Diffusion Wavelets and Applications

J.C. Bremer, R.R. Coifman, P.W. Jones, S. Lafon,
M. Mohlenkamp, MM, R. Schul, A.D. Szlam

Demos, web pages and preprints available at:
S.Lafon: www.math.yale.edu/~sl349 (eigenfunctions of Laplacian)
MM: www.math.yale.edu/~mmm82 (diffusion wavelets)

Goal

Do multiscale analysis intrinsically on manifolds, varifolds, "datasets".

Motivations

Analyze large amount of data, and functions on this data, intrinsically rather low-dimensional, embedded in high dimensions.

Paradigm: we have a large number of ***documents*** (e.g.: web pages, gene array data, (hyper)spectral data, molecular dynamics data etc...) and a way of measuring ***similarity*** between pairs. Model: a graph

(G,E,W)

In important practical cases: vertices are points in high-dimensional Euclidean space, weights are a function of Euclidean distance.

Difficulties

Data sets in high-dimensions are complicated, as are classes of interesting functions **on** them.

We want to do approximation and "learning" of such functions. Parametrize low dimensional data sets embedded in high-dimension. "Fast" algorithms

High dimensional data: examples

- Documents, web searching
- Customer databases
- Financial data
- Satellite imagery
- Transaction logs
- Social networks
- Gene arrays, proteomics data
- Art transactions data
- Traffic (automobilistic, network) statistics
-
-
-

Laplacian, diffusion geometries

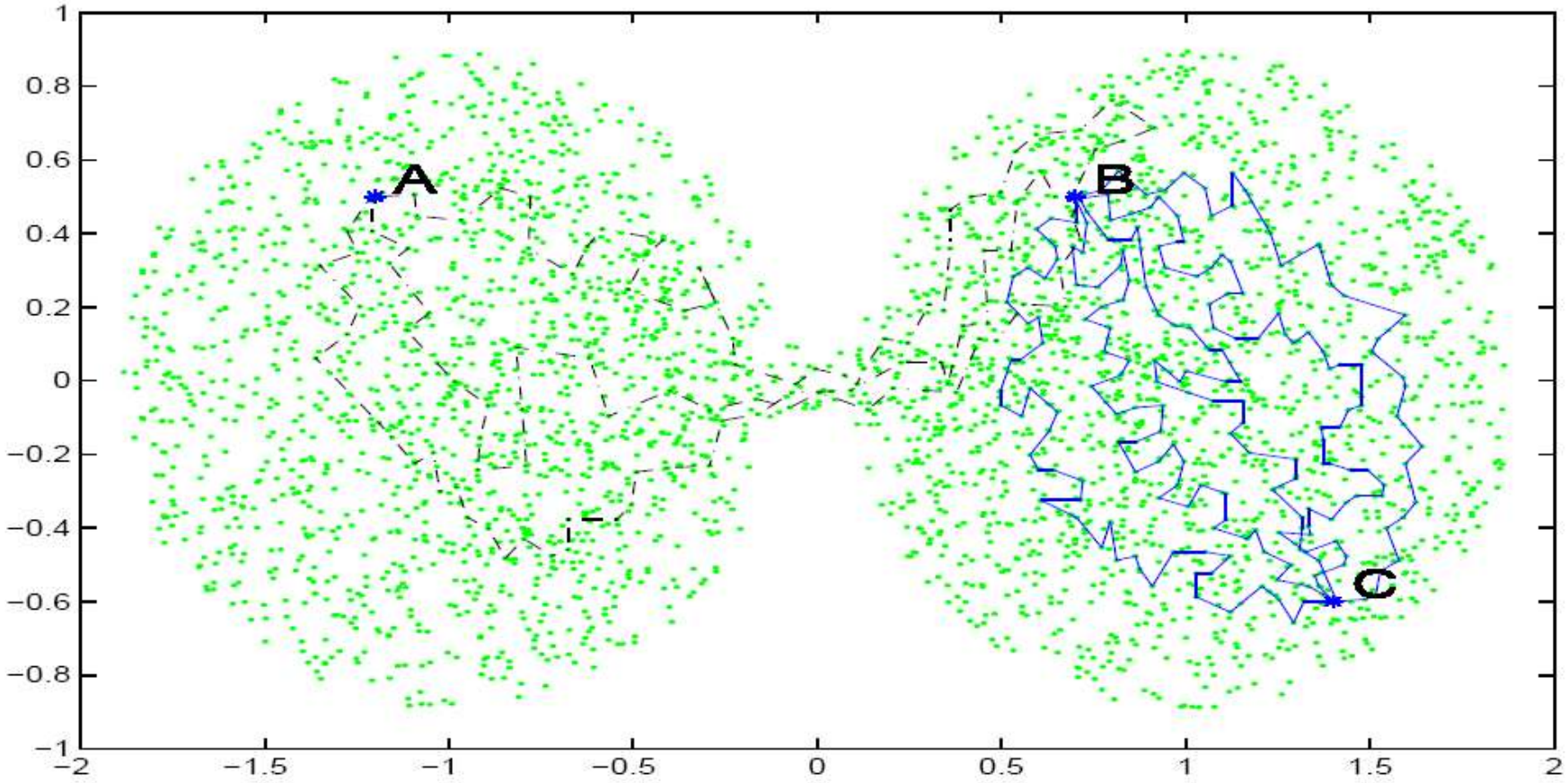
[RR Coifman, S. Lafon]

References: Belkin, Nyogi; Stephane's web page: www.math.yale.edu/~sl349

Part of the material in the next few slides is courtesy of Stephane Lafon.

From local to global: diffusion distances

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



A local "similarity" operator on the set

$X = \{x_1, x_2, ..., x_N\}$ data set.

$k(x, y)$ kernel defined on the data and being

symmetric $k(x, y) = k(x, y)$

positivity-preserving: $k(x, y) \geq 0$

positive semi-definite: $\sum_{x \in X} \sum_{y \in X} \alpha(x) \overline{\alpha(y)} k(x, y) \geq 0$

The kernel k describes the geometry of X by defining the relationship between the data points.

Examples of similarity kernels

Examples:

- If X lies in n-dimensional Euclidean space, k(x,y) could be:
  - exponentially weighted distance: $\exp(-(\|x-y\|/a)^2$
  - angle: $\langle x,y\rangle/(\|x\|\,\|y\|)$
  - harmonic potential: $1/(a+\|x-y\|)$, or powers of it
  - "feature distances": any of the above applied in the ragne f(X), f nonlinear, possibly mapping X to higher dimension

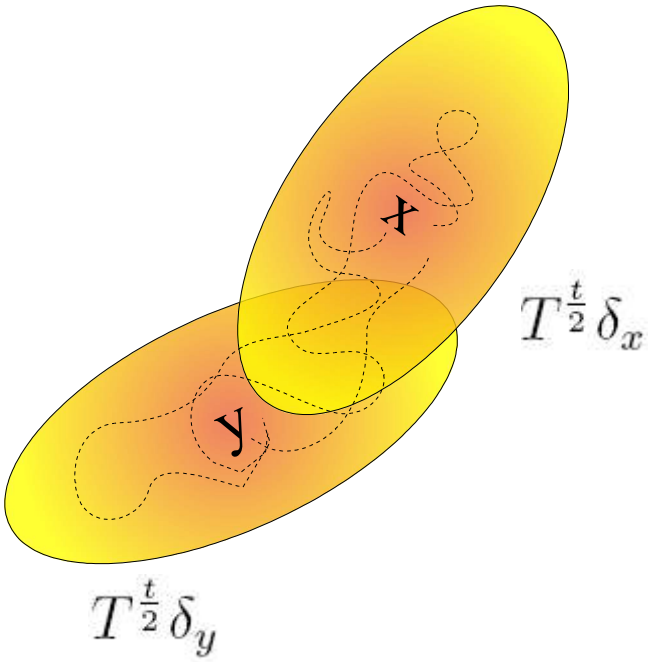
- If X is an abstract graph: we need to be given some weights on edges, measuring similarity.

This could wildly vary: from the graph obtained by discretizing a PDE, to ways of measuring similarity between two protein chains...

