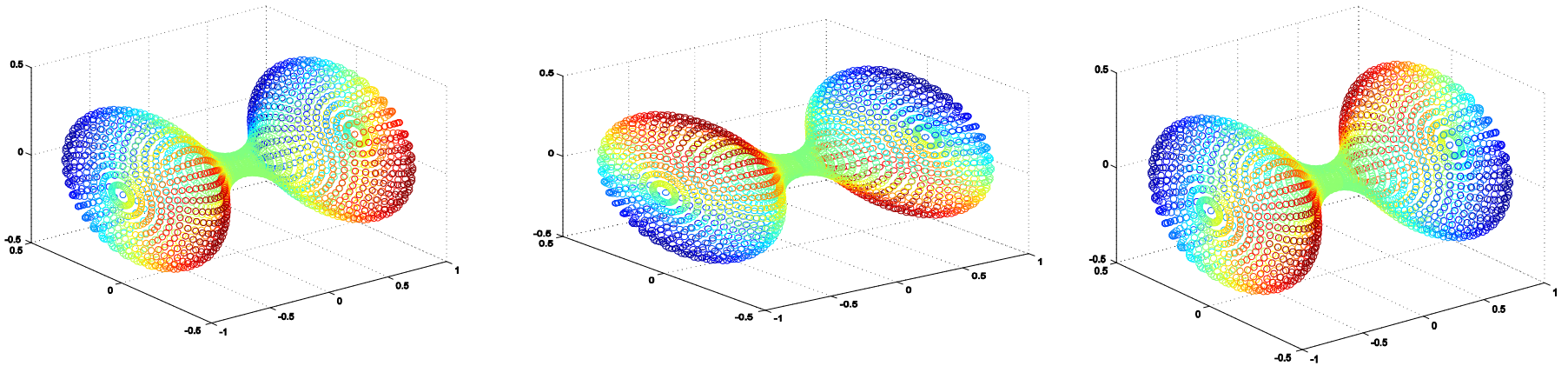
If a model (probabilistic, dynamic...) for the data is available, that can be used to generate kernels "consistent" with that model.

Spectral decomposition of the kernel

$$K^t(x,y) = \sum_{\lambda \in \sigma_T} \lambda^t \xi_\lambda(x) \xi_\lambda(y)$$



Eigenfunctions of a (small-scale) Gaussian kernel restricted to points on a dumbbell manifold



Spectral kernel methods

Recent kernel methods: LLE (Roweis, Saul 2000),
Laplacian Eigenmaps (Belkin, Niyogi 2002),
Hessian Eigenmaps (Donoho, Grimes 2003),
LTSA (Zhang, Zha 2002) ...

all based on the following paradigm: minimize $Q(f)$ where

$$Q(f) = \sum_{x \in X} Q_x(f)$$

$Q_x(f)$: quadratic form measuring local variation of f in a neighborhood of x

Solution: compute eigenfunctions $\{\varphi_l\}$ of Q and map data points via

$$x \quad (\varphi_0(x), \varphi_1(x), \ldots, \varphi_p(x))^T$$

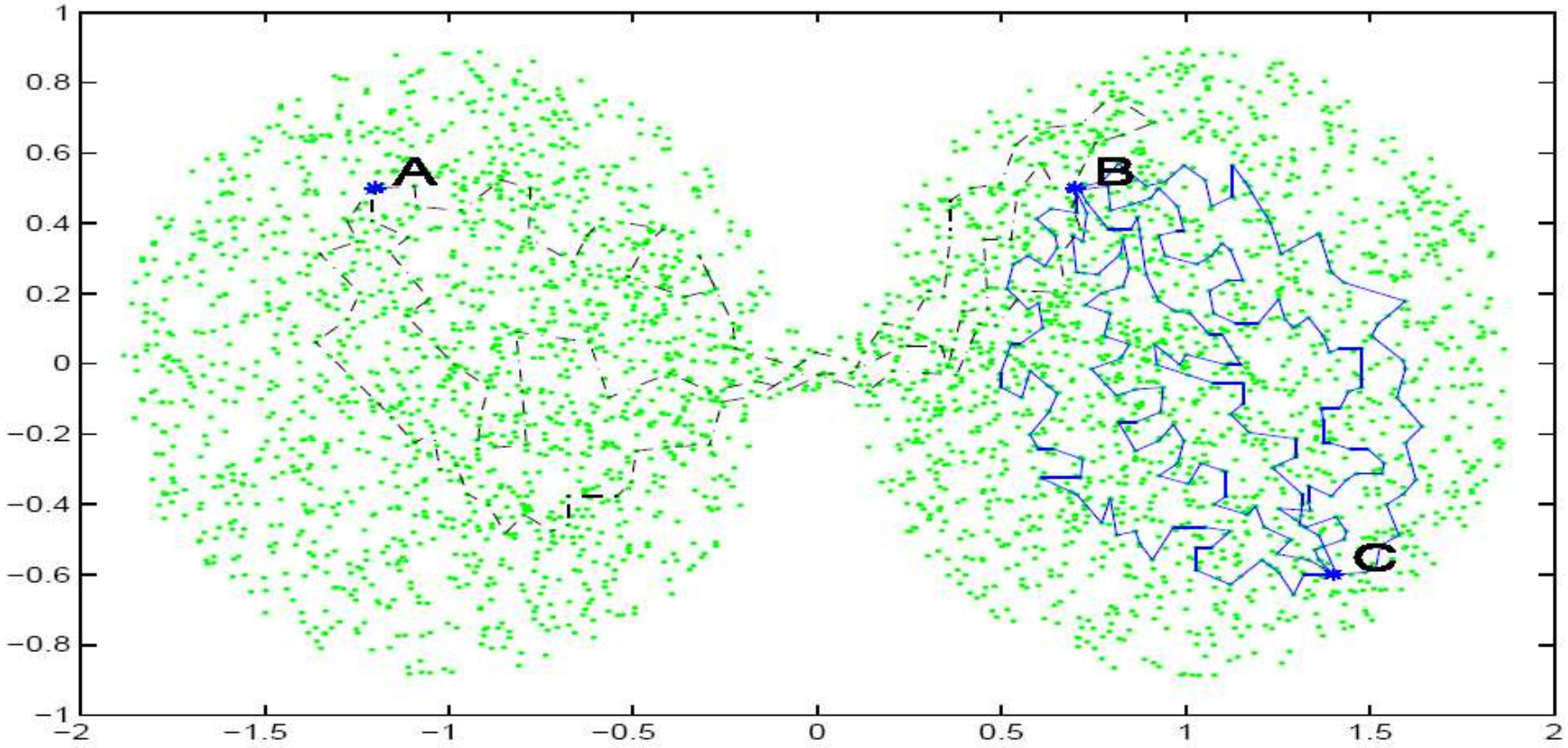
In our case we minimize $\sum_{x \in X} k(x, y)(f(x) - f(y))^2$

Laplacian on a graph, diffusion geometries

[M. Belkin, P. Nyogi, RR Coifman, S. Lafon]

From local to global: diffusion distances

Motto: Diffusion distance measures and averages connections of all lengths, uses a "preponderance of evidence"



Some pictures in the next few slides are courtesy of Stephane Lafon (www.math.yale.edu/~sl349)

An example: the Erdös number & the "small world effect"

Graph whose vertices are mathematicians, edge between two mathematicians if co-authored of a paper.
The Erdös number of a mathematician M is the length of a geodesic path between M and the vertex corresponding to P. Erdös.

The Erdös connected component has small diameter (13), the mean distance is <5 and the volume is concentrated in a ball of radius 8 around Erdös.
This is not very informative; it is called the "small world effect". One may try to argue that in some sense the "world of mathematicians" *is* small...however: diam(world population)~20, diam(www)~19.

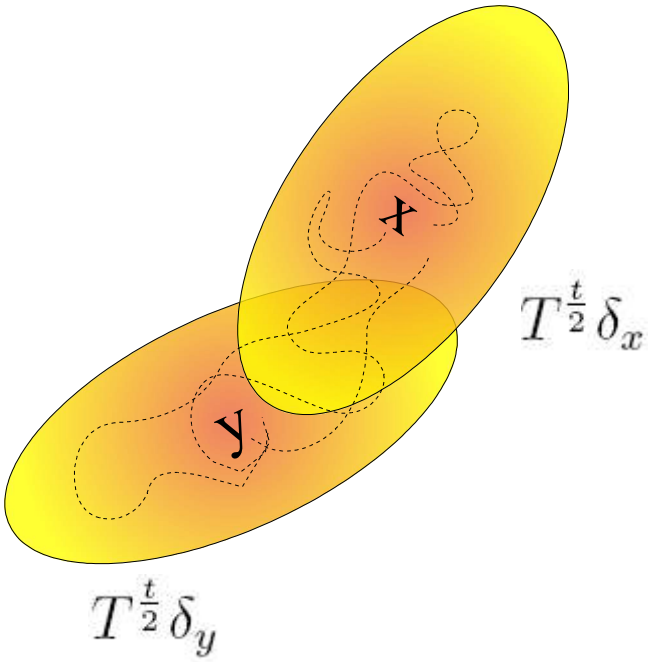
P. Erdös. Picture from
http://www.oakland.edu/enp/

The diffusion distance will take into account not only the shortest path from Erdös to a mathematician, but all the other paths, of all lengths, and takes a weighted average of them, refining the classifications of connections.

Diffusion Distances

$$K^t(x,y) = \sum_{\lambda \in \sigma_T} \lambda^t \xi_\lambda(x) \xi_\lambda(y)$$

$$\begin{aligned} d^{(t)}(x,y) &= \sqrt{\sum_{\lambda \in \sigma_T} \lambda^t \left(\xi_\lambda(x) - \xi_\lambda(y)\right)^2} \\ &= \sqrt{\langle \delta_x - \delta_y, T^t(\delta_x - \delta_y) \rangle} \\ &= ||T^{\frac{t}{2}}\delta_x - T^{\frac{t}{2}}\delta_y||_2. \\ &= \sqrt{K^t(x,x) + K^t(y,y) - 2K^t(x,y)} \end{aligned}$$

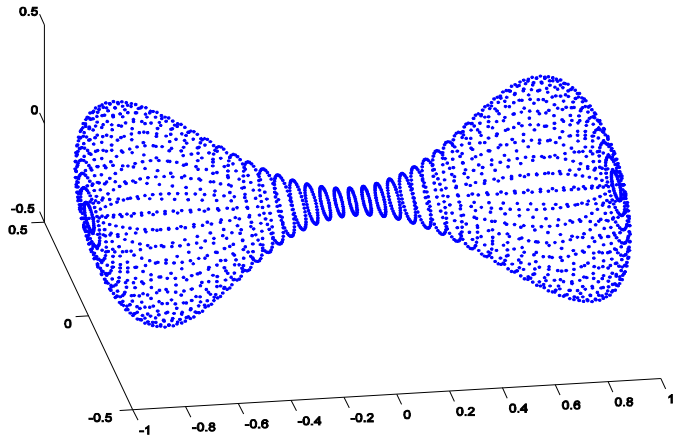


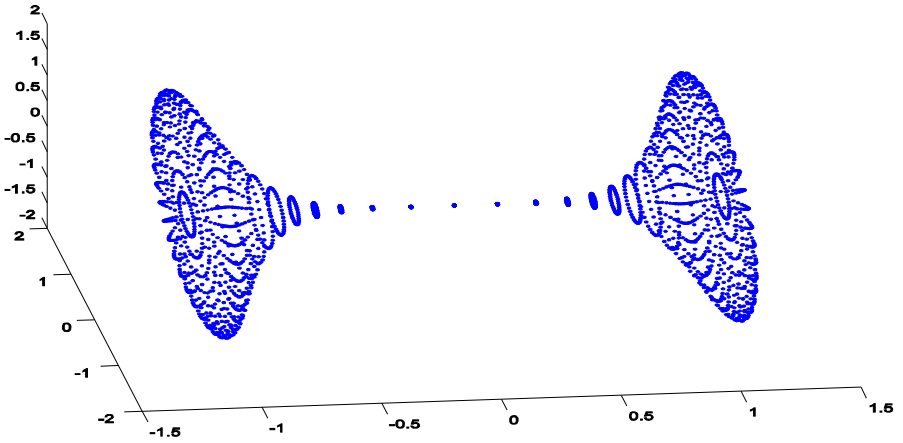
Diffusion embedding mapping ***X*** with diffusion distance into Euclidean space with Euclidean distance:

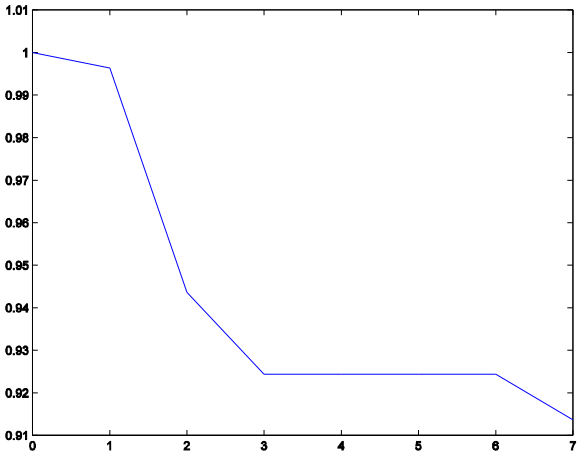
$$\Phi_m^{(t)}(x) = \left(\lambda_0^t \xi_0(x), \lambda_1^t \xi_1(x), \ldots, \lambda_{m-1}^t \xi_{m-1}(x)\right) \in \mathbb{R}^m$$

Diffusion embedding mapping $\boldsymbol{X}$ with diffusion distance into Euclidean space with Euclidean distance:

$$\Phi_m^{(t)}(x) = \left(\lambda_0^t \xi_0(x), \lambda_1^t \xi_1(x), \ldots, \lambda_{m-1}^t \xi_{m-1}(x)\right) \in \mathbb{R}^m$$







Applications

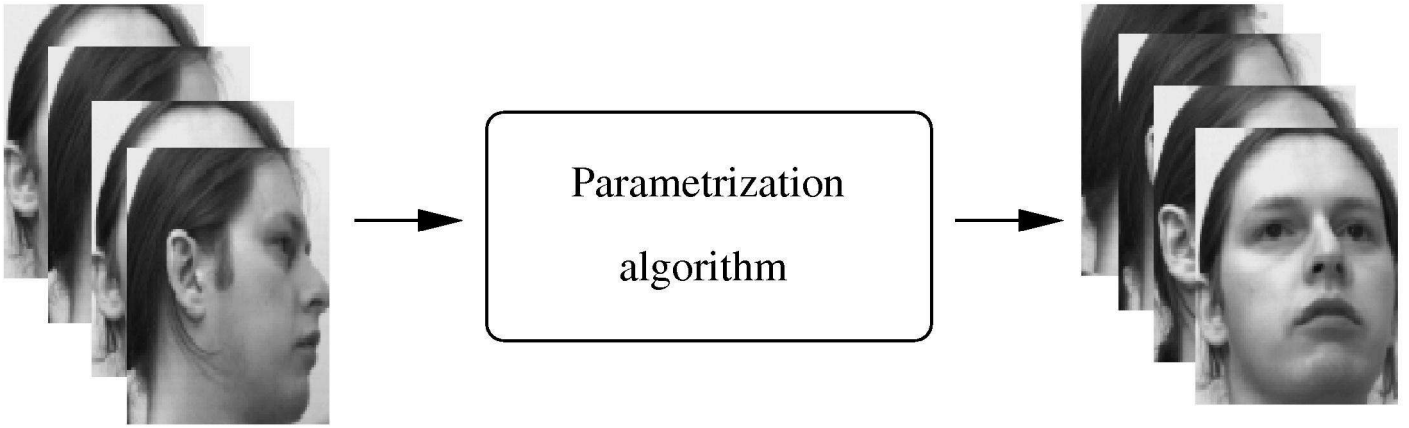
Many successful applications of spectral kernel methods.
For Laplacian eigenfunctions, the following work in particular:

- 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, MM, B. Nadler)

Example: curves

Umist face database: 36 pictures (92x112 pixels) of the same person being randomly permuted.

Goal: recover the natural organization of the data set (image ordering).

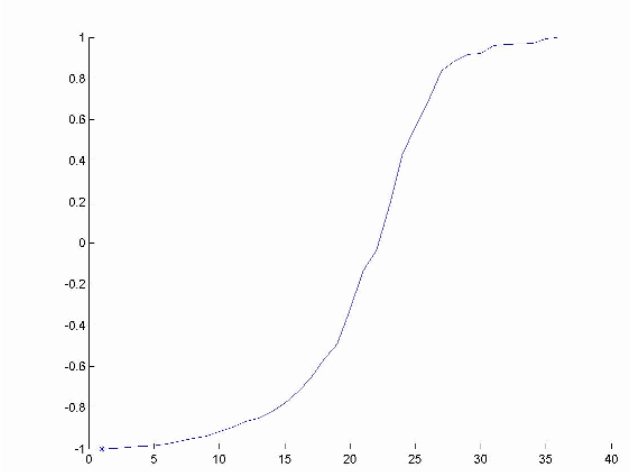


See Stephane Lafon's web page for this and other (animated!) examples.

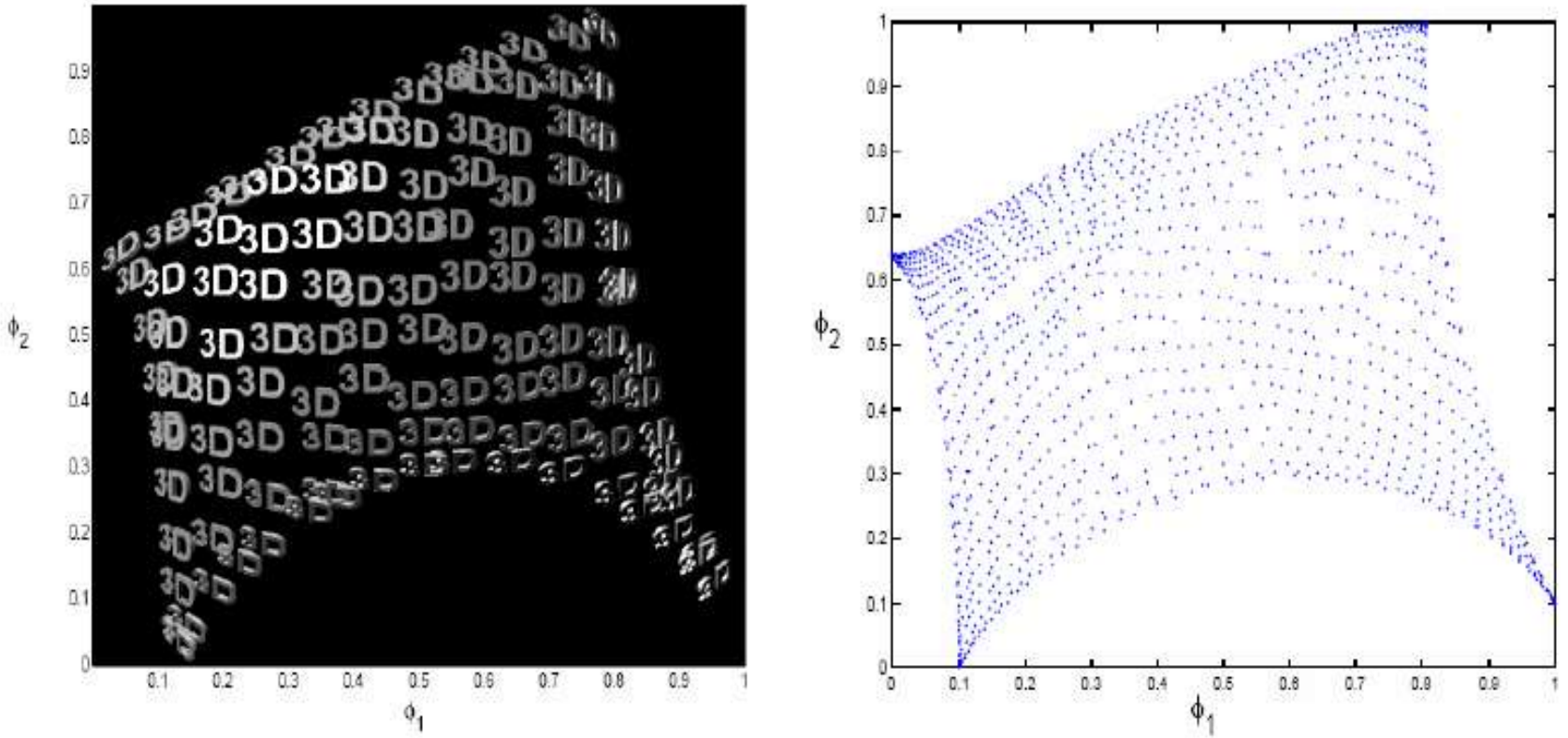
Original ordering

Re-ordering

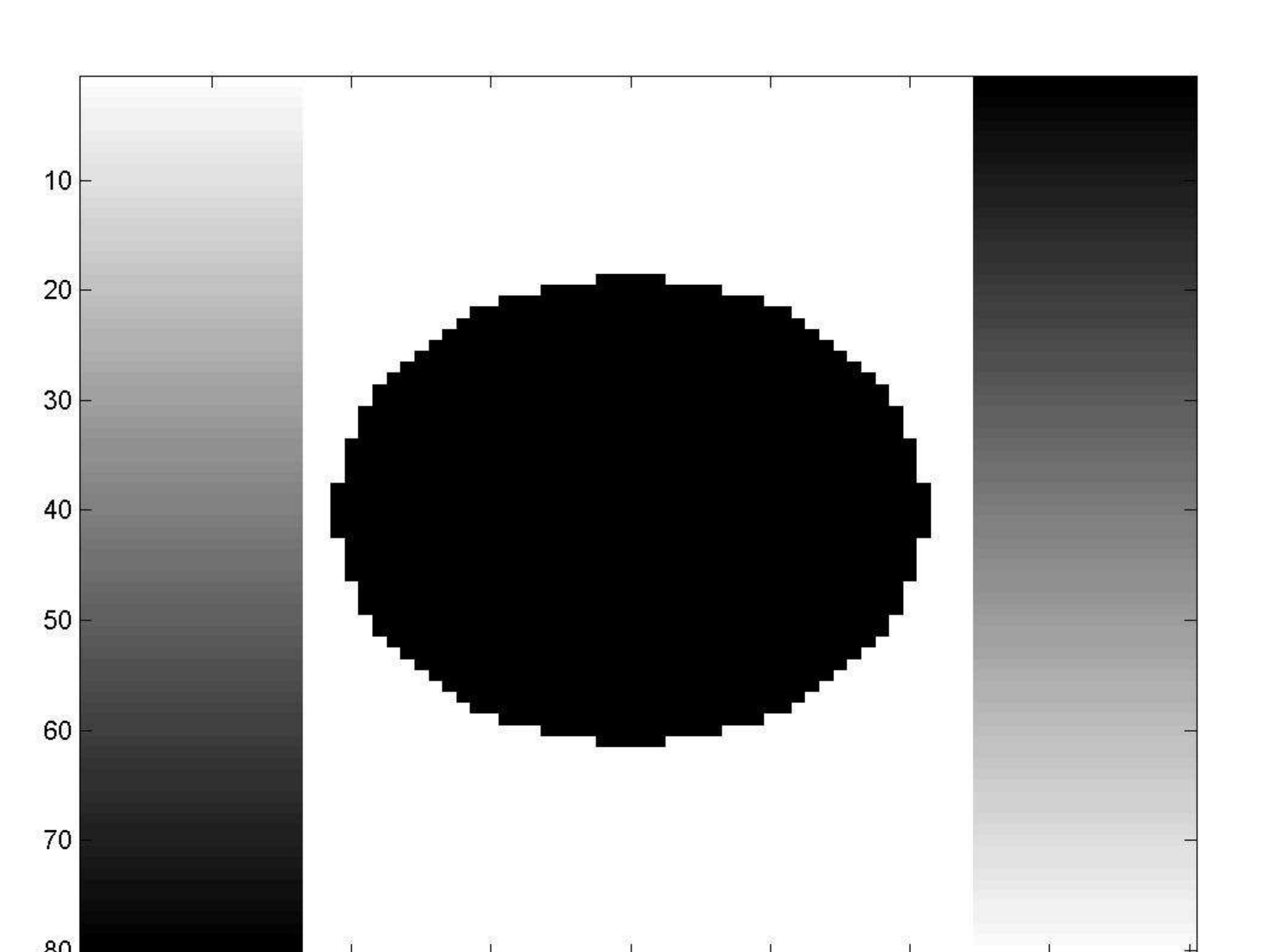
The second eigenfunction ϕ_1 assigns a realnumber to each image. When this set of numbers is re-ordered, one obtains a graph very similar to $\cos(t)$ on $[0,\pi]$.

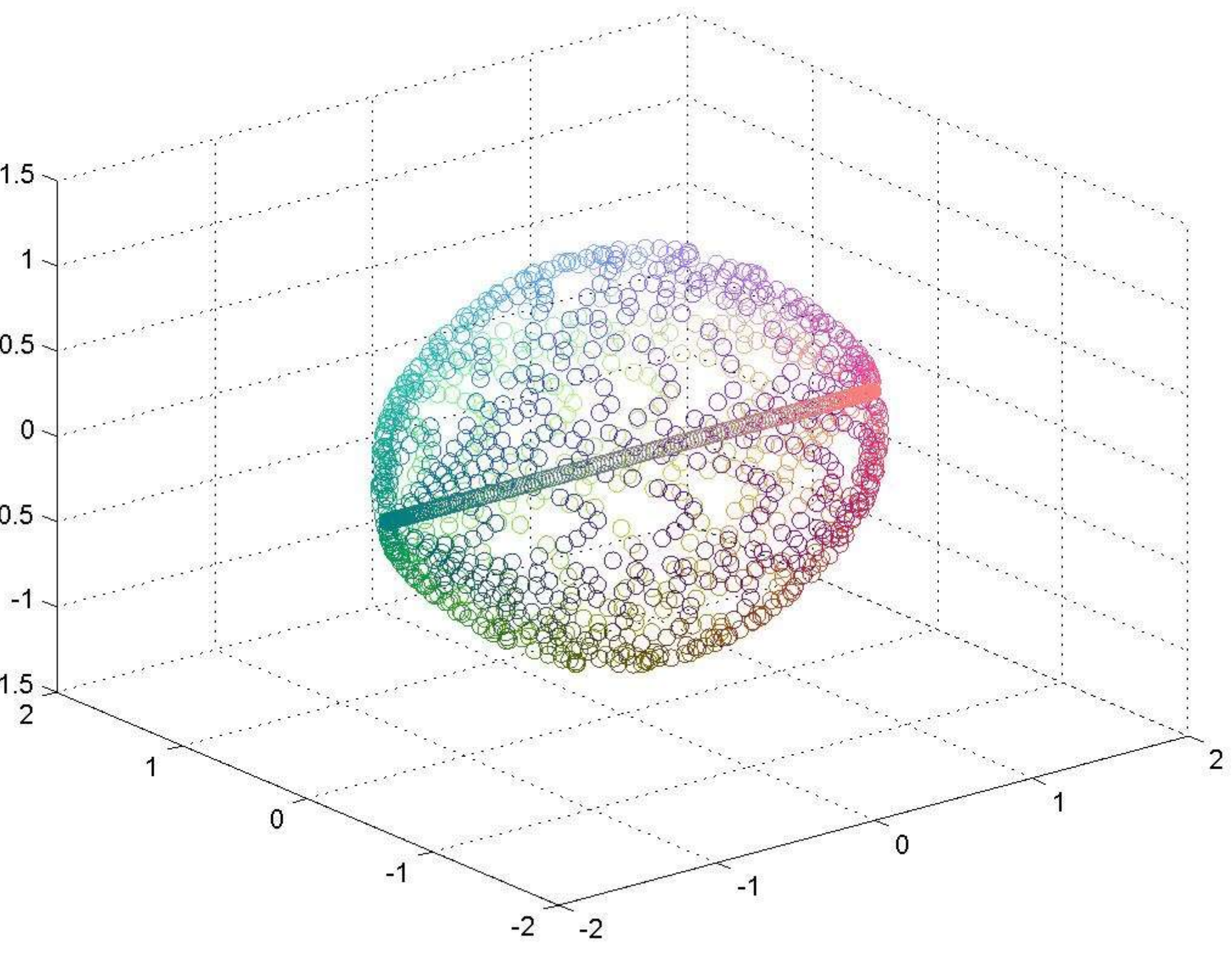


The natural parameter (angle of the head) is recovered, the data points are re-organized and the structure is identified as a curve with 2 endpoints.



The First two eigenfunctions organize the small images which were provided in random order


10
20
30
40
50
60
70
80


1.5
1
0.5
0
-0.5
-1
-1.5
2
1
0
-1
-2
-2
-1
0
1
2

Hyper-spectral Pathology Data

[MM, GL Davis, F Warner, F. Geshwind, A Coppi, R. DeVerse, RR Coifman]

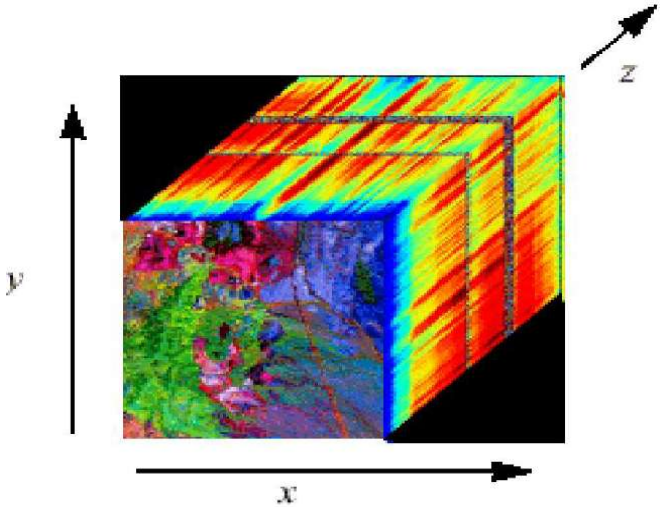


Figure 3: Hyperspectral data cube. (DataFusionCorp.)

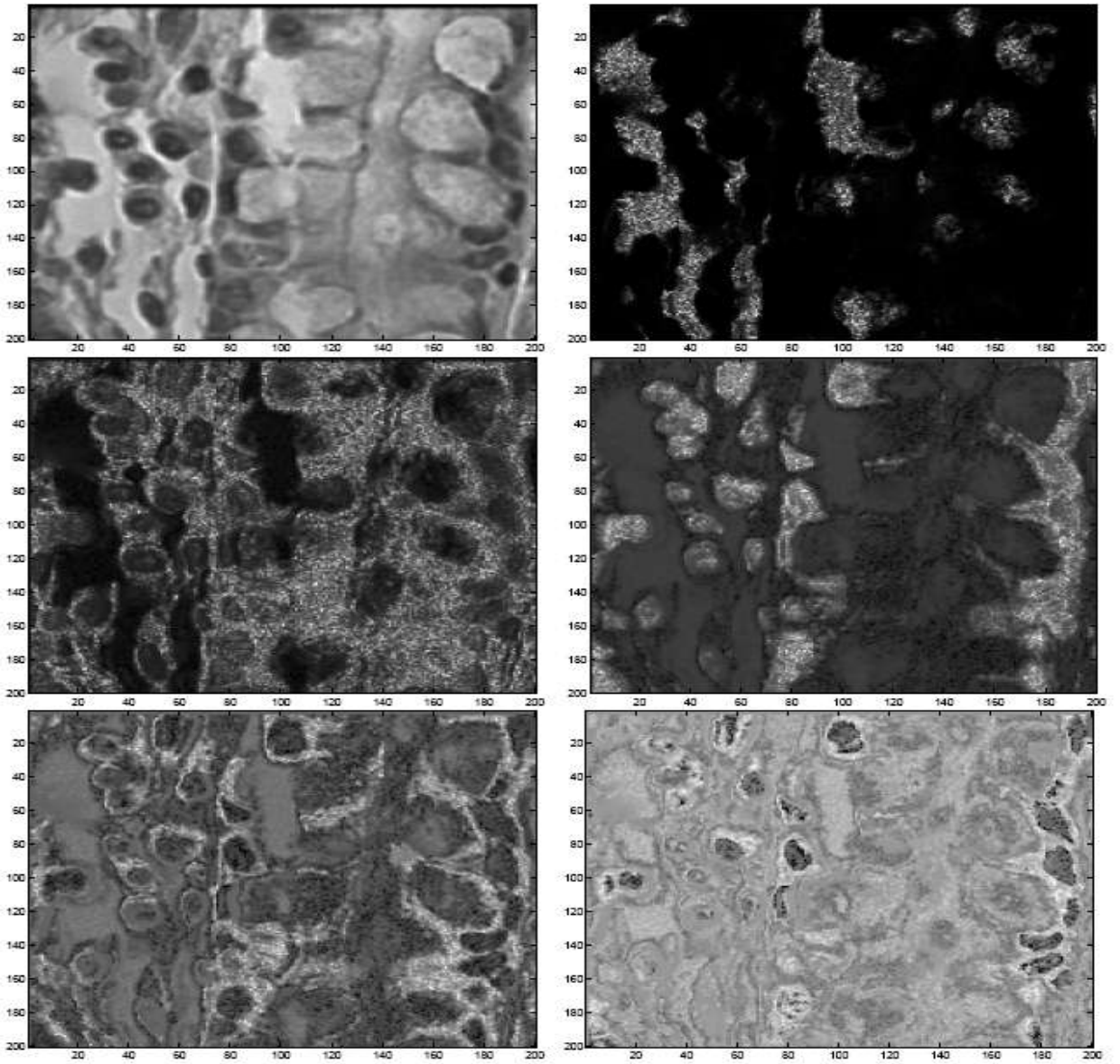
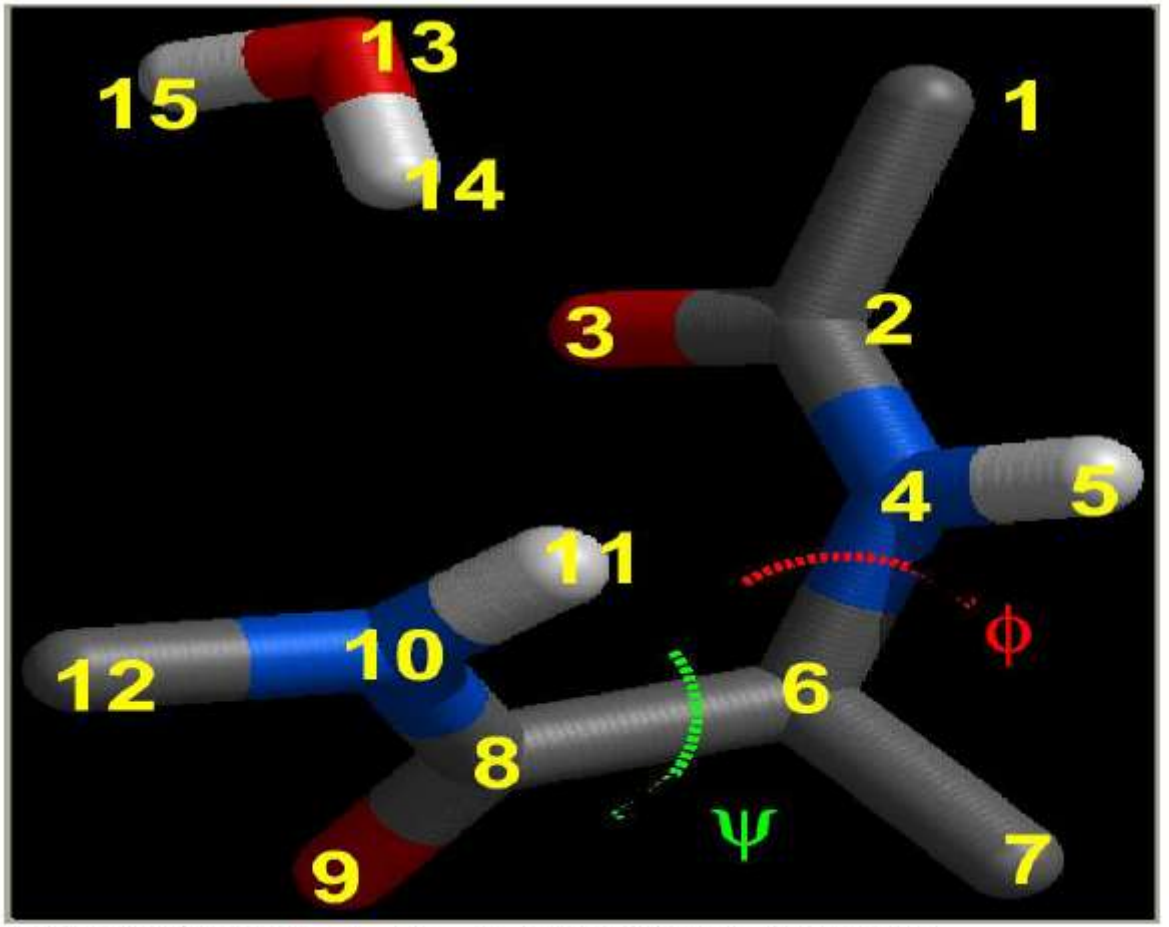