Diffusion Distances

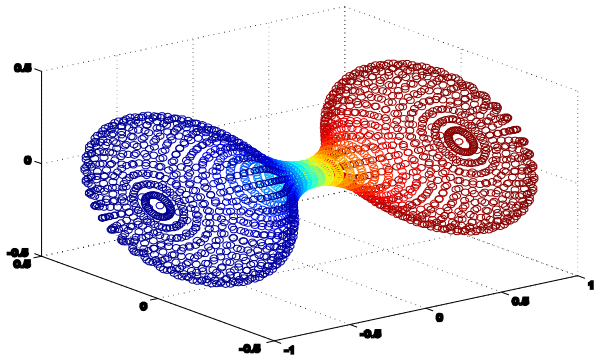
$$K^t(x,y) = \sum_{\lambda \in \sigma_T} \lambda^t \xi_\lambda(x) \xi_\lambda(y)$$

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

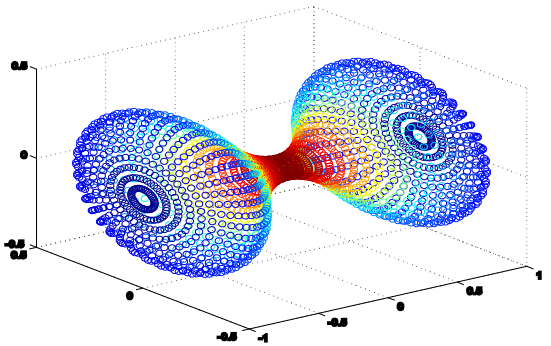


Diffusion embedding mapping ***X*** with diffusion distance into Euclidean space with Euclidean distance:

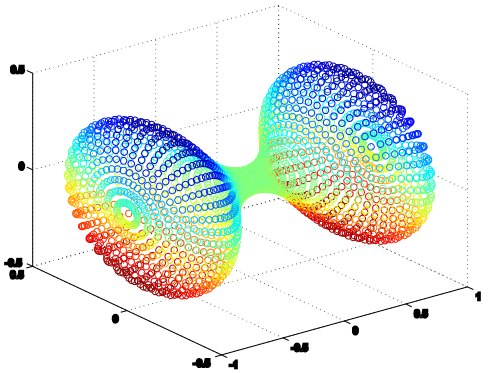
$$\Phi_m^{(t)}(x) = \left(\lambda_0^t \xi_0(x), \lambda_1^t \xi_1(x), \ldots, \lambda_{m-1}^t \xi_{m-1}(x)\right) \in \mathbb{R}^m$$