Fig. 2. Moving left to right, top to bottom, the eigenfunctions φ_k of the diffusion on the data set of spectra, for $k = 1, 2, 3, 4, 5, 6$ respectively, are mapped to the colors.

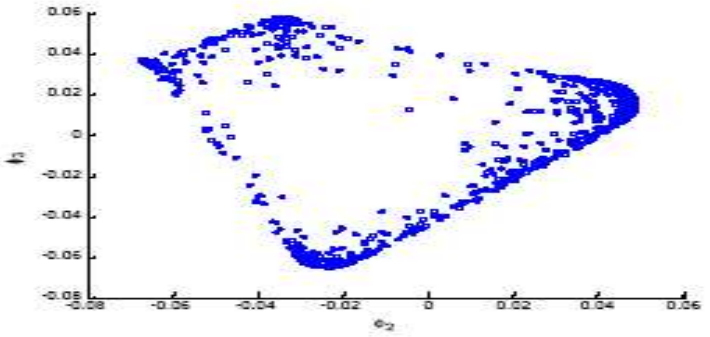
``Coarse molecular dynamics of a peptide fragment: free energy, kinetics and long time dynamics computations", G. Hummer and I.G.Kevrekidis {\it J. Chem. Phys.} {\bf 118}(23) pp. 10762-10773 (2003)

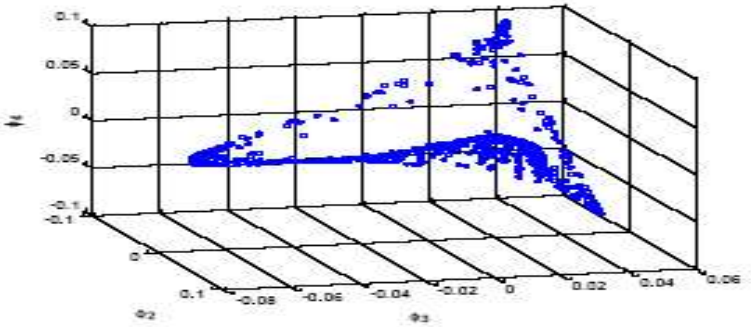


The two rotation angles ϕ and ψ are the main degrees of freedom.

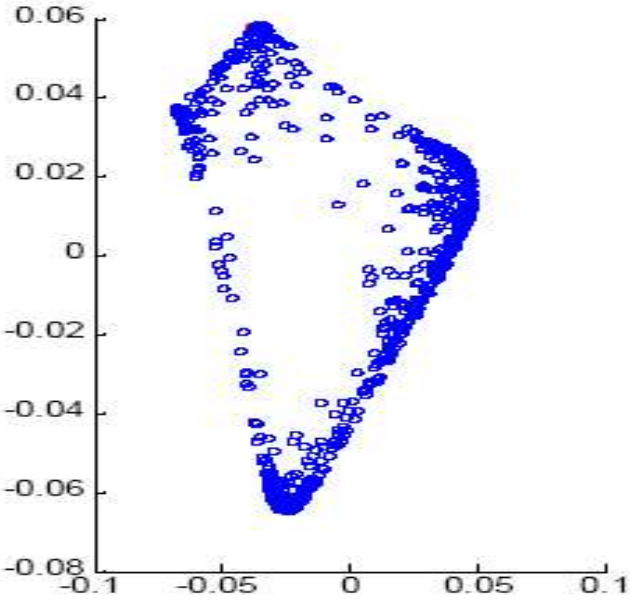
There is a special choice of kernel that uses in a consistent way the assumption the data comes from the simulation of physical systems driven by certain PDEs [Coifman-Lafon-Nadler].

The embedding in two and three dimensions look like in the following pictures:

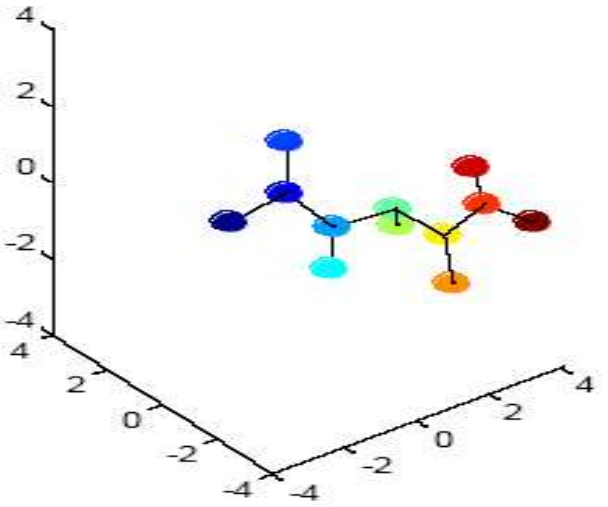


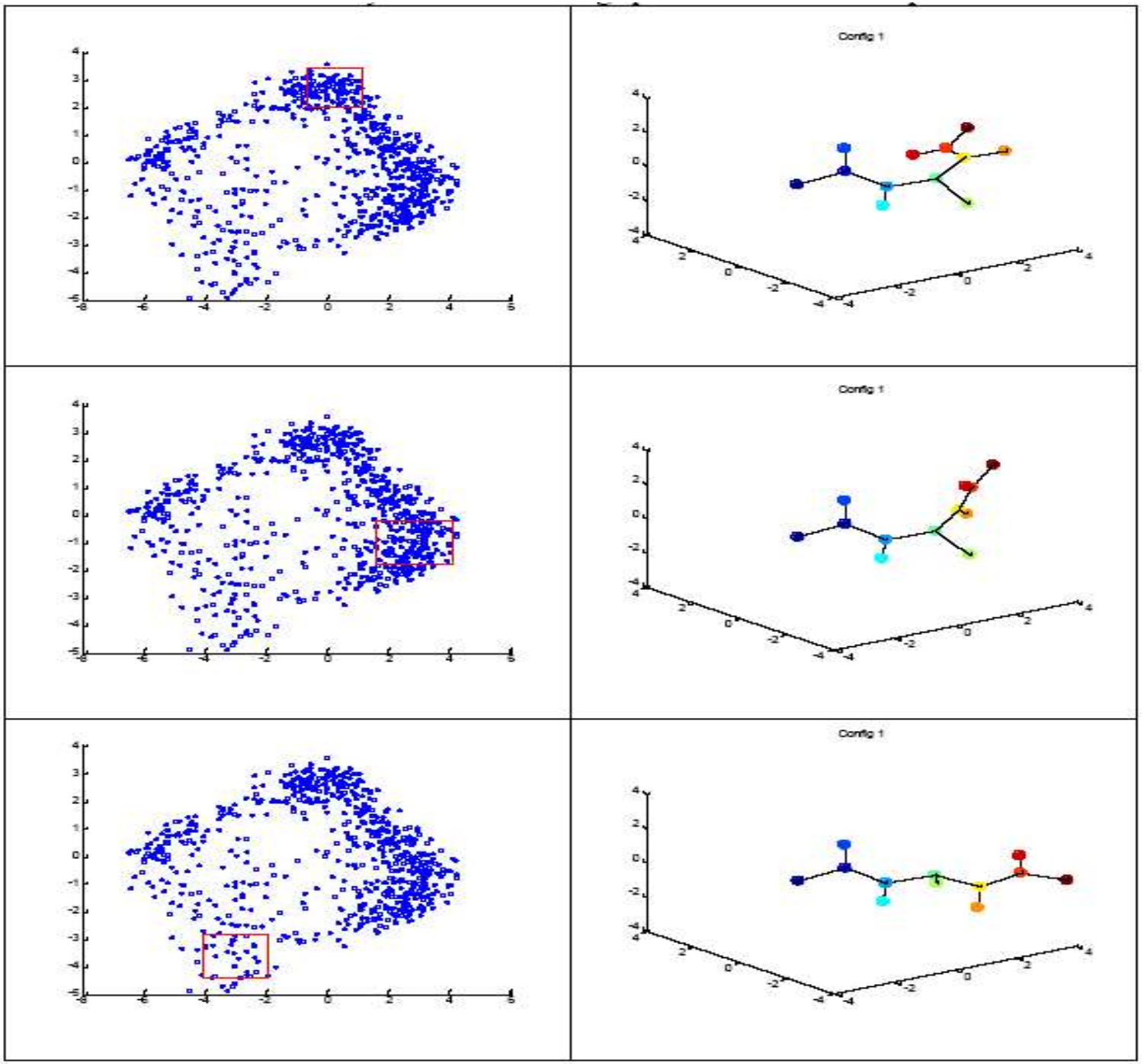


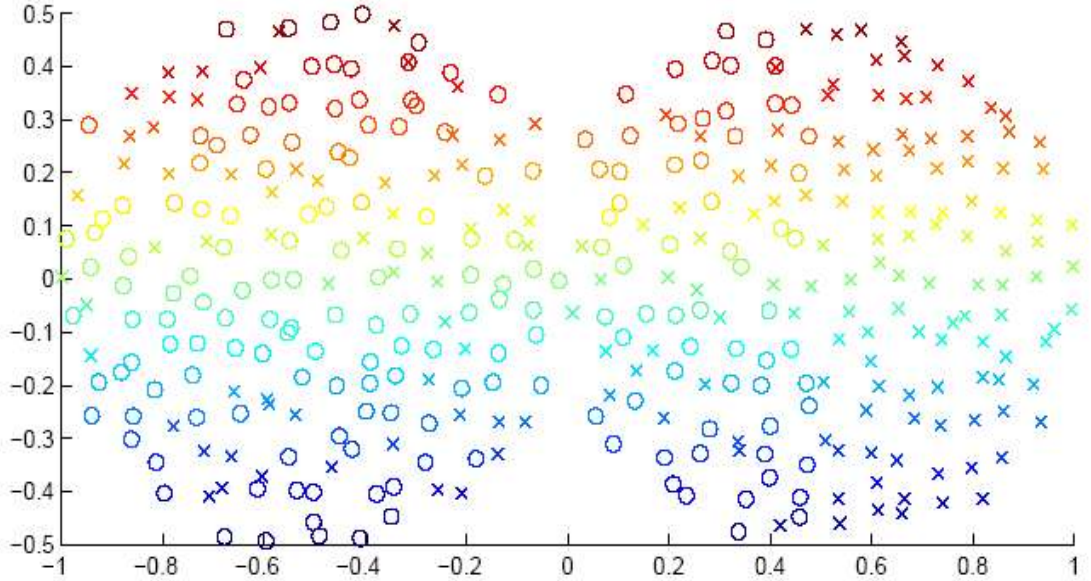
Let's browse through the points to see the corresponding configurations:

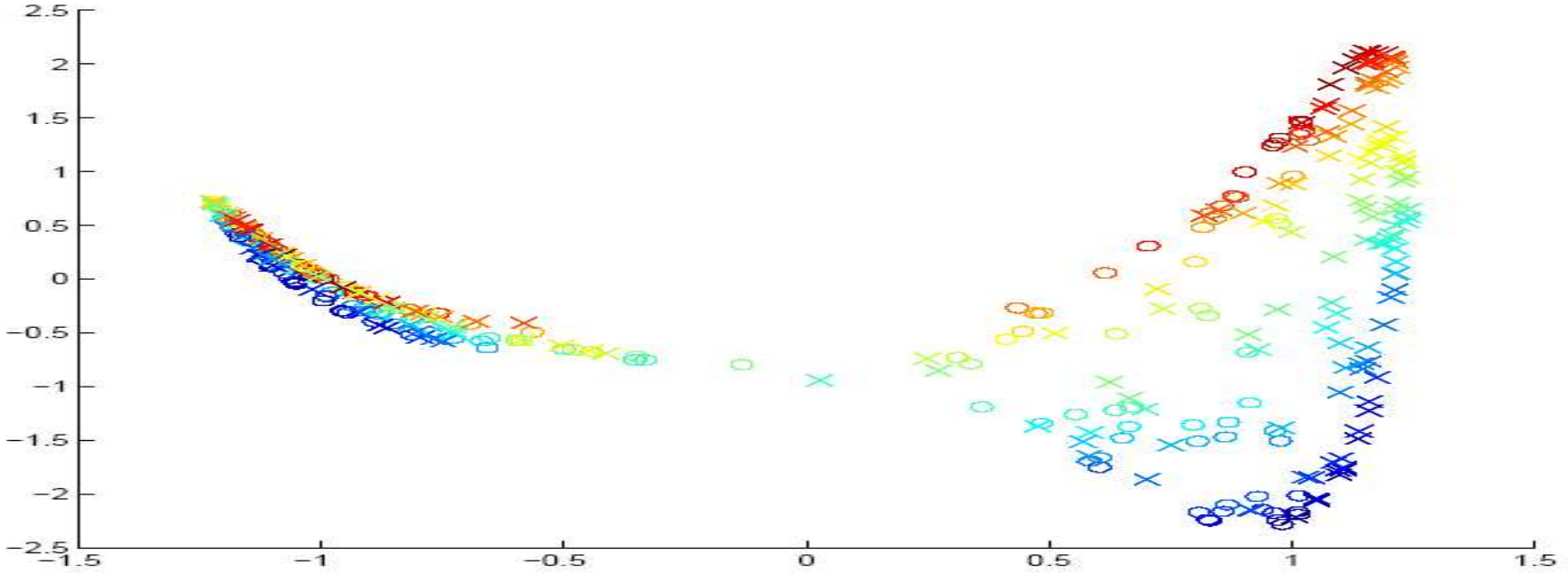


Config 1




Config 1
Config 1
Config 1


0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1


2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−1.5
−1
−0.5
0
0.5
1
1.5

Learning & Function Approximation on graphs and manifolds

In many cases learning algorithms (or learning itself?) can be viewed as an approximation problem, under smoothness constraints, on a set of points/state space represented as a manifold or graph. [think of SMVs as an example]

Approximation

Fitting the function of interest

Smoothness constraint

Want generalization power, complexity control, avoid overfitting

Manifold/graph

Belief: the structure of the state space has to do with the learning task.

Connections with Harmonic Analysis, I

In the framework just described, tools for efficiently working with functions on a manifold or graph: need efficient representation for such functions:

$$f = \sum_k \alpha_k \phi_k$$

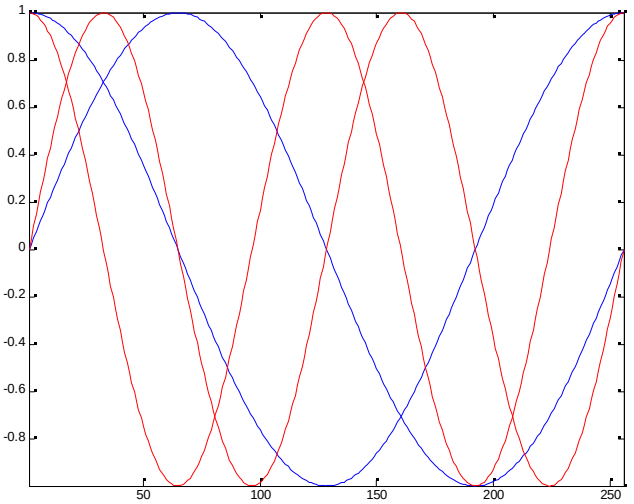
where f is a function in the class of functions we want to approximate, ϕ_k's are basis functions ("building blocks" or "templates"), and the coefficients α_k contain the information for putting together the "building blocks" in order to reconstruct (or approximate) f.