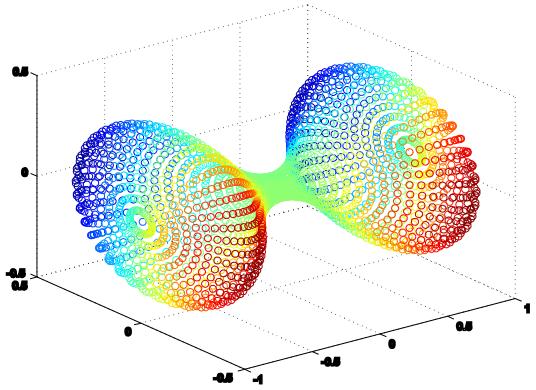
Phi1

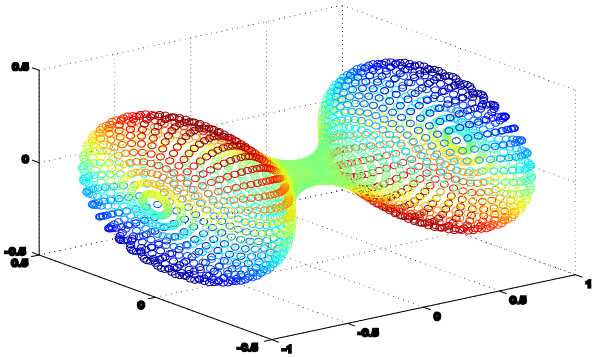


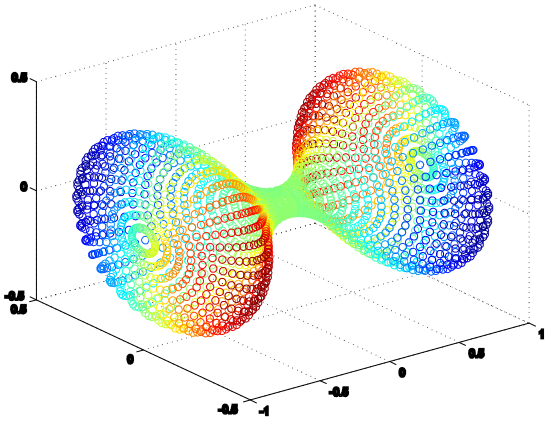
Phi2

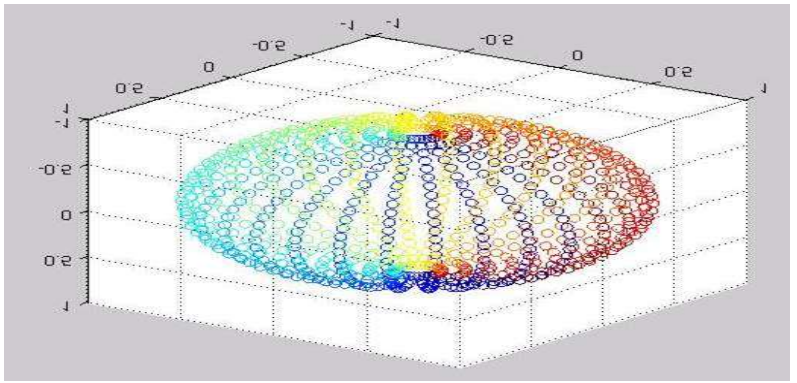


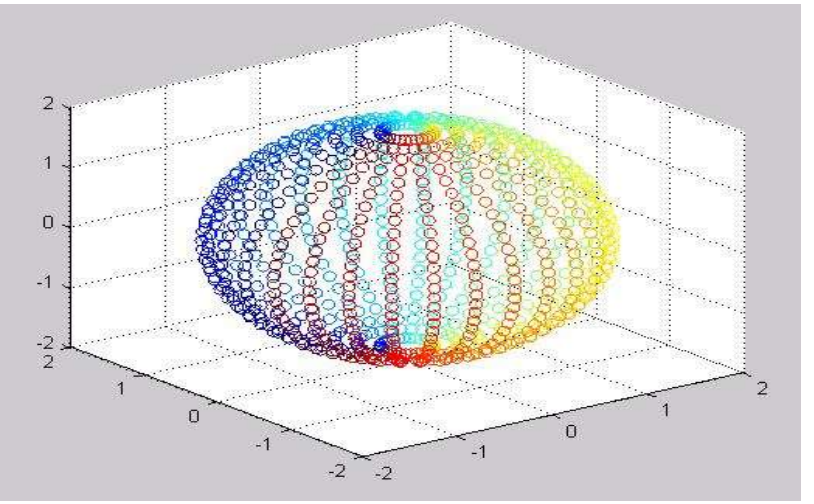
Phi3





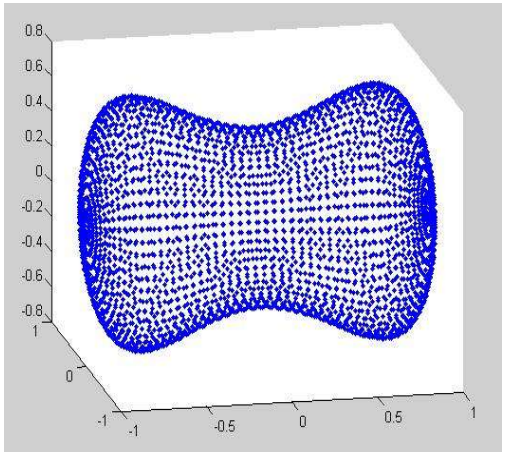


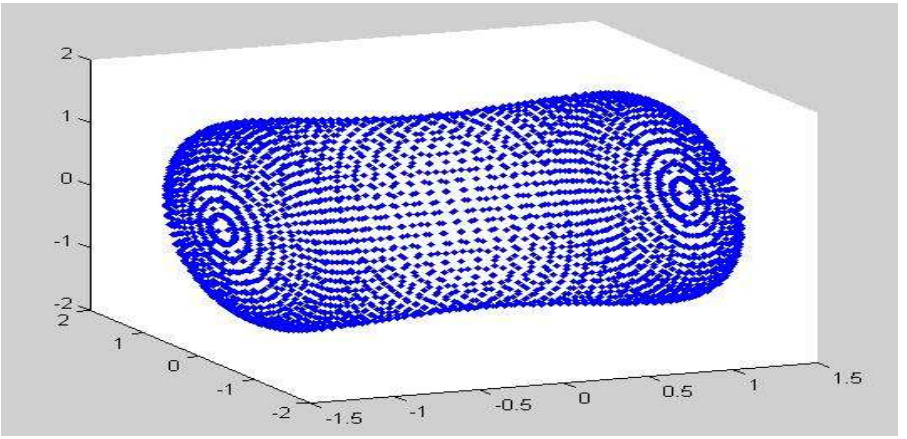


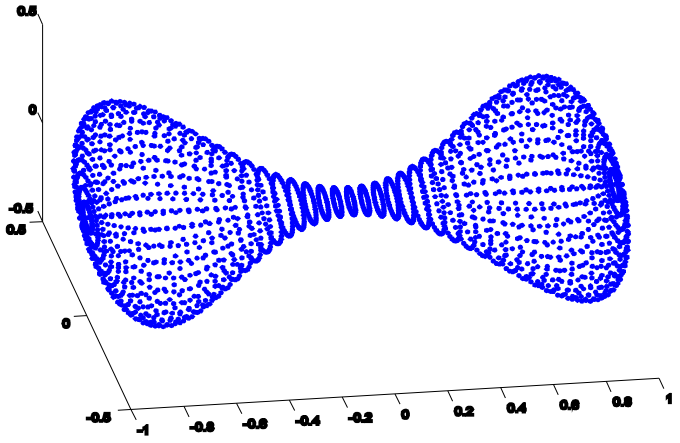


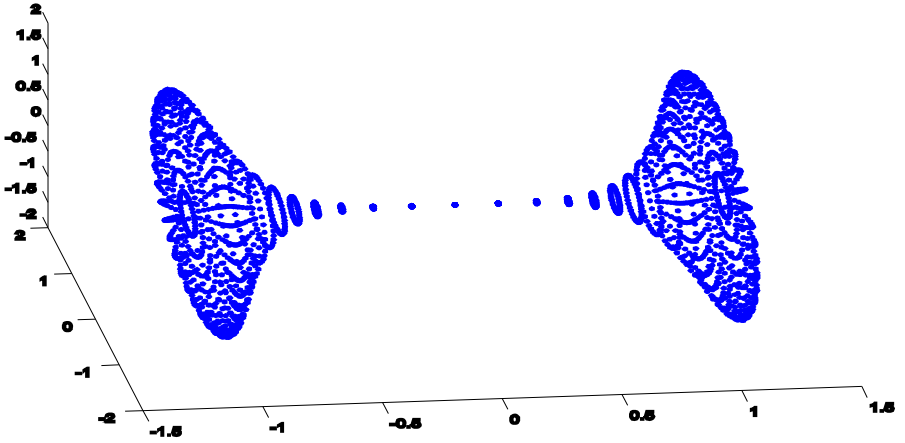
Original points

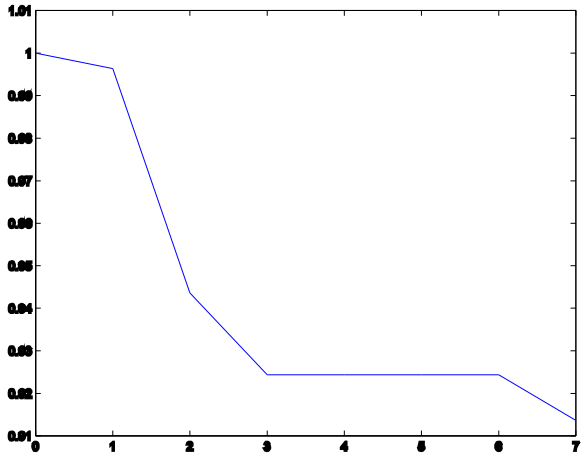
Embeddings






0.5
0
-0.5
0.5
0
-0.5
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


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


1.01
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0
1
2
3
4
5
6
7

Link with other 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

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

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

So far so good...

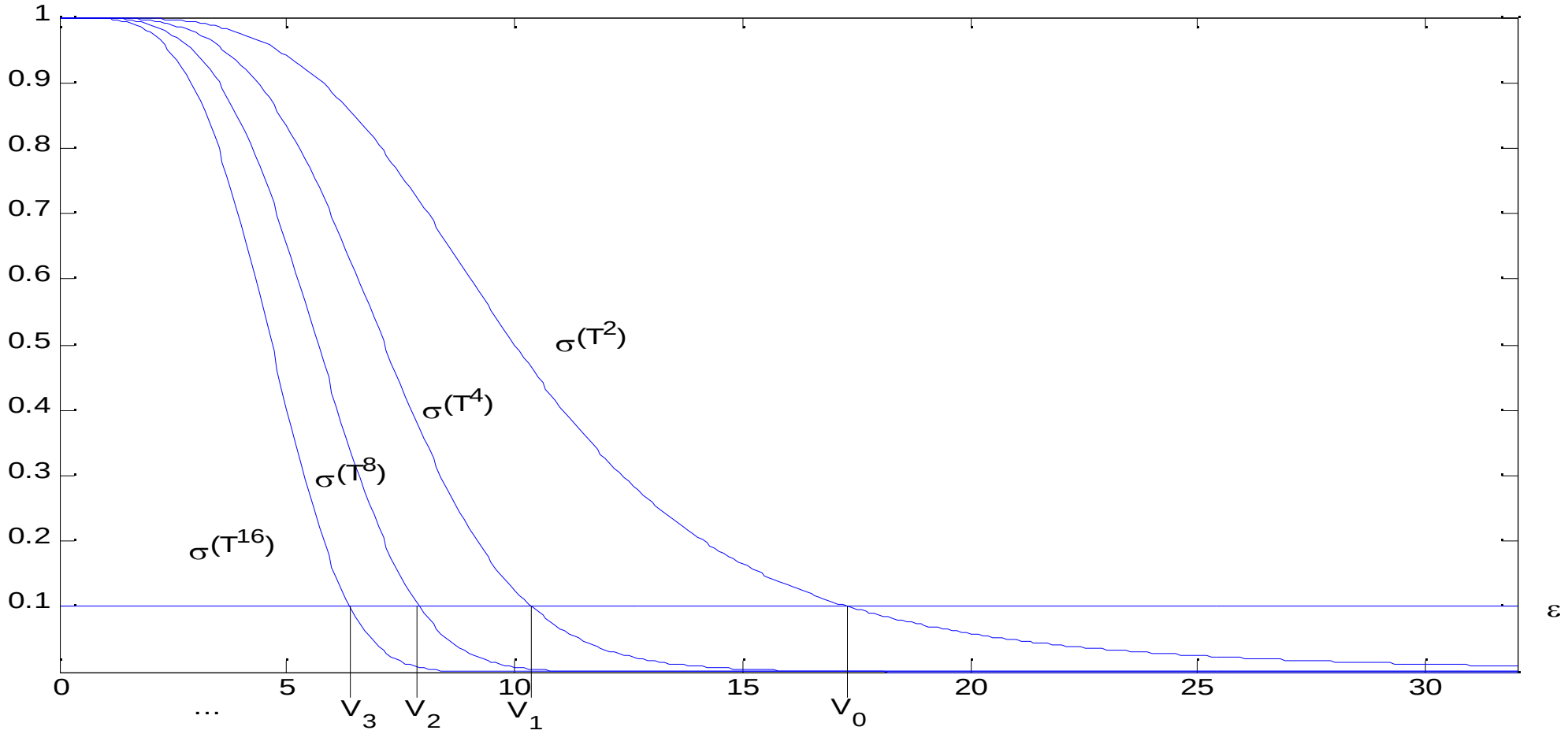
We see that:
- it seems useful to consider a framework in which, for a given data set, similarities are given only between very similar points
- it is possible to **organize** these local information by diffusion into global parametrizations,
- these parametrizations can be found by looking at the **eigenvectors** 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: **NO MULTISCALE SO FAR!**

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

From global Diffusion Geometries...

Recall: we are given a graph X with weights W. There is a natural random walk P on X induced by these weights. P maps probability distributions on X to probability distributions on X. We let T be P renormalized to have largest eigenvalue 1. The spectra of T and its powers look like:

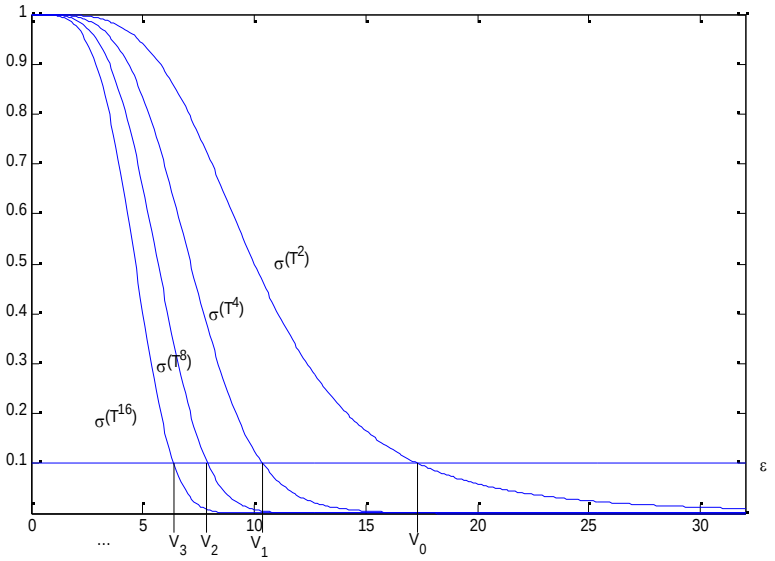


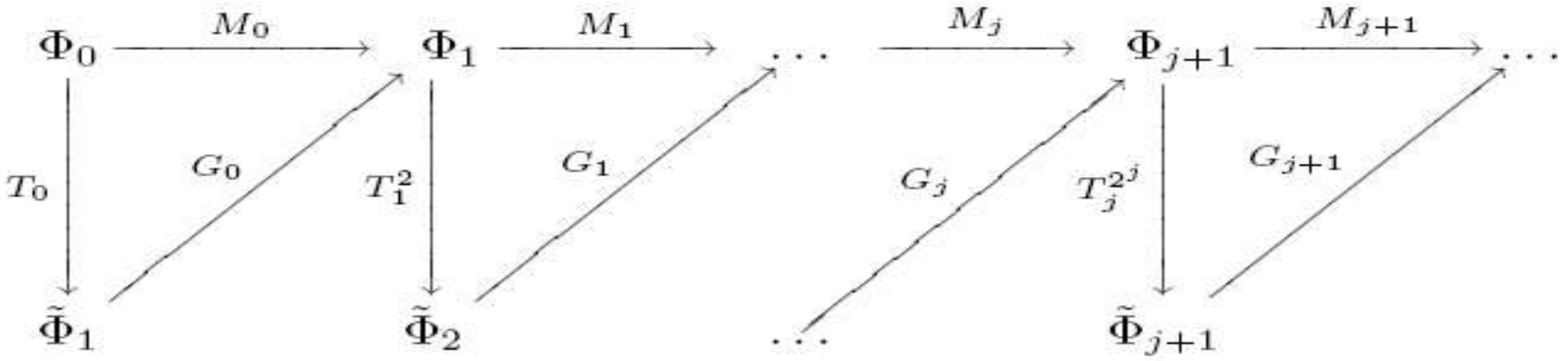
... to Multiresolution Diffusion

[Coifman,MM]

The decay in the spectrum of $\boldsymbol{T}$ says powers of $\boldsymbol{T}$ are low-rank, hence compressible.

Random walk for one step, collect together random walkers into representatives, let the representatives random walk twice, collect them into representatives, and so on....





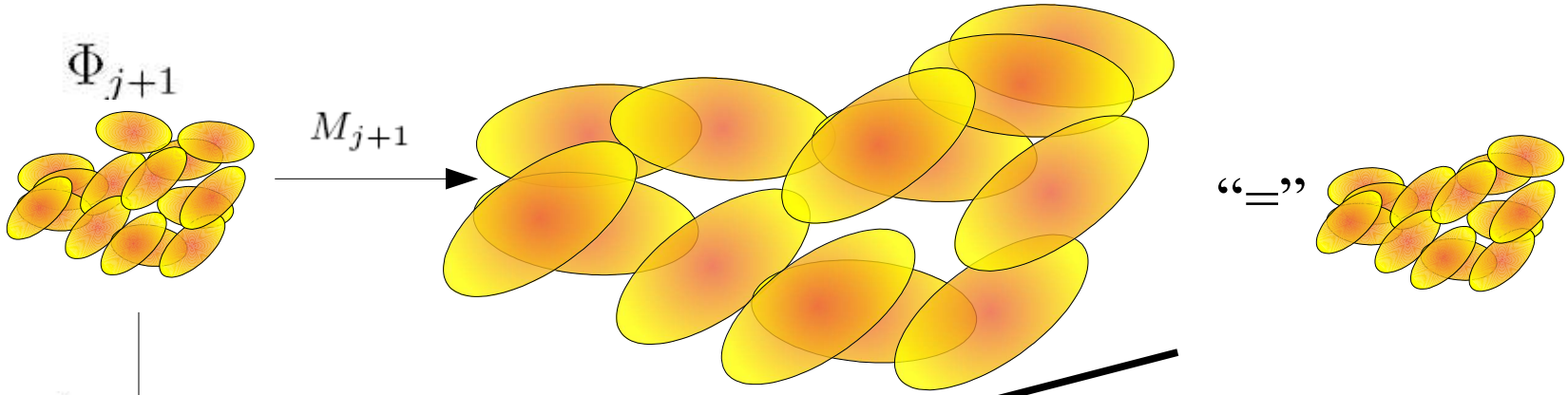
Dilations, translations, downsampling

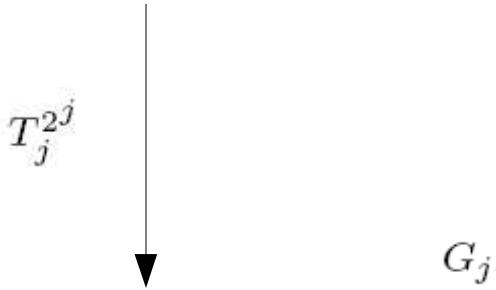
We have **frequencies**: the eigenvalues of the diffusion T.
What about dilations , translations , downsampling?
We may have minimal information about the geometry, and only locally. Let's think in terms of functions on the set X.

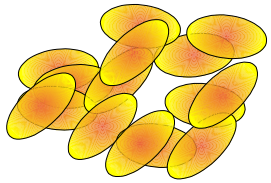
Dilations:
Use the diffusion operator T and its dyadic powers as dilations.

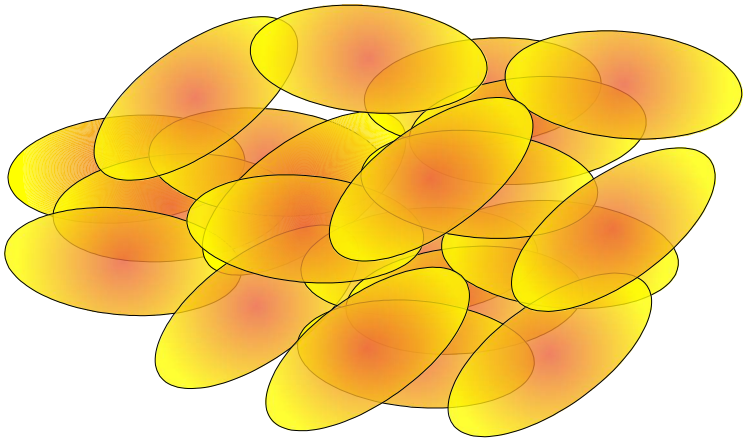
Translations and downsampling:
Idea: diffusing a basis of "scaling functions" at a certain scale by a power of T should yield a redundant set of coarser "scaling functions" at the next coarser scale: reduce this set to a Riesz (i.e. well-conditioned)-basis. This is downsampling in the function space,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