What does **efficient** mean? Few, in proportion to how "complicate" f is, and efficiently-organized coefficients α_k. Smoothness constraints become **sparsity** constraints.

Connections with Harmonic Analysis, II

[Enter Fourier]

(i) Fourier: approximate solutions of the heat equation on an interval 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, ϕ_k's are global approximants, and α_k are not as sparse as one may wish.

Semi-supervised learning with eigenfunctions of the Laplacian

[Belkin-Niyogi]

- Given all the points G you will ever want to classify, compute the ϕ_k's.

- Given a function f on a subset $G' \subset G$, write $\tilde{f} := \sum_{k \in \mathcal{K}} \alpha_k \phi_k$, so that $\tilde{f}|_{G'} \sim f$.

- To "guess" f at a point $x \in G \setminus G'$, simply let $f(x) = \sum_{k \in \mathcal{K}} \alpha_k \phi_k(x)$ (the α_k's are the ones computed above).

Usually $\mathcal{K}$ chosen as the top K eigenfunctions, or the K with largest α_k. Optimal choice of K not always easy: one the one hand one wants good approximation on G' (good fit), but good prediction on G (no overfit).

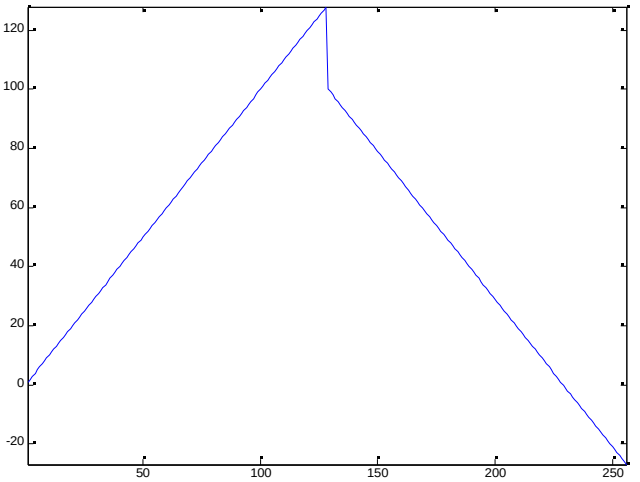
Connections with Harmonic Analysis, III

Wavelets and Multiresolution Analysis:

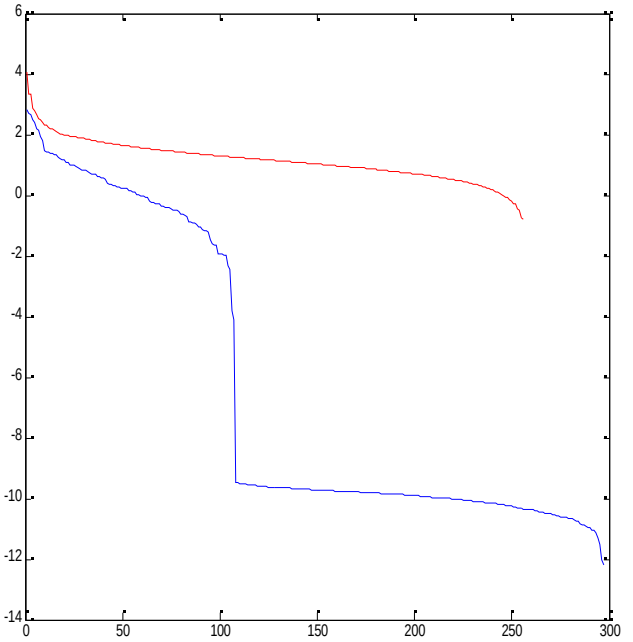
Wavelets are concentrated both in time and frequency. Wavelets have to indices $\phi_{j,k}$ is an "atom" concentrated in time at position k, width about 2^{-j}, and concentrated around frequency 2^j. They provide essentially the best possible building blocks for interesting and large classes of functions,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

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].



Function with a discontinuity



Rate of decay of coefficients onto Fourier (red) and wavelets (blue)

So far...

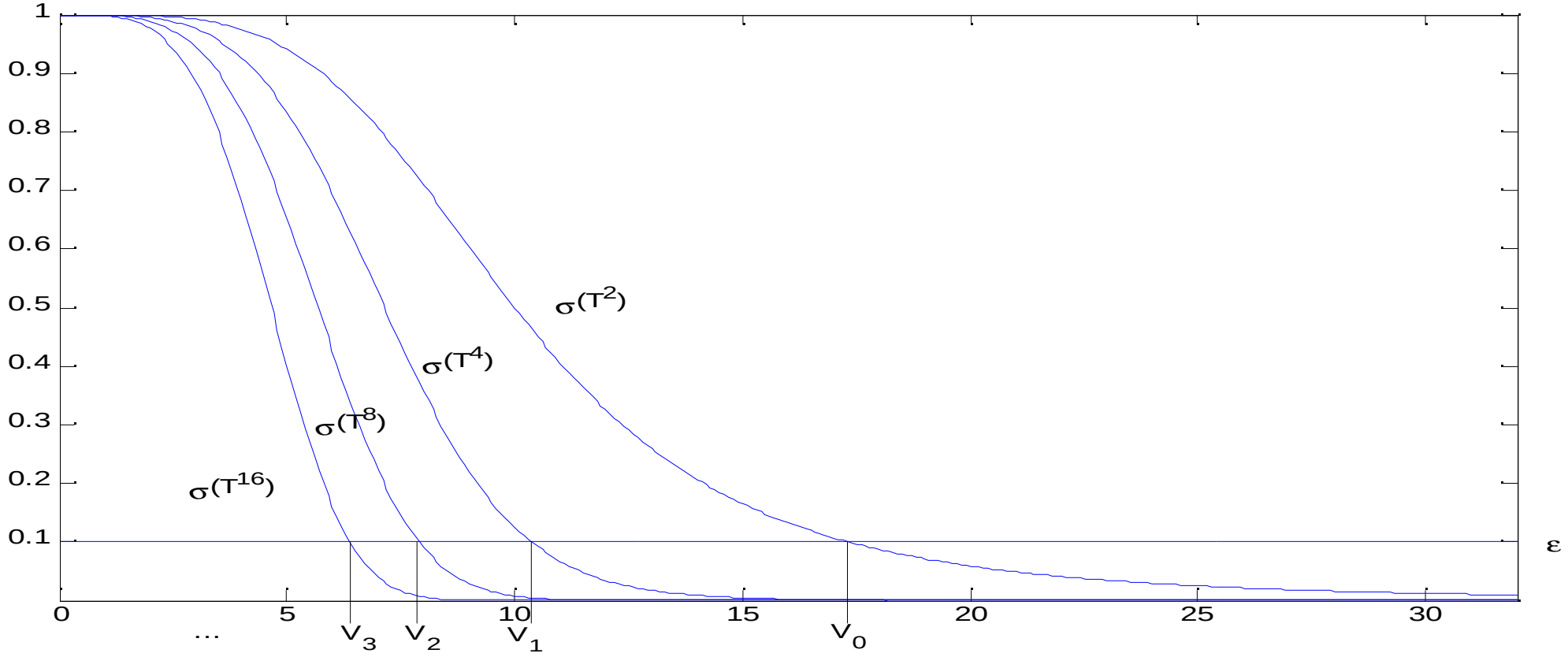
..we have seen that:
- it seems useful to consider a framework in which, for a given data set, ***local similarities*** are given only between very similar points
- it is possible to ***organize*** these local information by diffusion into global parametrizations,
- these parametrizations can be found by looking at the ***eigenvectors*** of a ***diffusion*** operator,
- these eigenvectors in turn yield a ***nonlinear embedding*** into low-dimensional Euclidean space,
- the eigenvectors can be used for global Fourier analysis on the set/manifold

PROBLEM:
Either very local information or very global information: in many problems the intermediate scales are very interesting! Would like ***MULTISCALE*** information!

Solution 1: proceed ***bottom up***: repeatedly cluster together in a multi-scale fashion, in a way that is "faithful" to the operator: diffusion wavelets.
Solution 2: proceed ***top bottom***: cut greedily according to global information, and repeat procedure on the pieces: recursive partitioning, local cosines...
Solution 3: do ***both***!

From global Diffusion Geometries...

We are given a graph X with weights W. There is a natural ***random walk*** P on X induced by these weights (P is just a normalized version of k). P maps probability distributions on X to probability distributions on X. A reversibility condition on P implies that P is conjugate to a self-adjoint operator T, which we renormalize to have 2-norm 1.
The spectra of powers of T look like:

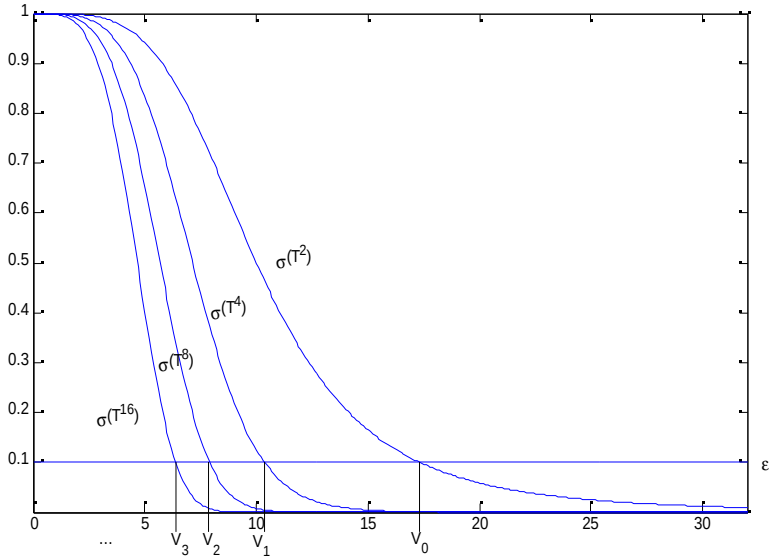


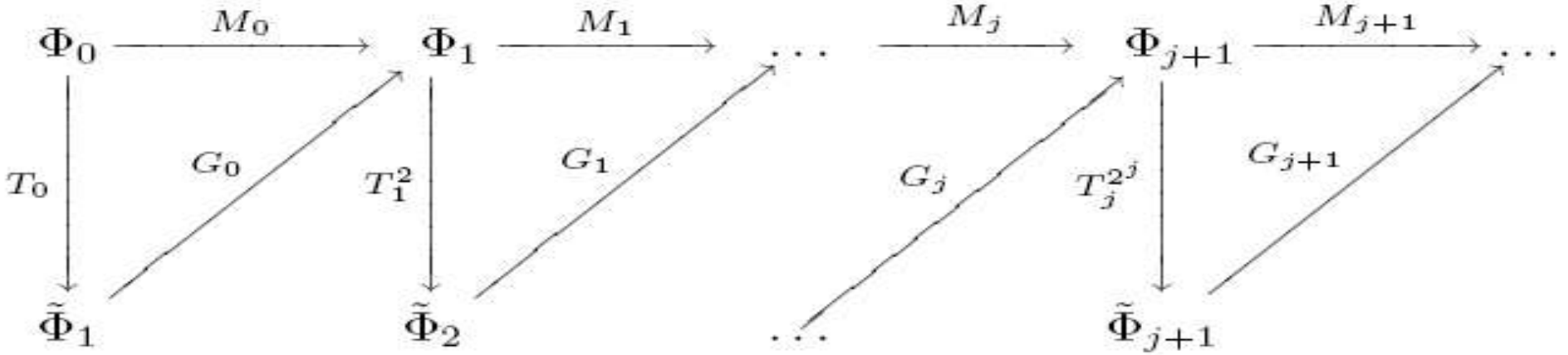
... to Multiresolution Diffusion

[Coifman,MM]

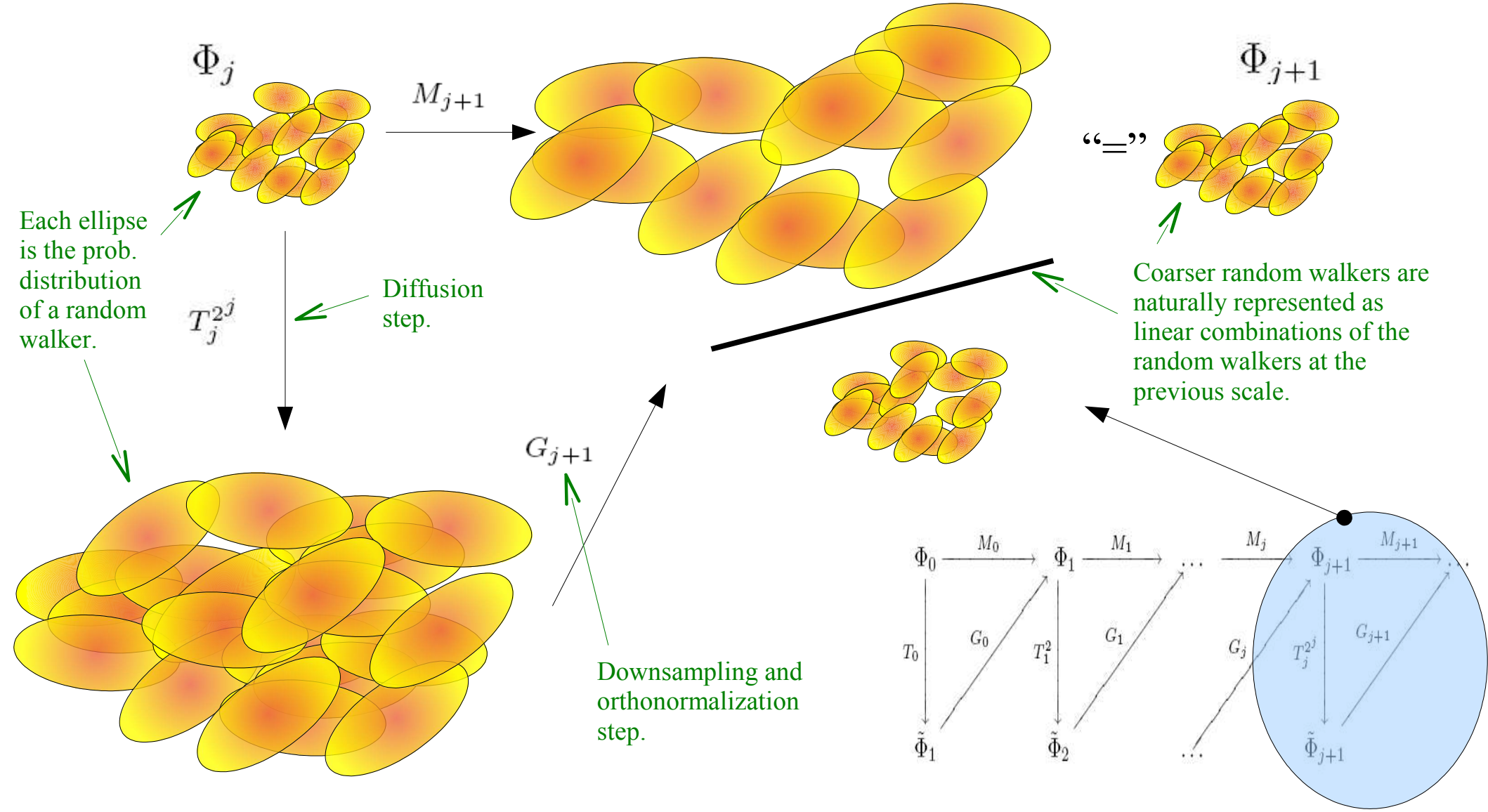
Multiscale compression scheme:
- Random walk for 1=2^0 step
- Collect together random walkers into "representatives"
- Write a random walk on the representatives
- Let the representatives random walk 2=2^1 steps...

The decay in the spectrum of ***T*** says powers of ***T*** are low-rank, hence compressible.





Multiscale Random Walkers



Dilations, translations, downsampling

We have **frequencies**: the eigenvalues of the diffusion T.
What about *dilations , translations , downsampling*?
We may have minimal information about the geometry, and only locally. Let's think in terms of functions on the set ***X***.

Dilations:
Use the diffusion operator ***T*** and its dyadic powers as dilations.

Translations and downsampling:
Idea: diffusing a basis of "scaling functions" at a certain scale by a power of T should yield a redundant set of coarser "scaling functions" at the next coarser scale: reduce this set to a Riesz (i.e. well-conditioned) or even orthonormal basis. This is downsampling in the function space, and corresponds to finding a well-conditioned subset of "translates".