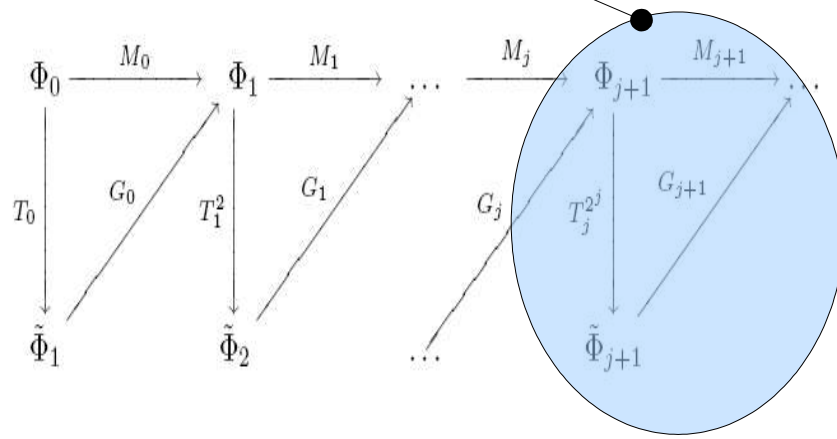
Multiscale Random Walkers





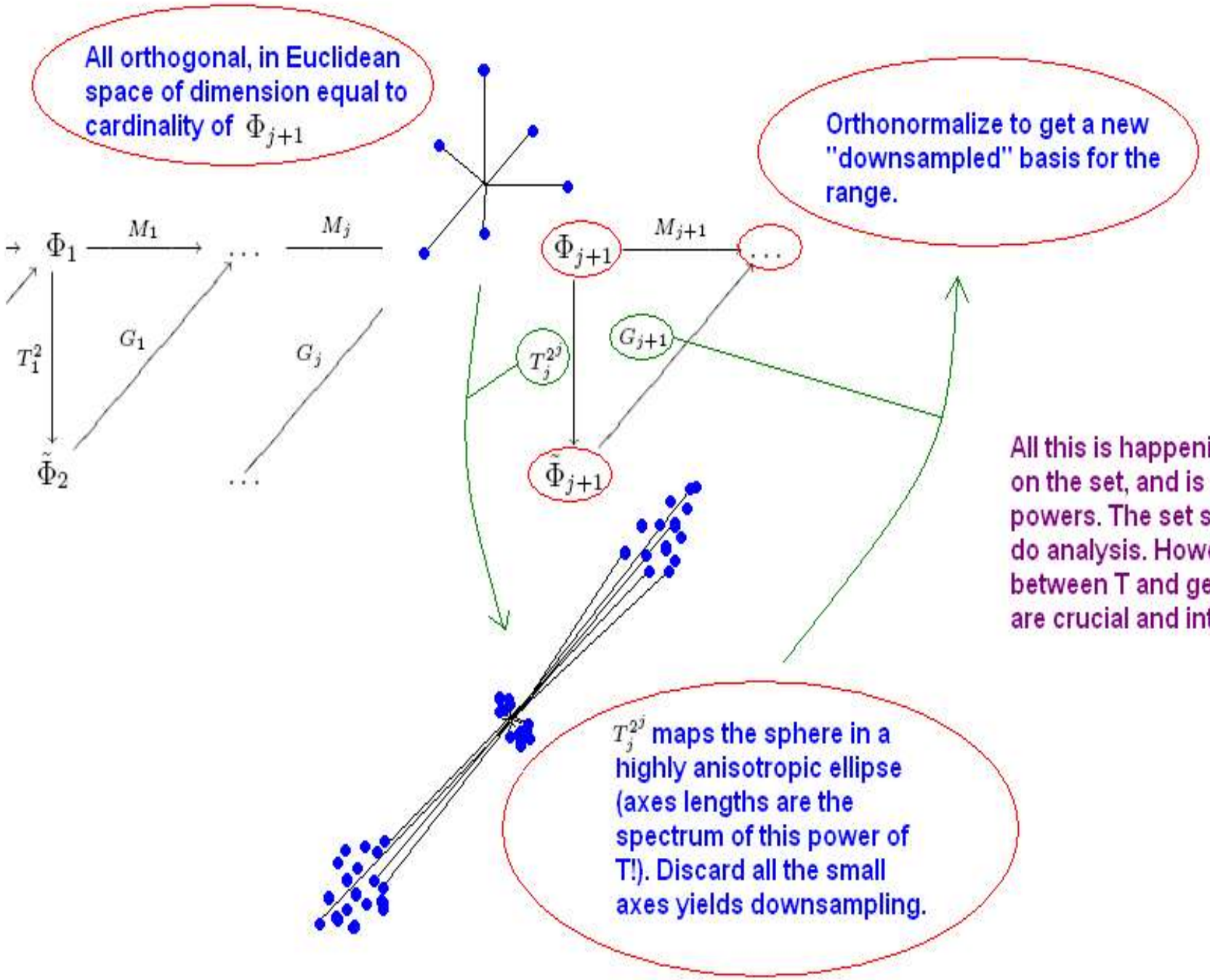






$$\varphi_{j,l,k}(x) = M_{j-1}\varphi_{j-1,l,k}(x) \quad , x \in X_{j-1}$$

$$= M_{j-1}M_{j-2} \cdot \ldots \cdot M_0\, \varphi_{0,l,k}(x) = \left(\prod_{l=0}^{j-1} M_l\right) \varphi_{0,l,k}(x) \quad , x \in X_0$$


All orthogonal, in Euclidean space of dimension equal to cardinality of Φ_{j+1}
Orthonormalize to get a new "downsampled" basis for the range.
Φ_1
M_1
M_j
M_{j+1}
Φ_{j+1}
T_1^2
G_1
G_j
G_{j+1}
$T_j^{2^j}$
$\tilde{\Phi}_2$
$\tilde{\Phi}_{j+1}$
All this is happening in the function space on the set, and is driven by T and its (dyadic) powers. The set seem almost an "excuse" to do analysis. However the interaction between T and geometric properties of set are crucial and interesting.
$T_j^{2^j}$ maps the sphere in a highly anisotropic ellipse (axes lengths are the spectrum of this power of T!). Discard all the small axes yields downsampling.

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

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

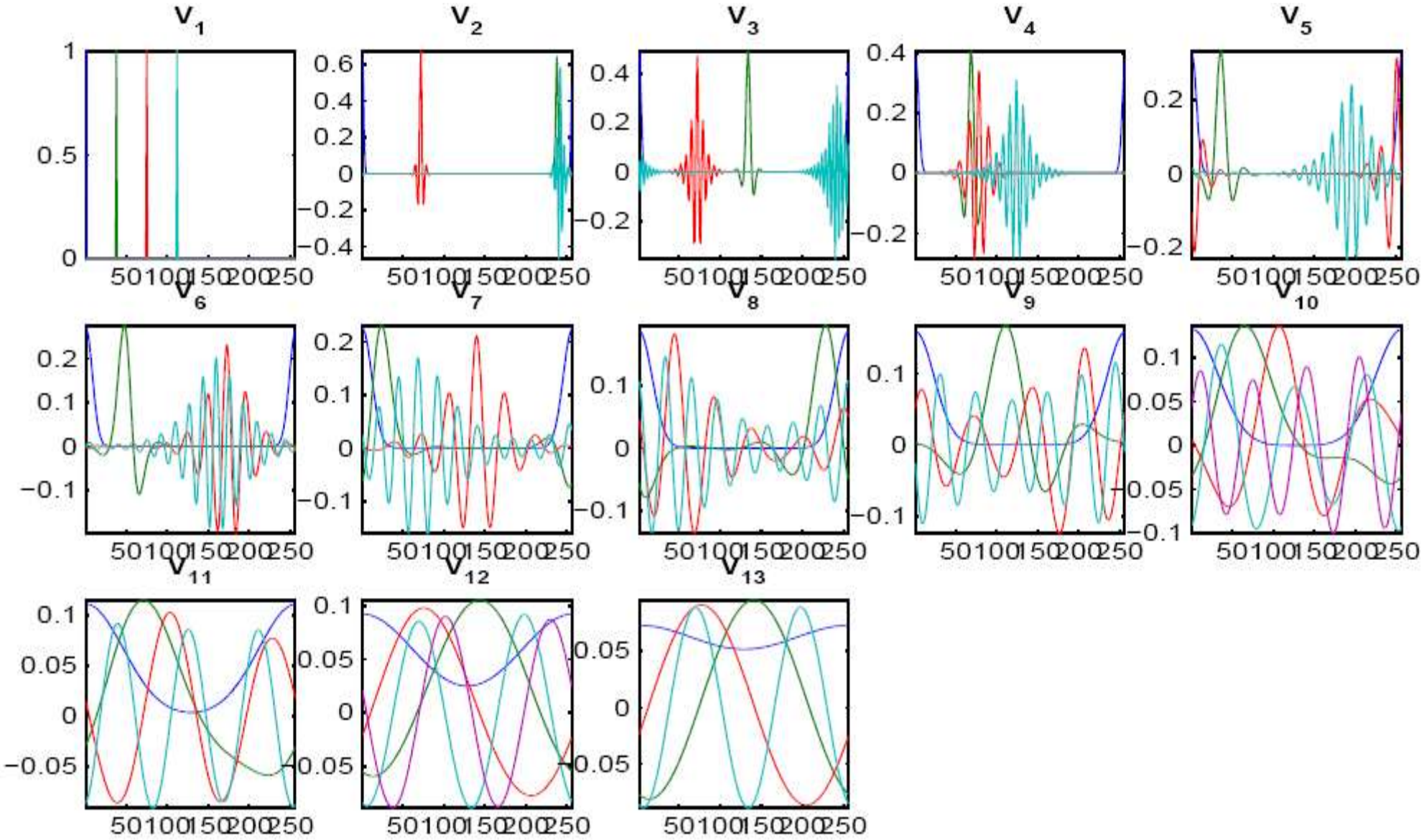


Fig. 3. Multiresolution Analysis on the circle. We consider 256 points on the unit circle, start with $\varphi_{0,k} = \delta_k$ and with the standard diffusion. We plot several scaling functions in each approximation space V_j.