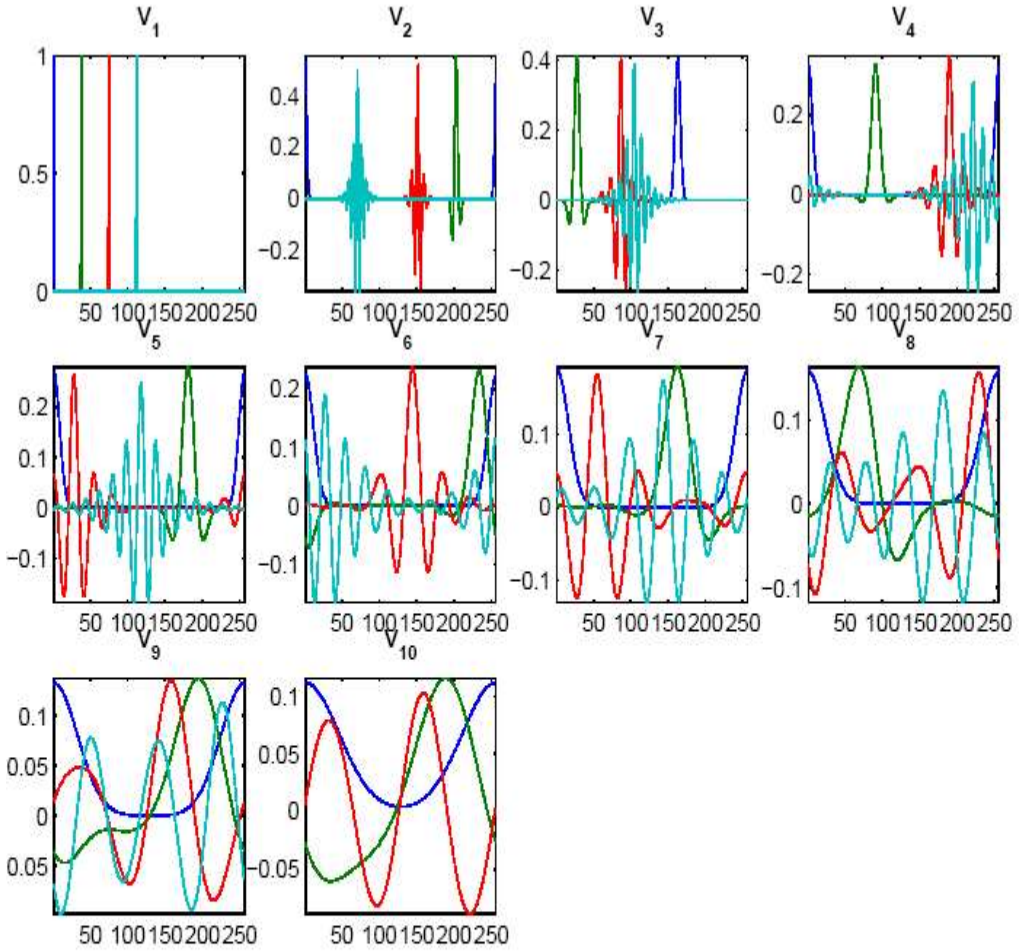


FIGURE 2. Diffusion Multiresolution Analysis on the circle. We consider 256 points on the unit circle, starting with $\varphi_{0,k} = \delta_k$ and with the standard diffusion. We plot several scaling functions in each approximation space V_j.

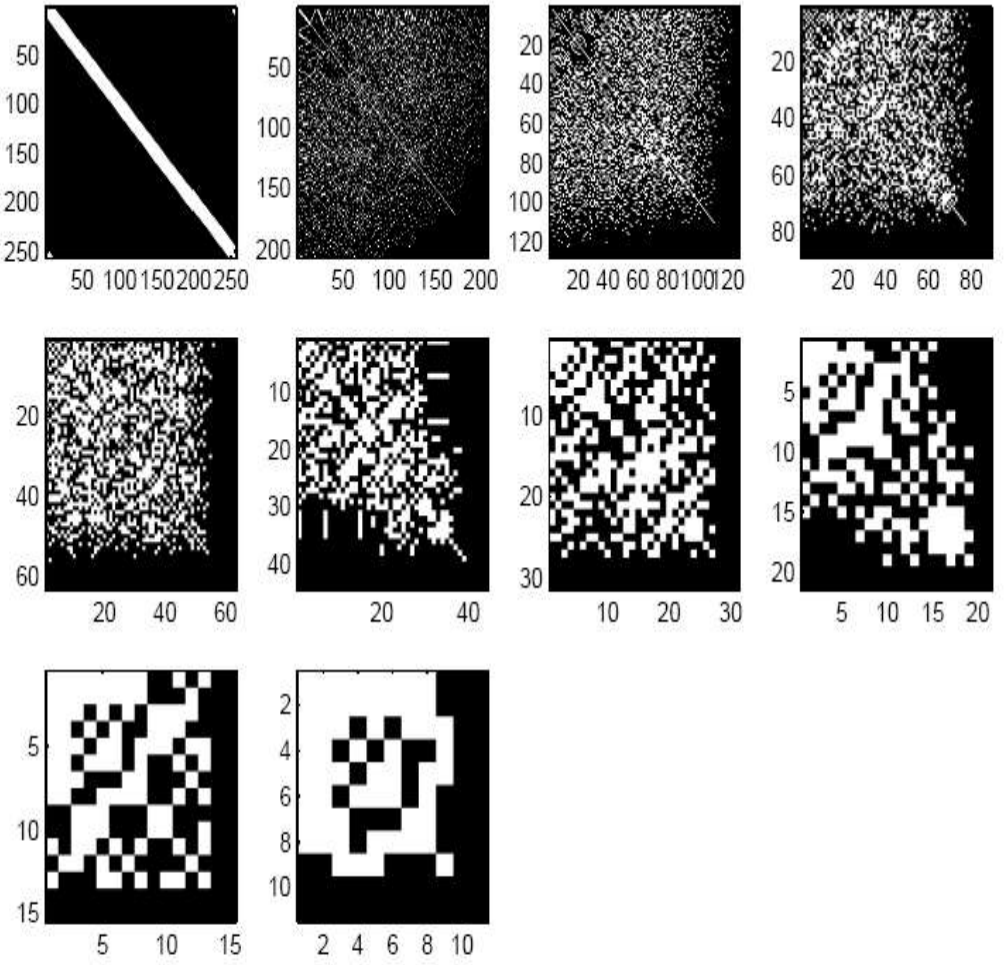


FIGURE 3. Diffusion Multiresolution Analysis on the circle: we plot the compressed matrices representing powers of the diffusion operator, in white are the entries above working precision (here set to 10^{-8}). Notice the shrinking of the size of the matrices which are being compressed at the different scales.

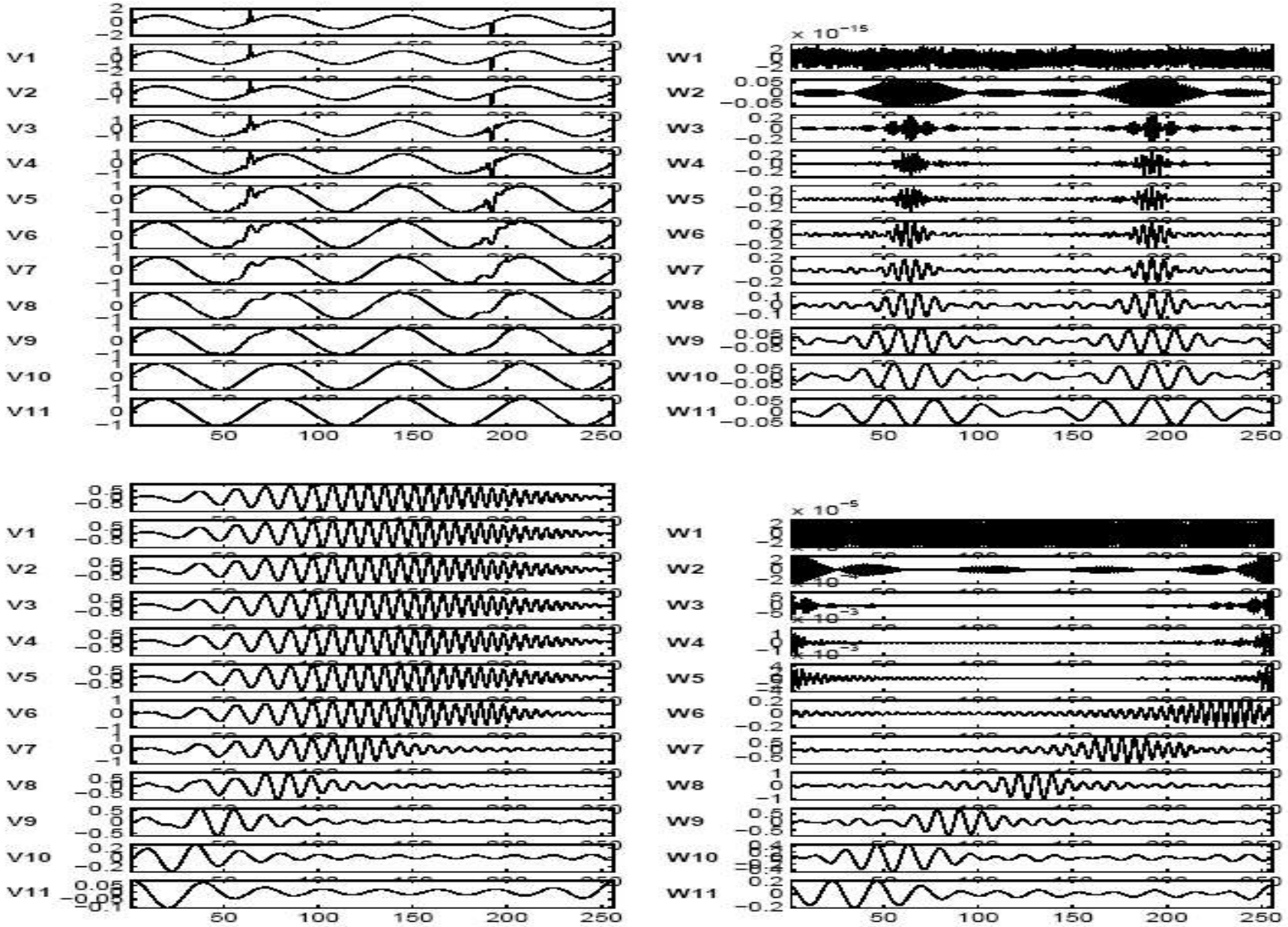


FIGURE 4. Multiresolution Analysis on the circle. In the same setting as for Figure 3, we compute the multiscale transform of a periodic signal on the circle, contaiminated by two δ-impulses (top) and of windowed chirp (bottom). In the first column we plot the projections onto coarses and coarses scaling spaces, in the second column we plot the projection on the corresponding wavelet subspaces. Computations here were done to 5 digits of precision.

Diffusion wavelets on a dumbbell manifold

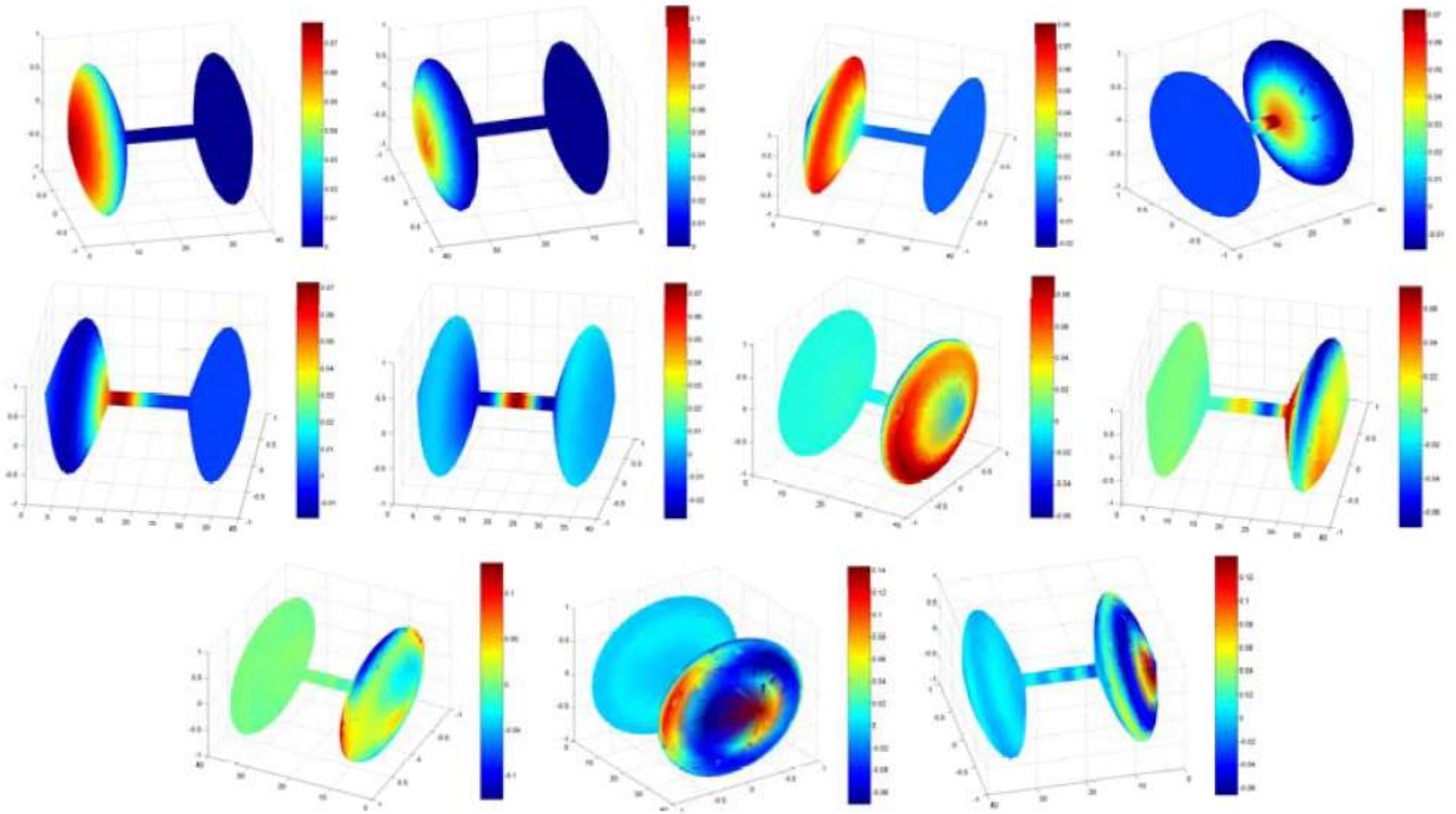
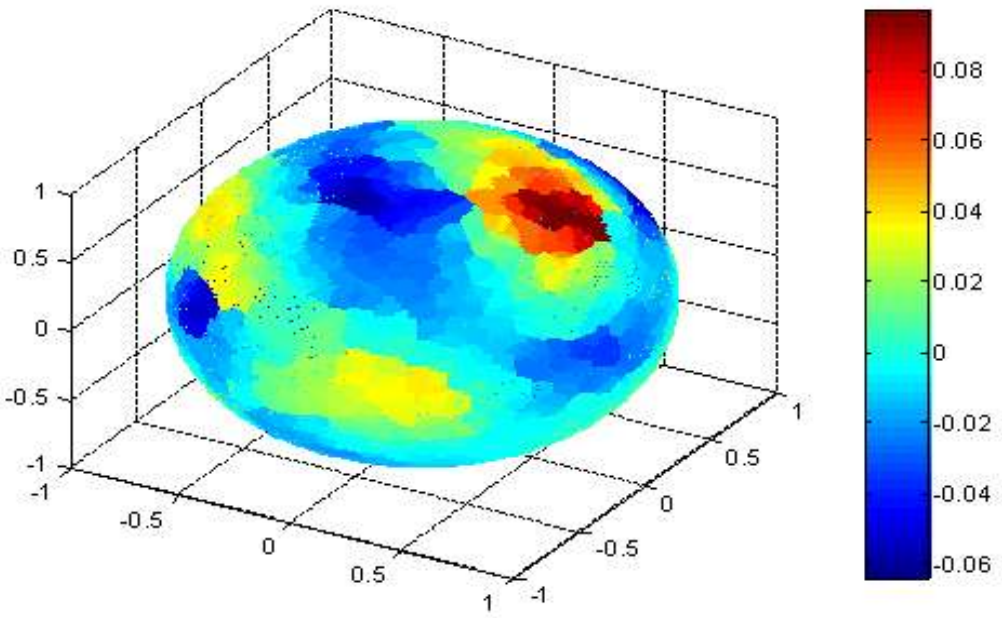
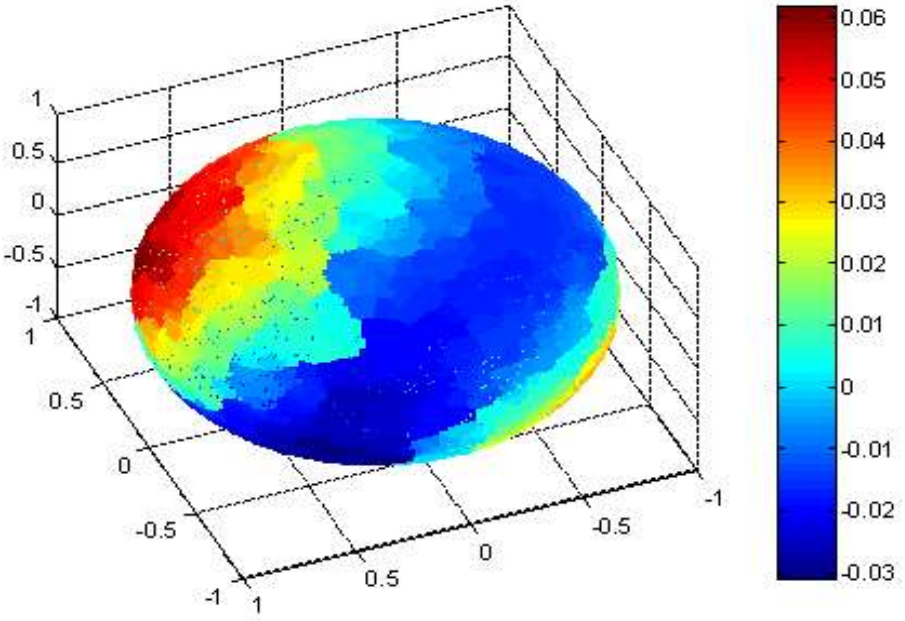
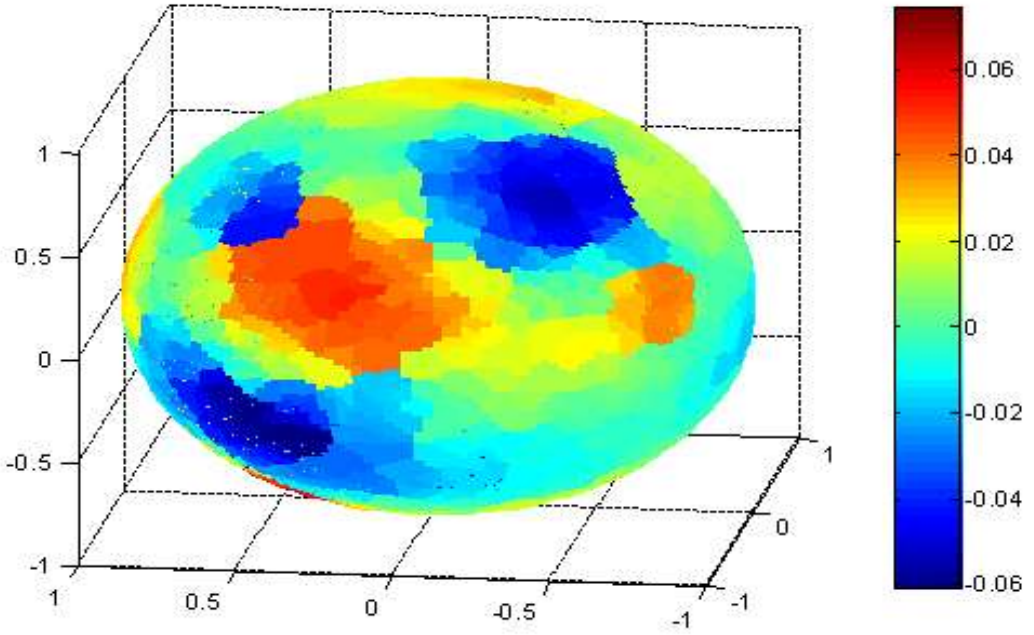