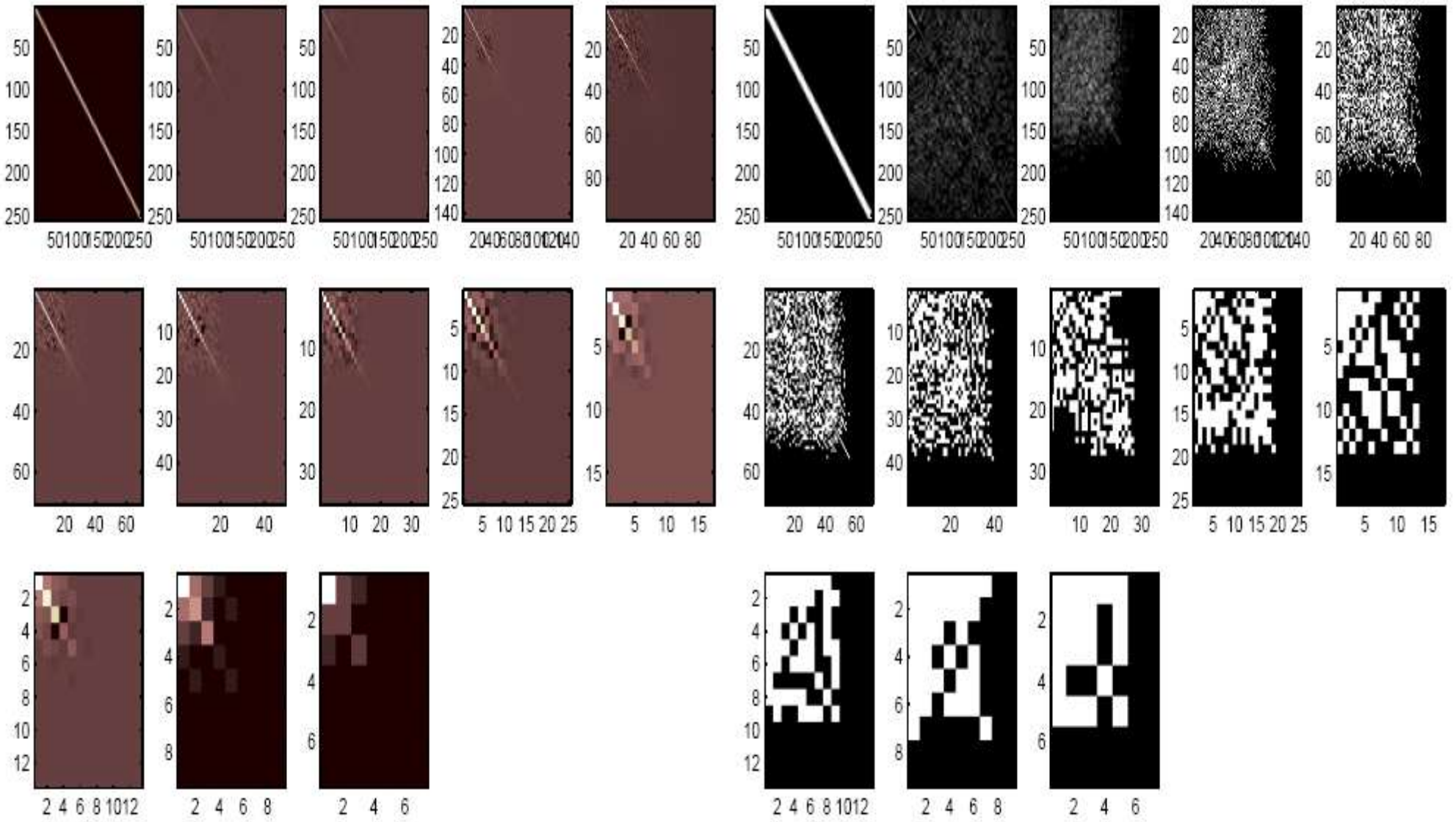


Fig. 4. Multiresolution Analysis on the circle: on the left we plot the compressed matrices representing powers of the diffusion operator, on the right we plot the entries of the same matrices which are above working precision.

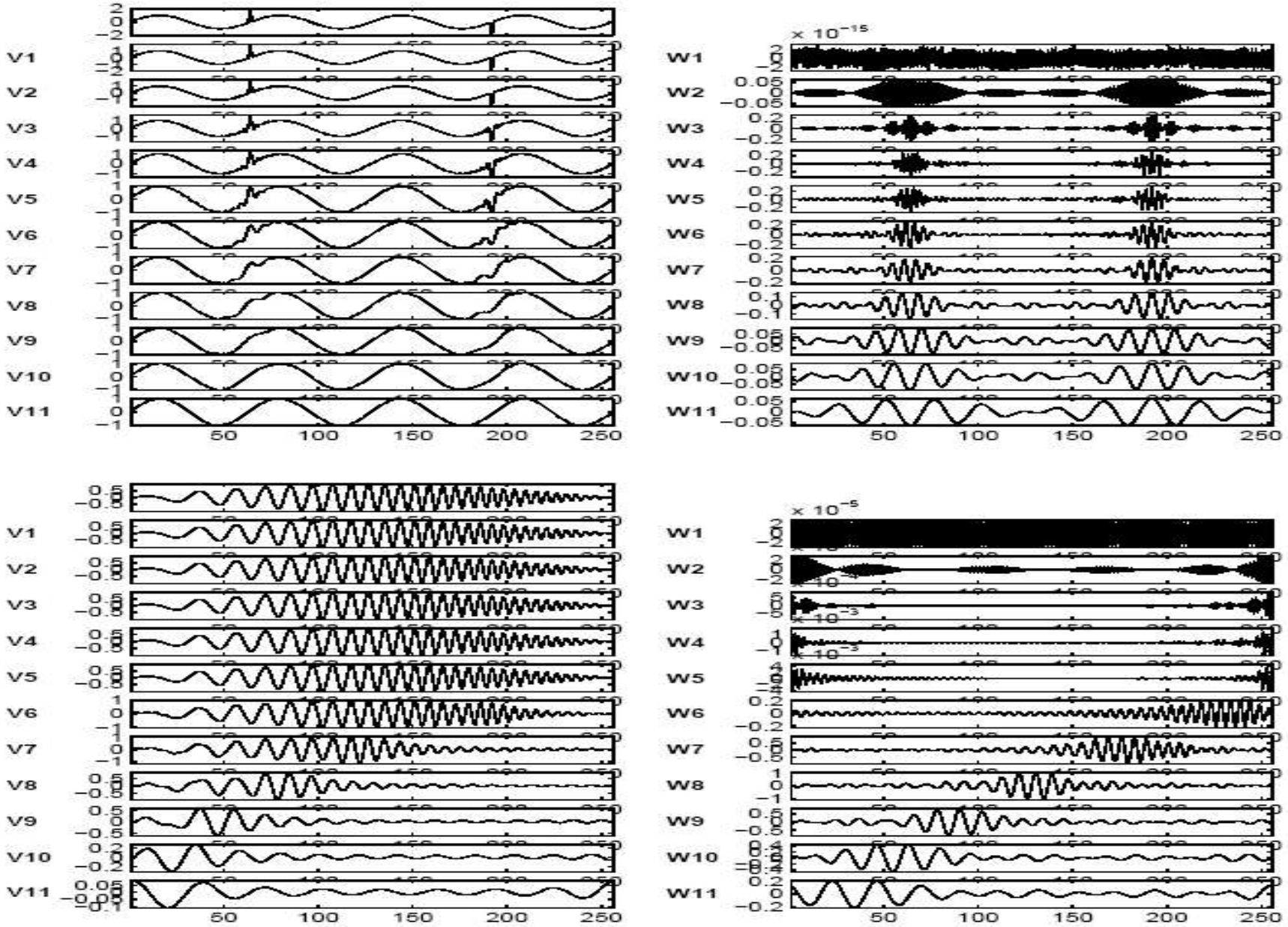
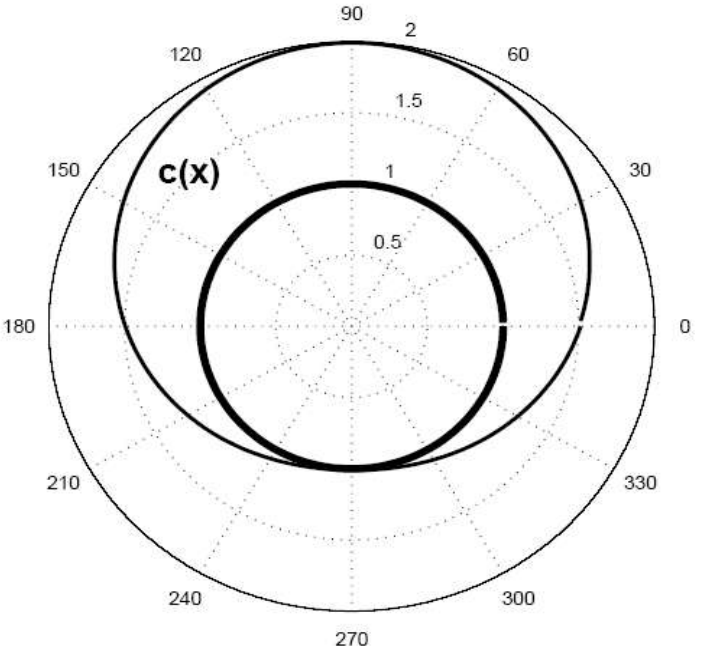
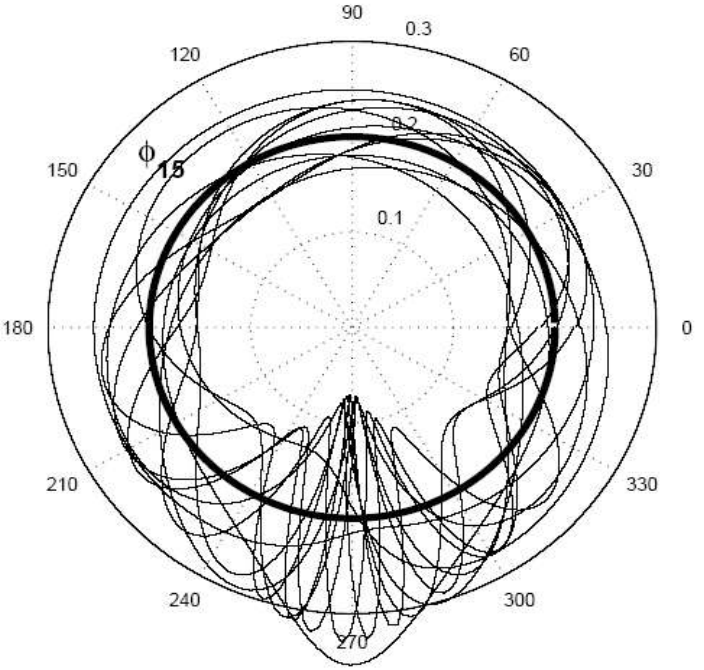
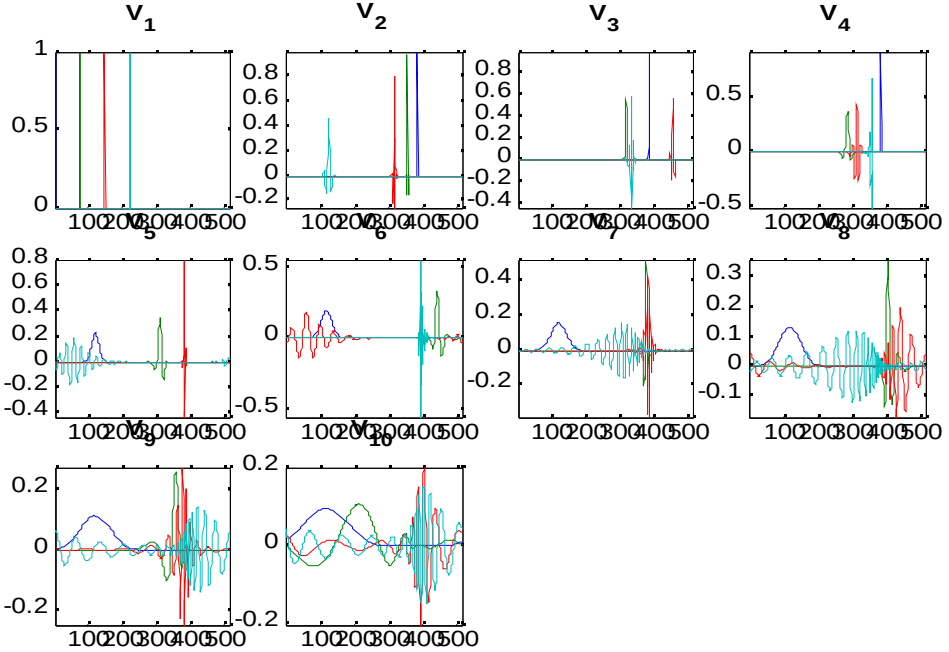


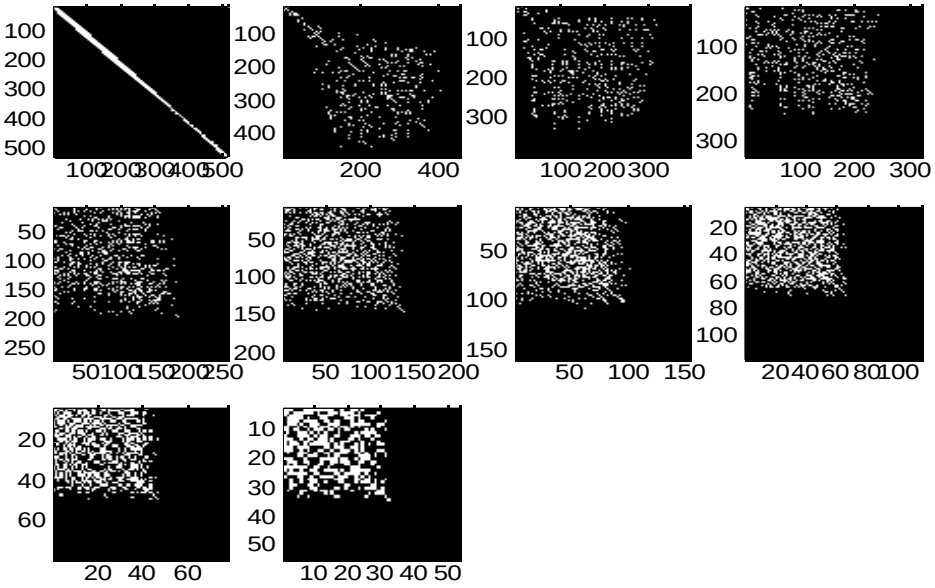
FIGURE 4. Multiresolution Analysis on the circle. In the same setting as for Figure 3, we compute the multiscale transform of a periodic signal on the circle, contaiminated by two δ-impulses (top) and of windowed chirp (bottom). In the first column we plot the projections onto coarses and coarses scaling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