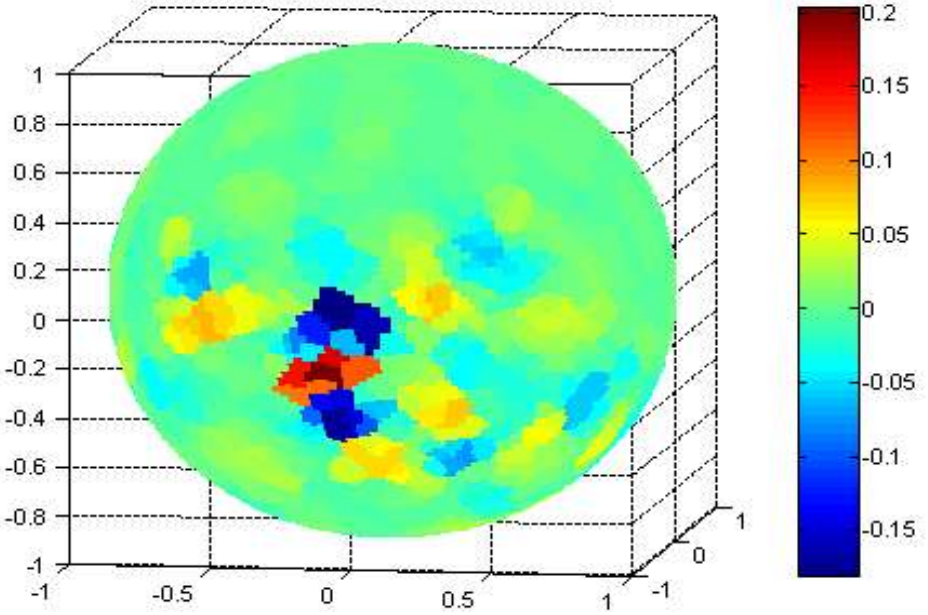
Fig. 8. Some diffusion scaling functions and wavelets at different scales on a dumbbell-shaped manifold sampled at 1400 points.

Diffusion Wavelets on the sphere









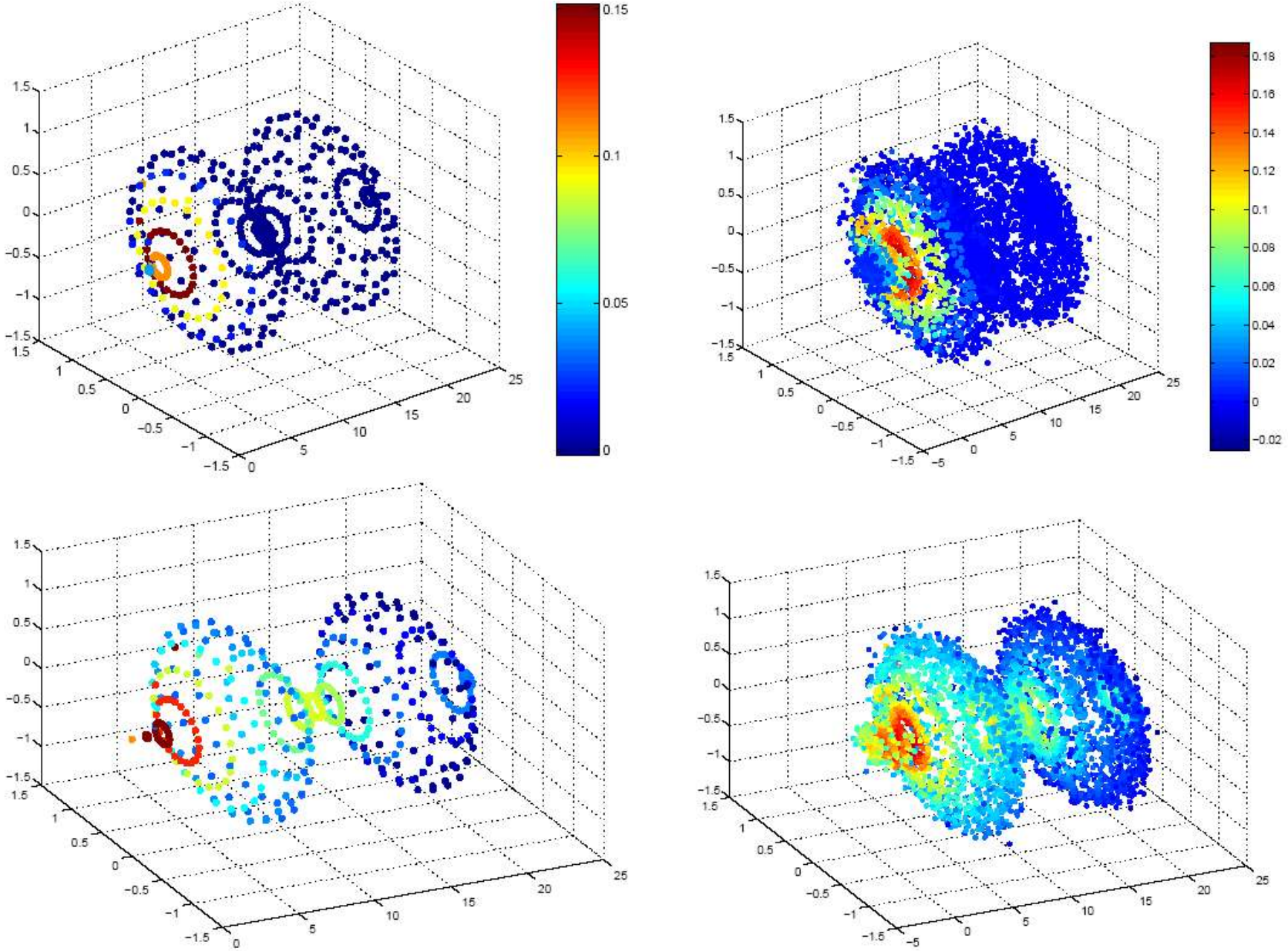


Fig. 11. Top left: the diffusion scaling function $\varphi_{3,3}$ color mapped on the dumbbell-shaped manifold sampled at 500 points. Top right: extension of $\varphi_{3,3}$ to 5000 points obtained by adding white Gaussian noise to the points the dumbell. Bottom left and right: as the corresponding pictures above, but for $\varphi_{10,1}$.

Potential Theory, Green's function

$$\frac{\partial \psi}{\partial t} = -L\psi(t)$$

The semigroup generator acts by

$$e^{-tL} f = \sum_i e^{-\lambda_i t} < \phi_i, f > \phi_i$$

Averaging over all times:

$$\frac{1}{L} = \int_0^{+\infty} e^{-tL} dt$$

[small catch: L has a kernel, one has to work in the complement.]

$$(I - T)^{-1} f = \sum_{k=1}^{+\infty} T^k f$$

and, if $S_K = \sum_{k=1}^{2^K} T^k$, we have

$$S_{K+1} = S_K + T^{2^K} S_K = \prod_{k=0}^{K} \left(I + T^{2^k} \right) f \, .$$

Many connections...

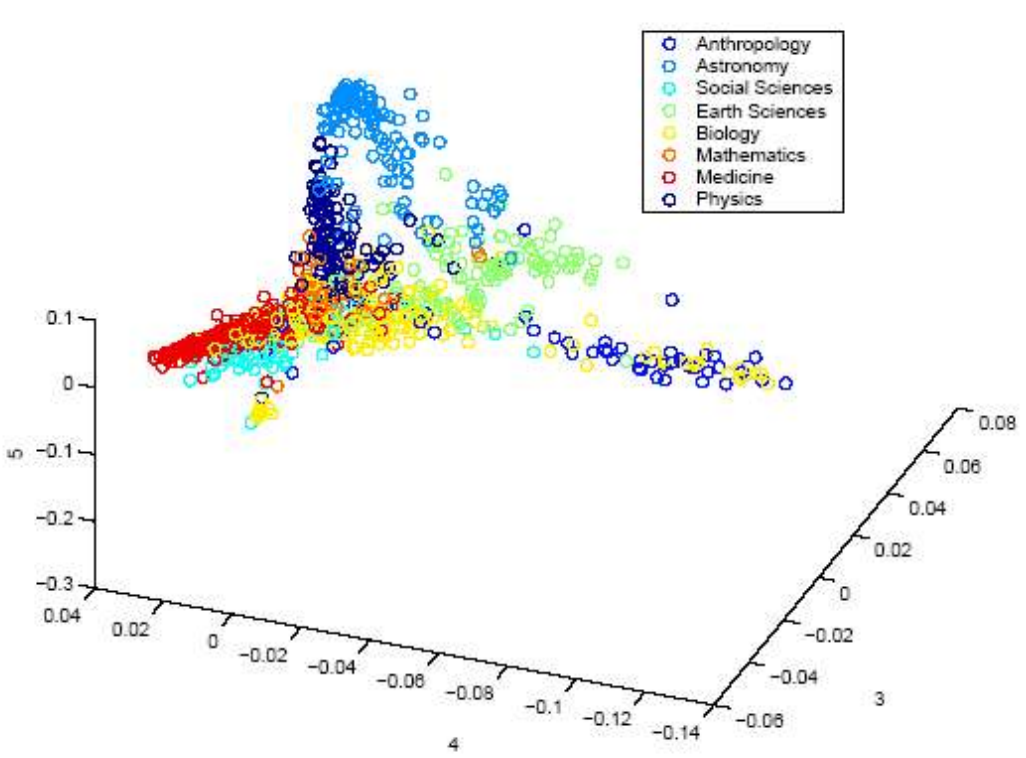
Wavelets:
- Lifting (W. Sweldens, I. Daubechies,...)
- Continuous wavelets from group actions (Grossman, Morlet, Ali, Antoine, Gazeau, many physicists...)

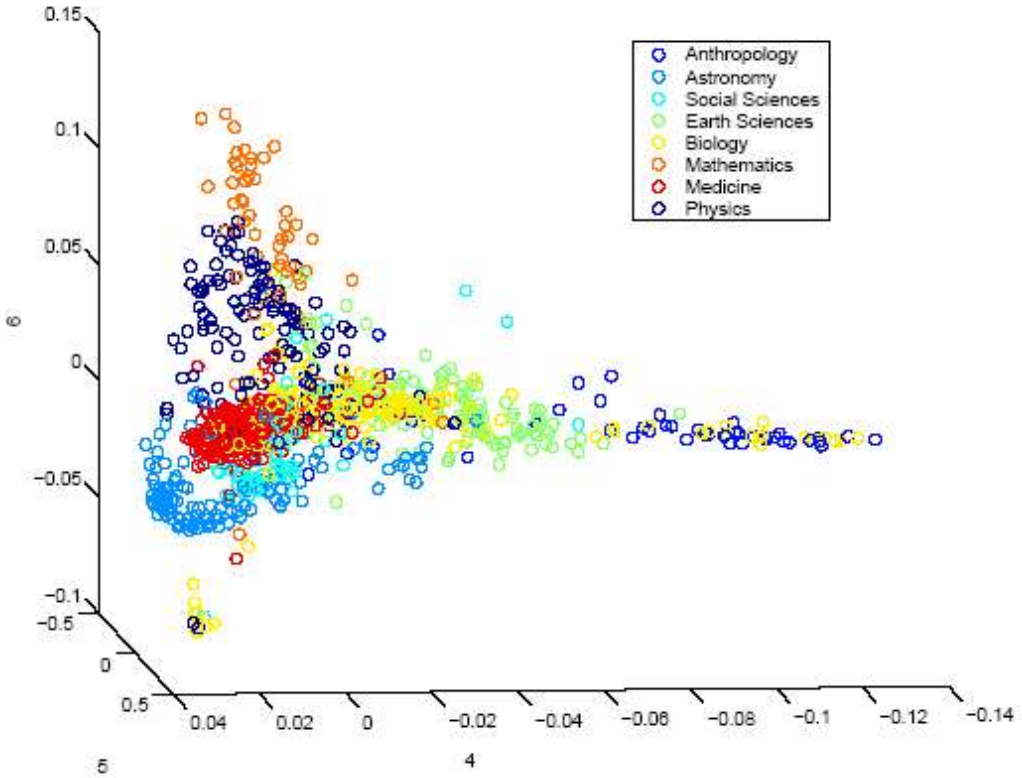
Classical Harmonic Analysis:
- Littlewood-Paley on semigroups (Stein), Markov diffusion semigroups (large literature)
- Martingales associated to the above, Brownian motion
- Heat kernel estimates, generalized Heisenberg principles (A. Nahmod)
- Harmonic Analysis of eigenfunctions of the Laplacian on domains, manifolds and graphs
- Atomic decompositions (Coifman, Weiss, Rochberg...)

Numerics:
- Algebraic Multigrid (Brandt, large literature since the 80's...)
- Kind of “inverse” FMM (Rohklin, Beylkin-Coifman-Rohklin, ...)
- Multiscale matrix compression techniques (W. Stewart, Gu-Eisenstat, Chen-Gimbutas-Martinsson-Rohklin,...)
- Randomizable
- FFTs?

Analysis of a document corpora

Given 1,000 documents, each of which is associated with 10,000 words, with a value indicating the relevance of each word in that document.
View this as a set of 1,000 in 10,000 dimensions, construct a graph with 1,000 vertices, each of which with few selected edges to very close-by documents.





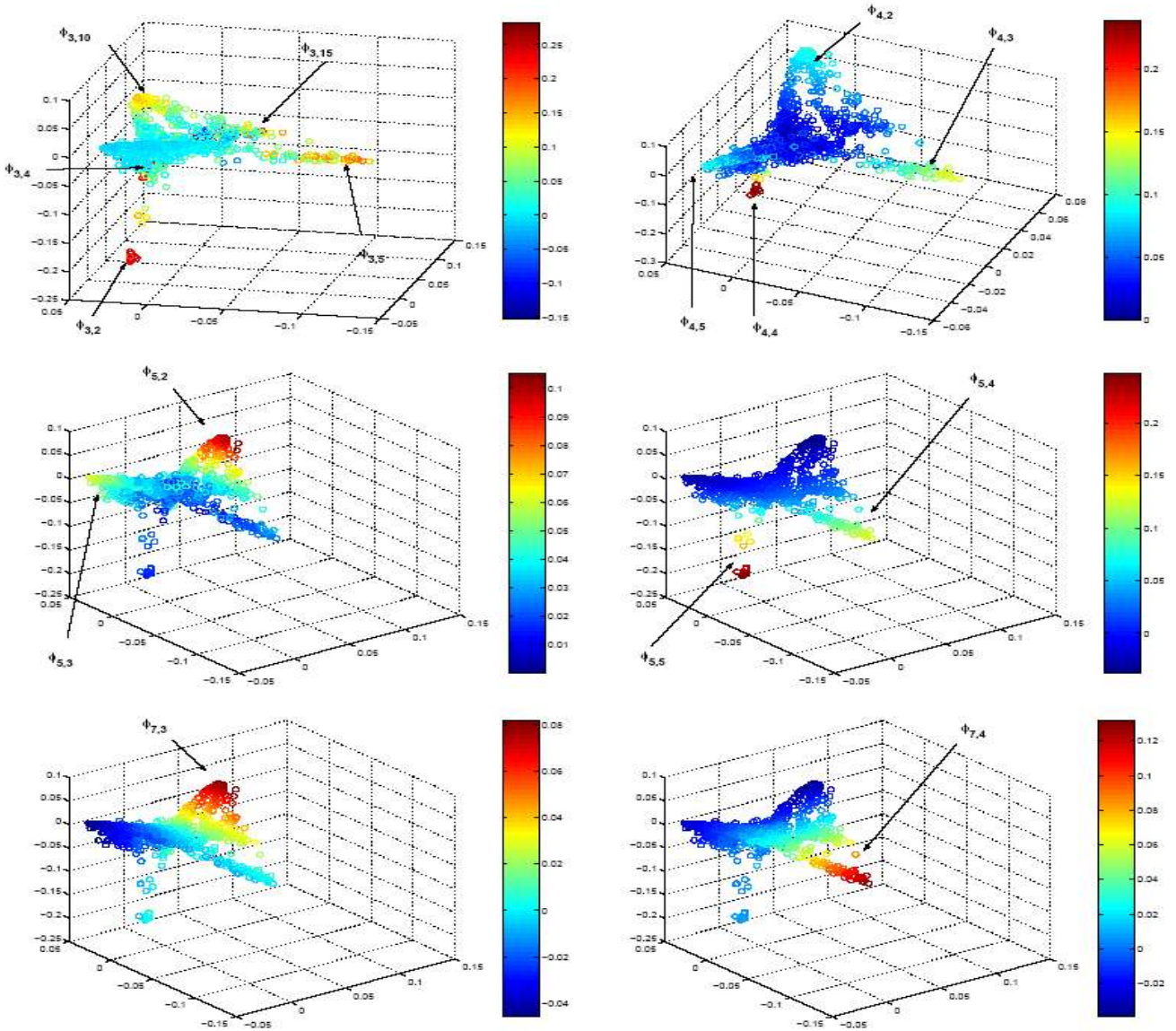


Fig. 6. Scaling functions at different scales represented on the set embedded in $\mathbb{R}^3$ via $(\xi_3(x), \xi_4(x), \xi_5(x))$.

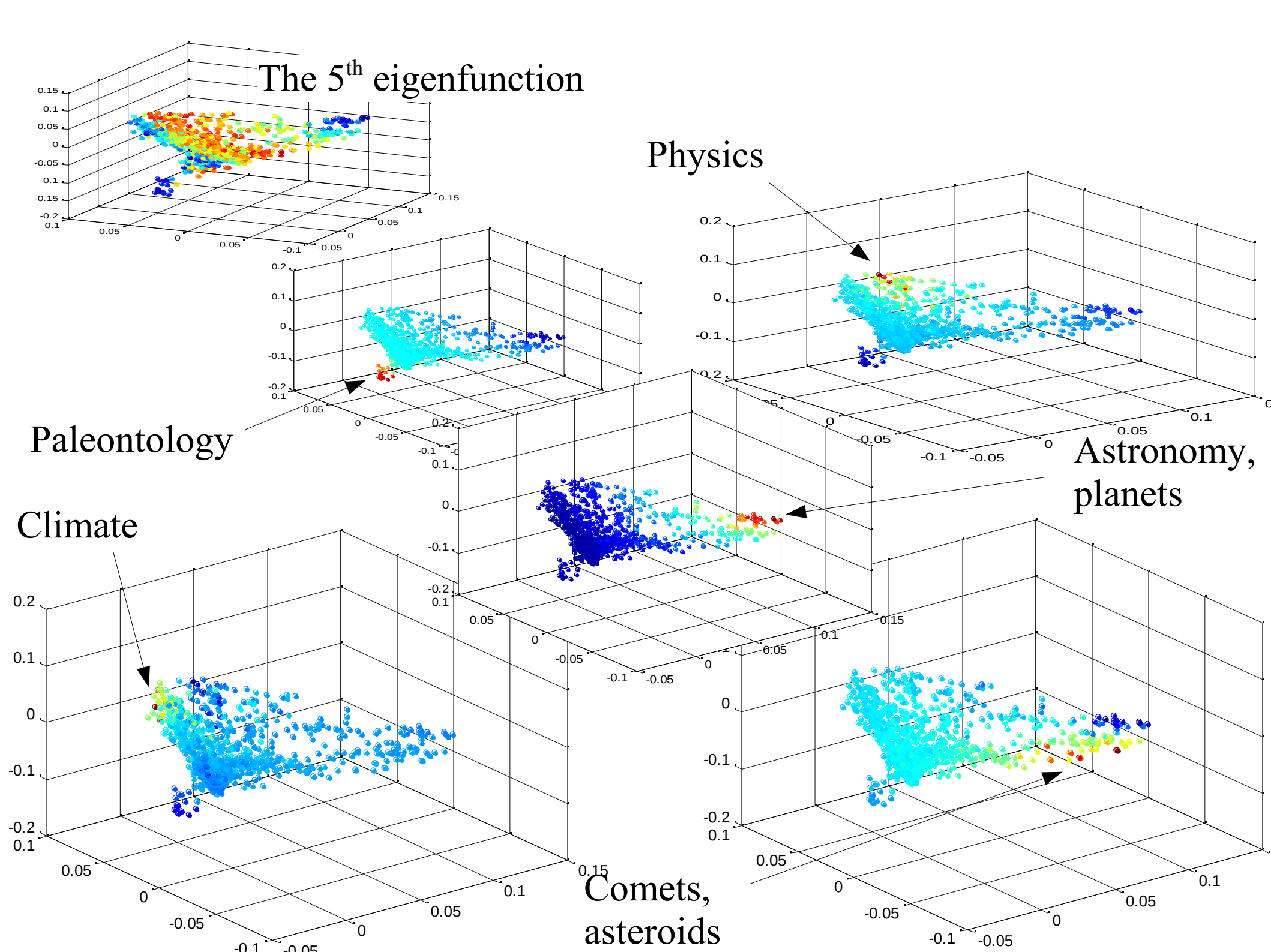
$\phi_{3,4}$ is about Mathematics, but in particular applications to networks, encryption and number theory;

$\phi_{3,10}$ 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.


The 5th eigenfunction
Physics
Paleontology
Astronomy, planets
Climate
Comets, asteroids

Diffusion Wavelet Packets

[JC Bremer, RR Coifman, MM, AD Szlam]

We can split the wavelet subspaces further, in a hierarchical dyadic fashion, very much like in the classical case. The splittings are generated by "numerical kernel" and "numerical range" operations.

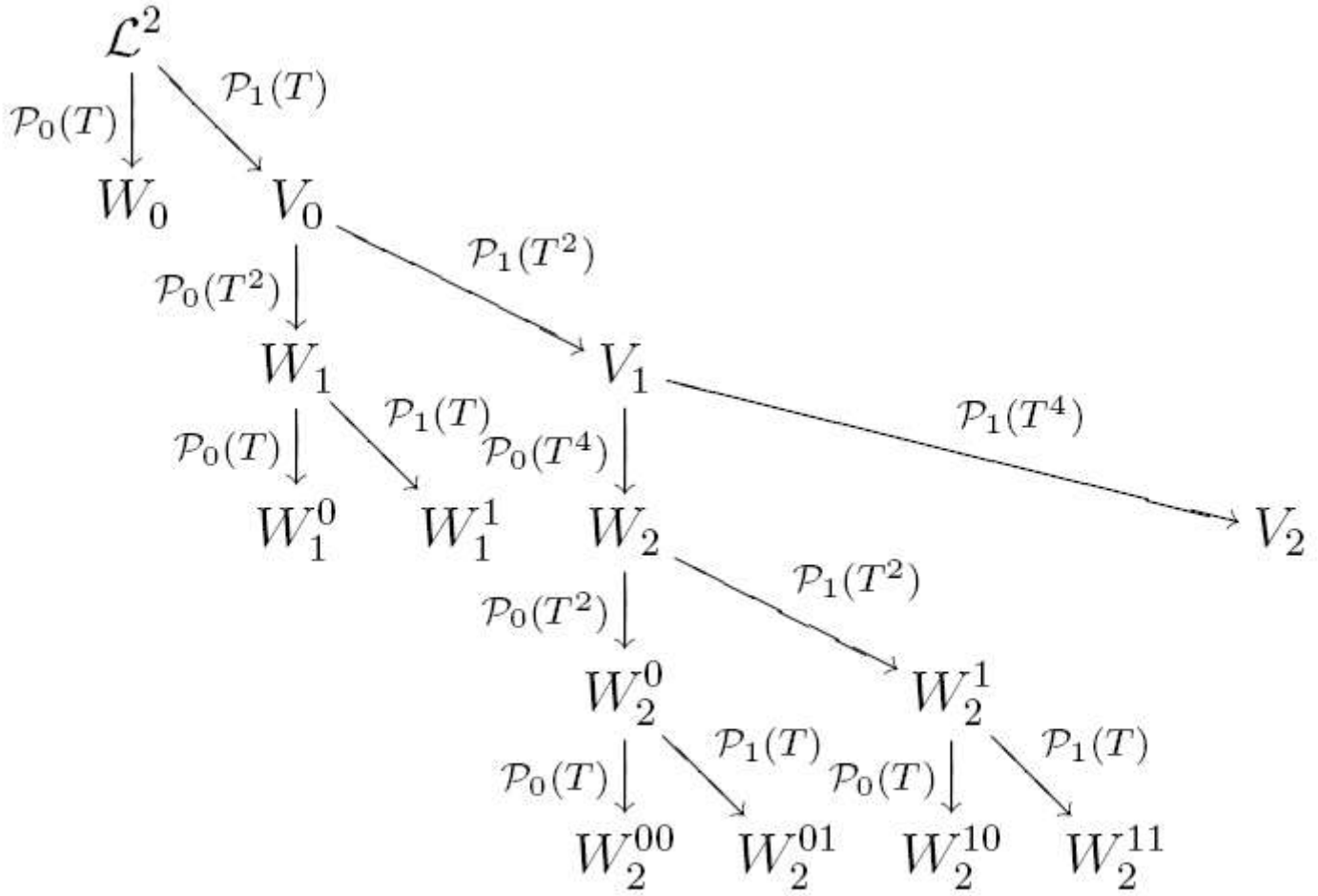


Fig. 2. Diagram for wavelet packet construction

Wavelet packets, best basis & compression

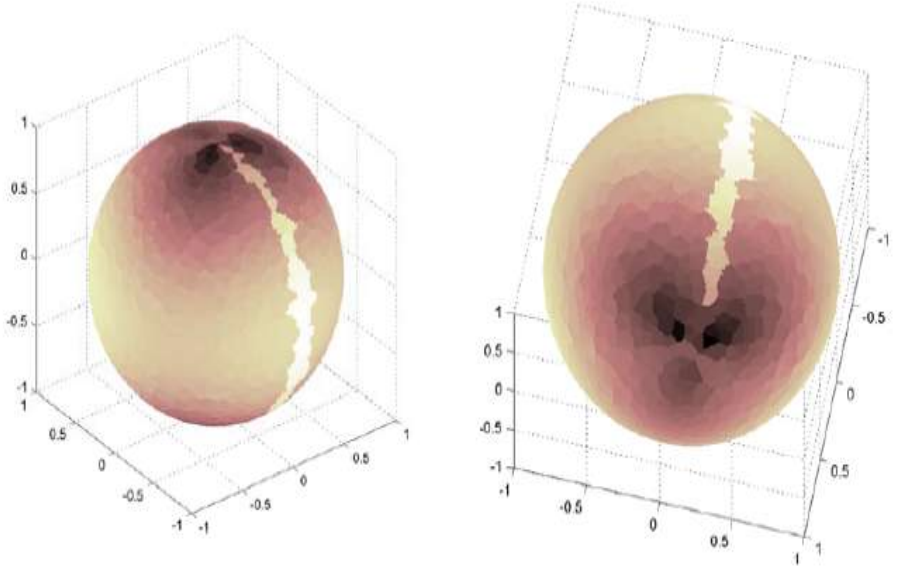


Fig. 9. Two different views of the function F on the sphere.

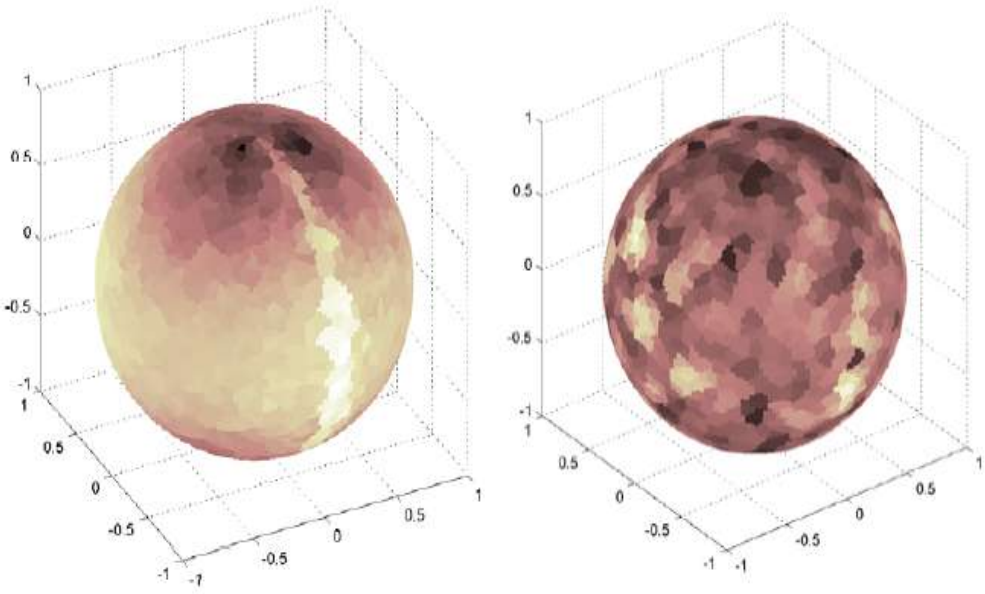


Fig. 11. Left: reconstruction of the function F with top 50 best basis packets. Right: reconstruction with top 200 eigenfunctions of the Beltrami Laplacian operator.

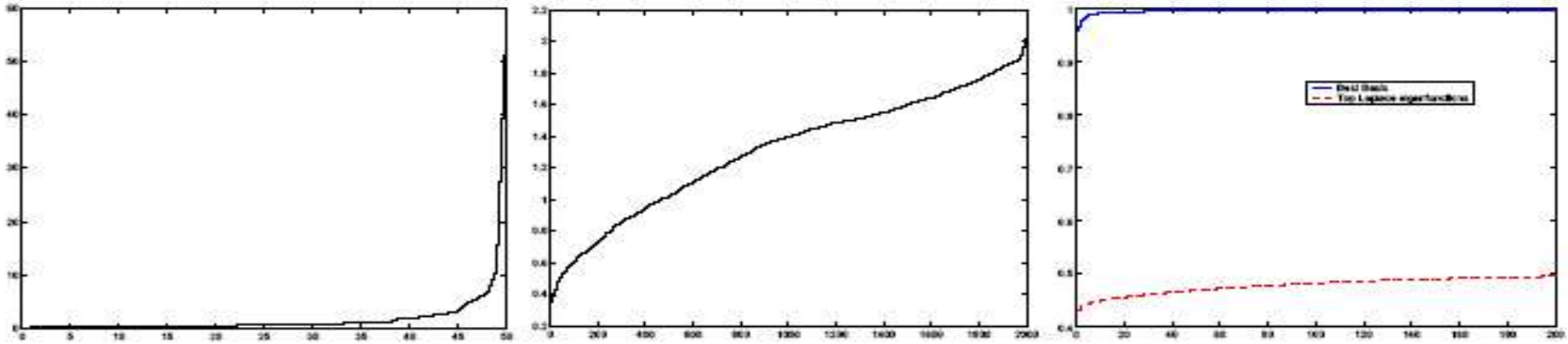


Fig. 10. Left to right: 50 top coefficients of F in its best diffusion wavelet basis, distribution coefficients$F$ in the delta basis, first 200 coefficients of F in the best basis and in the basis of eigenfunctions.

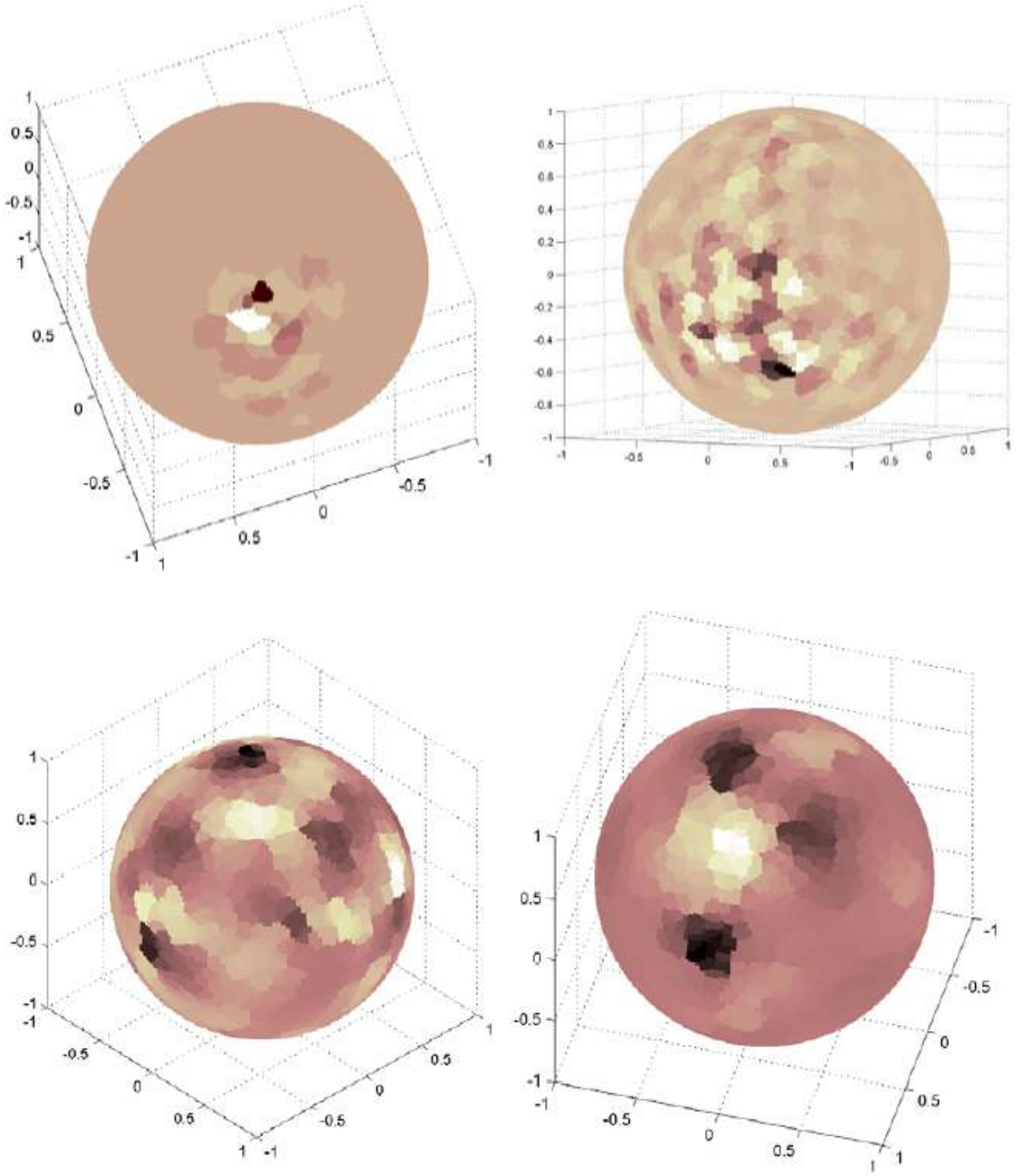


Fig. 6. Some diffusion wavelets and wavelet packets on the sphere, sampled randomly uniformly at 2000 points.

Compression example II

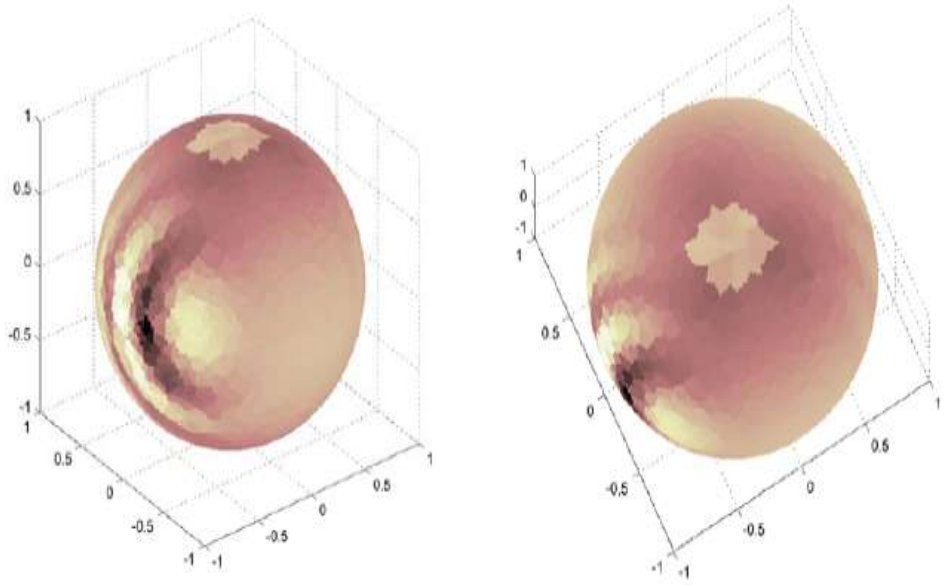


Fig. 12. Two different views of the function F on the sphere.

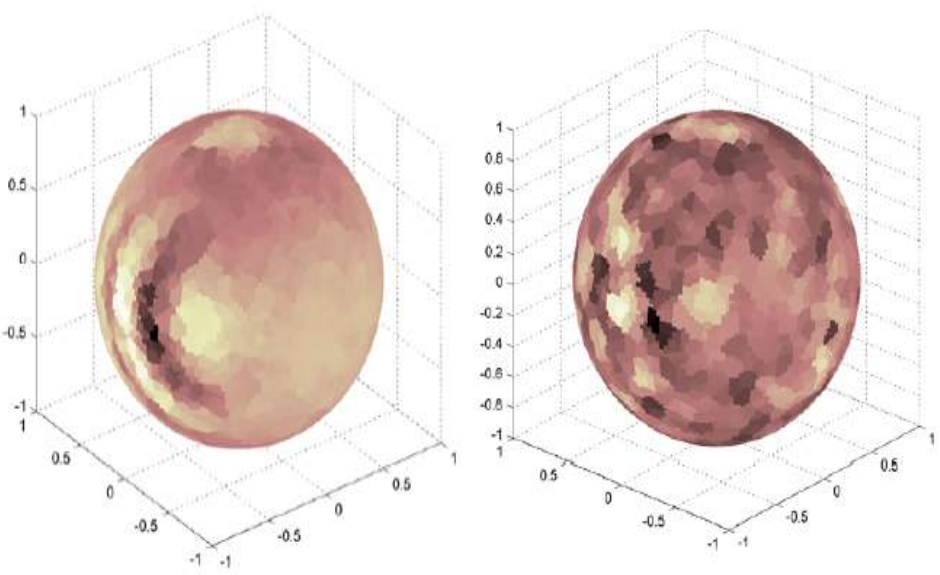


Fig. 14. Left: reconstruction of the function F from 200 best basis diffusion wavelet packet coefficients. Right: reconstruction from top 200 eigenfunctions.

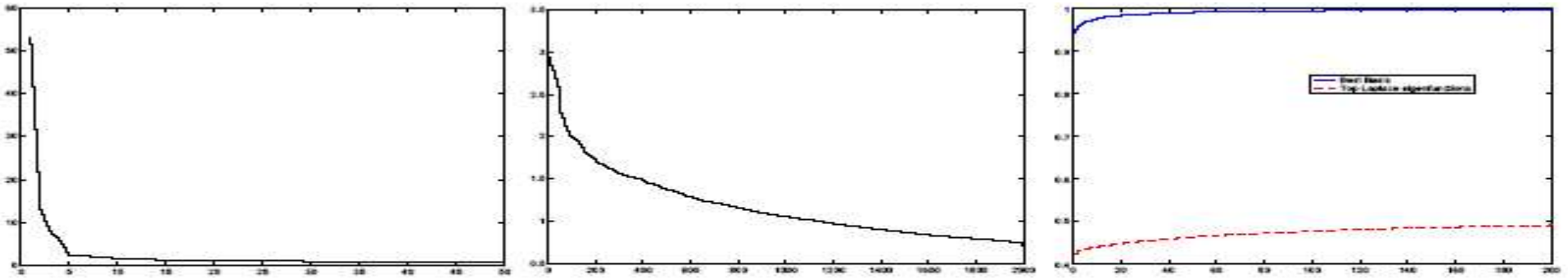


Fig. 13. Left to right: 50 top coefficients of F in its best diffusion wavelet basis, distribution coefficients$F$ in the delta basis, first 200 coefficients of F in the best basis and in the basis of eigenfunctions.

Best basis for denoising

[à la Donoho-Johnstone/Coifman-Wickerhauser]

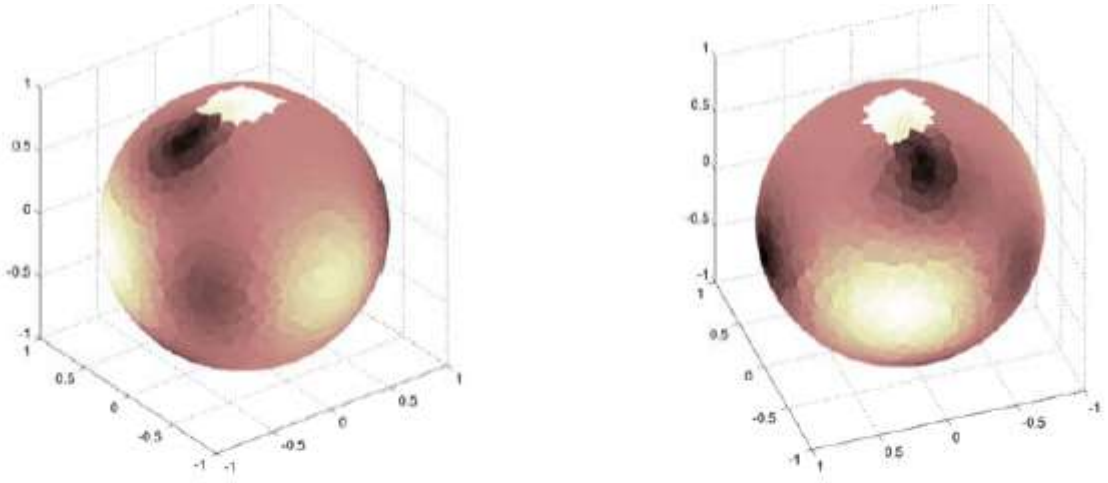


Fig. 15. Two views of G.

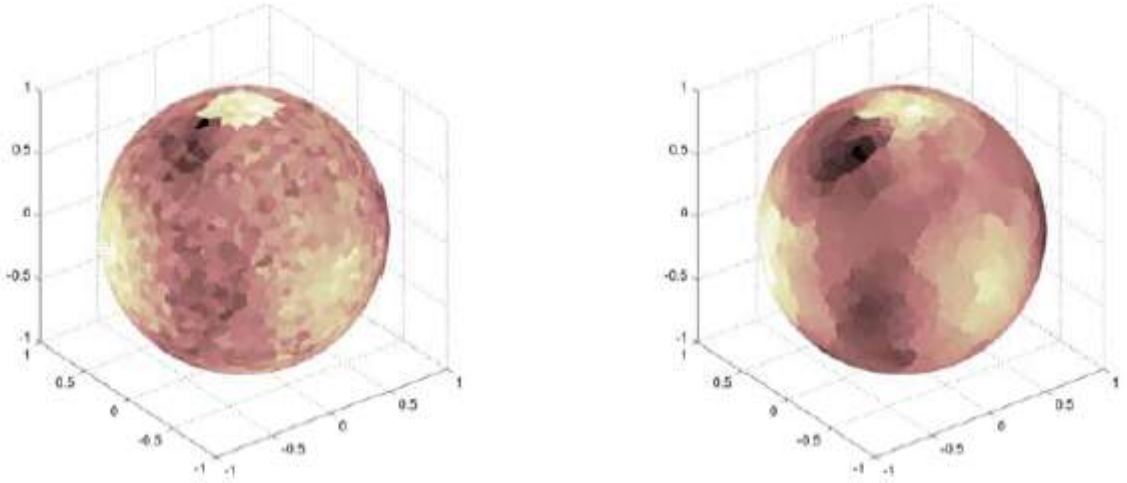


Fig. 16. Left: G with noise; right: G denoised

Best basis for discrimination

[à la Coifman-Saito]

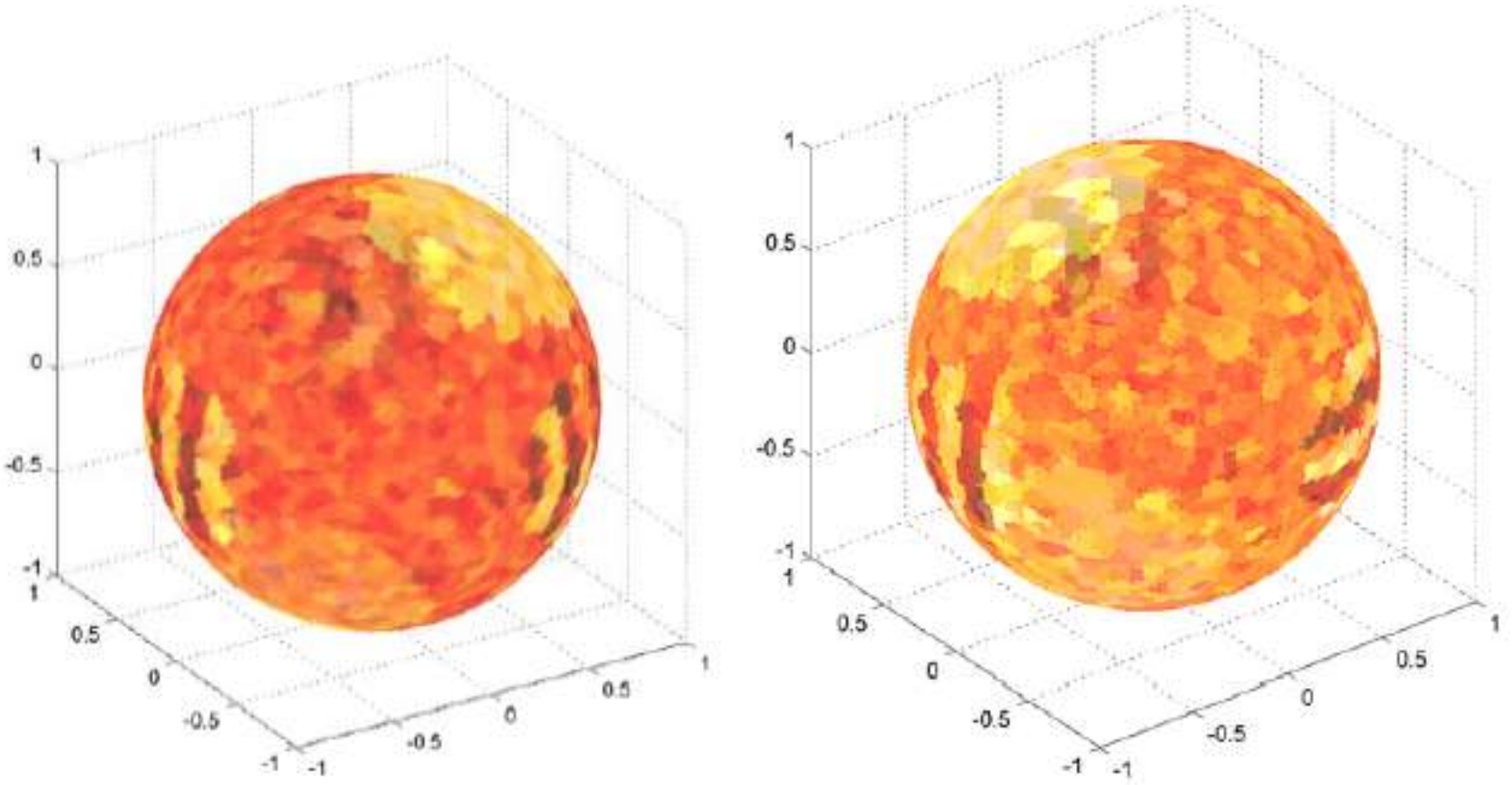


Fig. 19. Left to right, a realization of a function from class 1 and 2 respectively. Note that the third smooth texture patch is on the back side of the sphere, and can be viewed in semitransparency. The other two smooth patches are decoys in random non-overlapping positions.

Comments, Applications, etc...

- This is a ***wavelet analysis*** on manifolds, graphs, Markov chains, certain fractals, adapted to a given diffusion operator. Laplacian eigenfunctions (and other spectral kernel methods) do the corresponding Fourier Analysis.
- We are "***compressing***": powers of the operator, functions of the operators, subspaces of the function subspaces on which its powers act. This yields sampling formulas, quadrature formulas.
- Does not require the diffusion to be self-adjoint, nor eigenvectors.
- A ***biorthogonal*** version of the transform (better adapted to studying Markov chains) is in the works.
- The multiscale spaces are a natural scale of complexity spaces for ***learning*** empirical functions on the data set.
- Diffusion scaling functions ***extend outside the set***, in a natural multiscale fashion.
- Exploring ties with measure-geometric considerations used for ***embedding*** metric spaces in Euclidean spaces with small distortion.
- Application to Markov decision processes.

Current & Future Work

- Biorthogonal construction (better localization, better constants).
- Multiscale embeddings for graphs, measure-geometric implications.
- Martingale aspects and efficient Brownian motion simulation in nonhomogeneous media.
- Applications to learning and regression on manifolds.
- Robustness, perturbations, extensions.
- Compression of data sets.
- Going nonlinear.

This talk, papers, Matlab code available at:

www.math.yale.edu/~mmm82

Thank you!