Diffusion on nonhomogenous circle









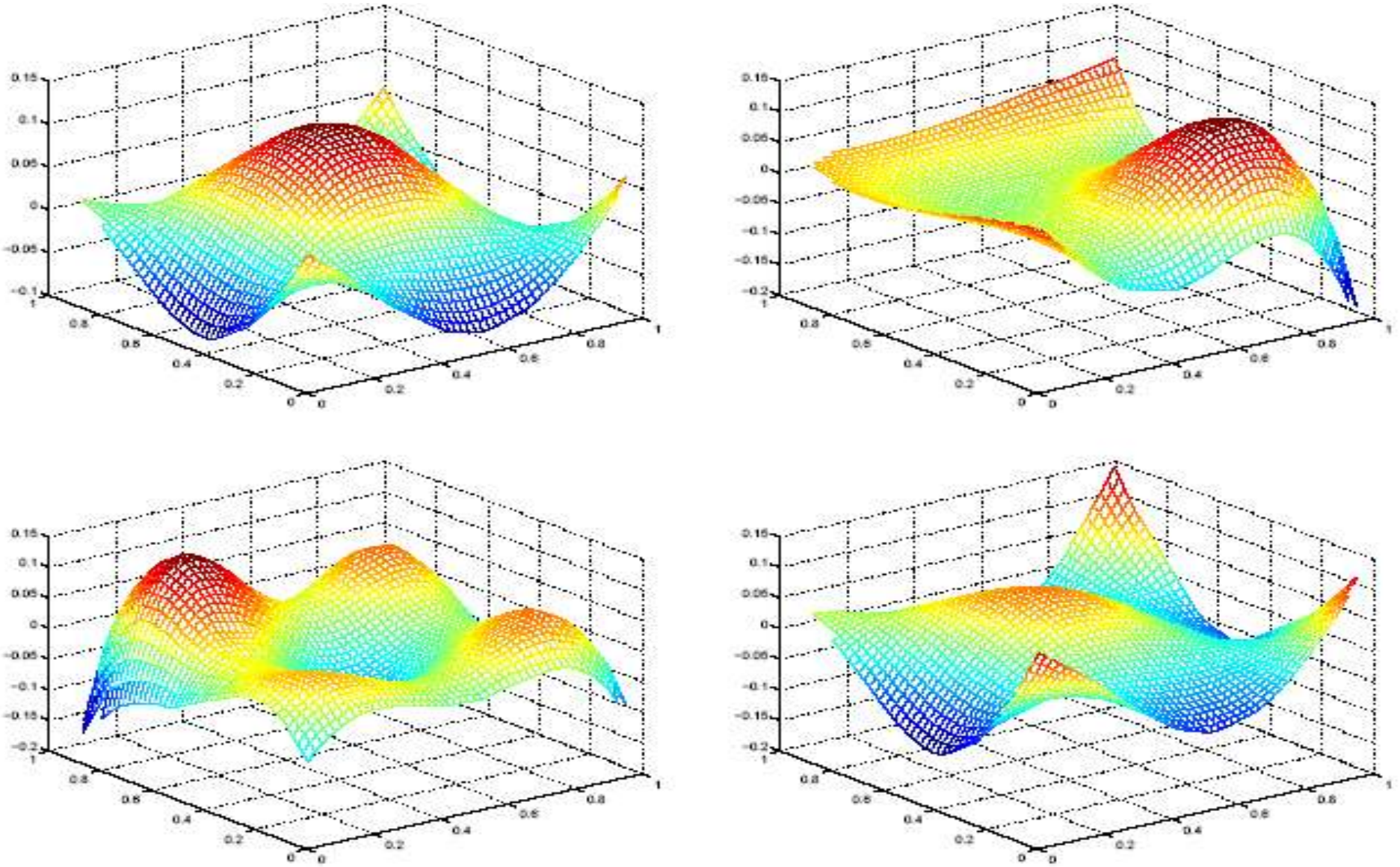
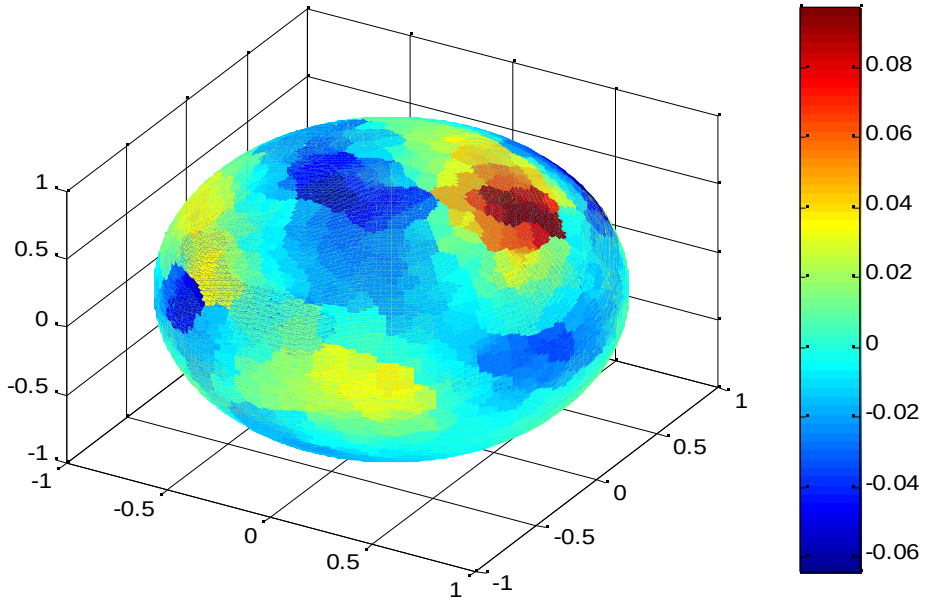
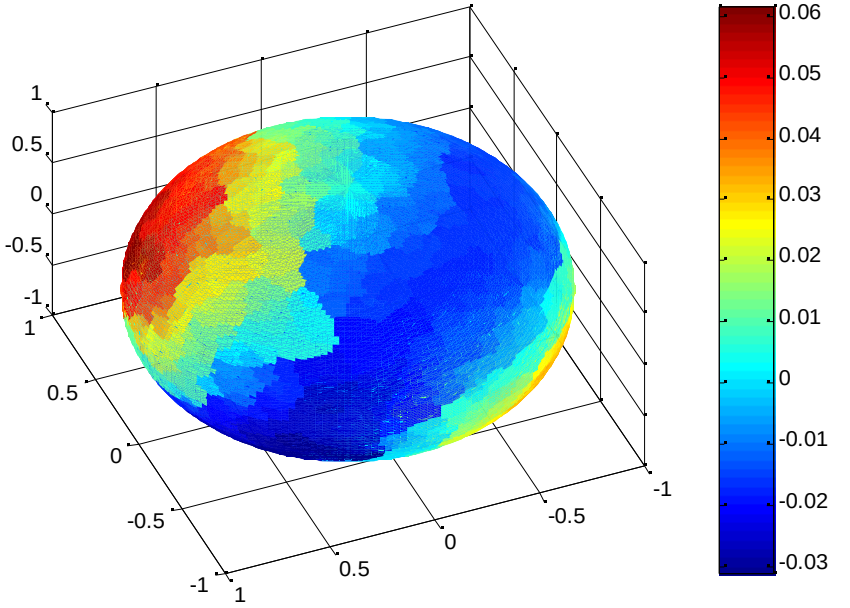
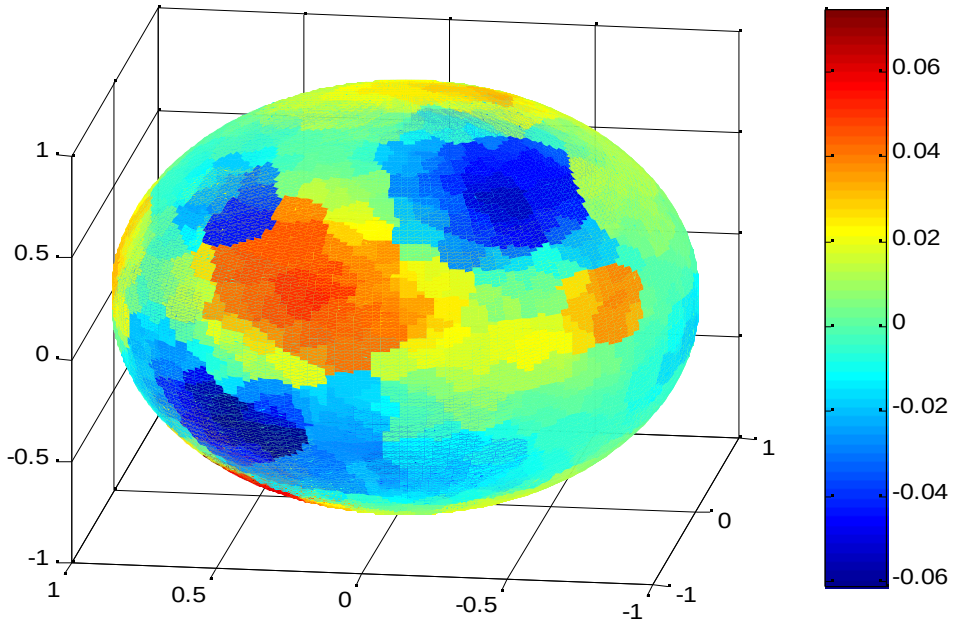


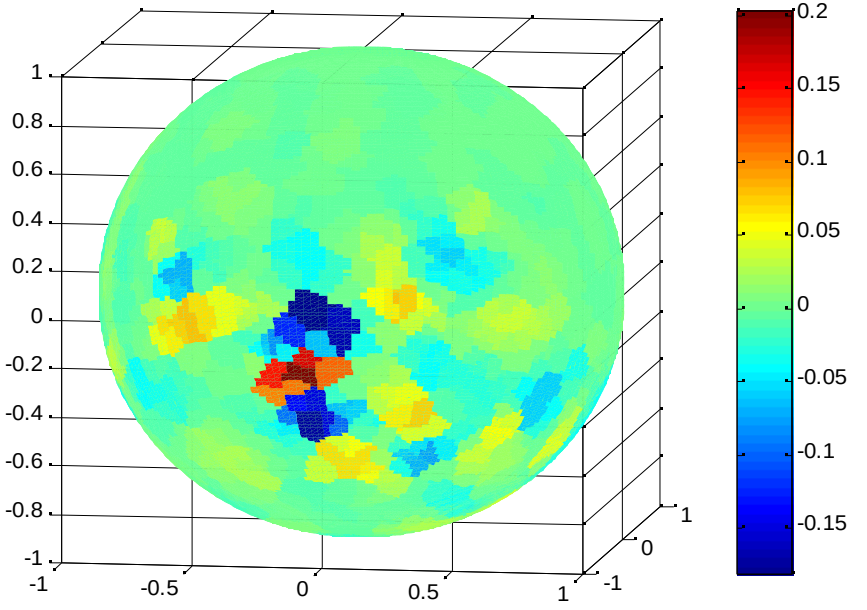
FIGURE 6. Example of scaling functions at coarse level associated with a Beltrami diffusion on randomly distributed points on the unit square. For graphical reasons, we are plotting a smooth extension of these scaling functions on a uniform grid by cubic interpolation.

Diffusion Wavelets on the sphere









Diffusion Wavelets on a dumbbell

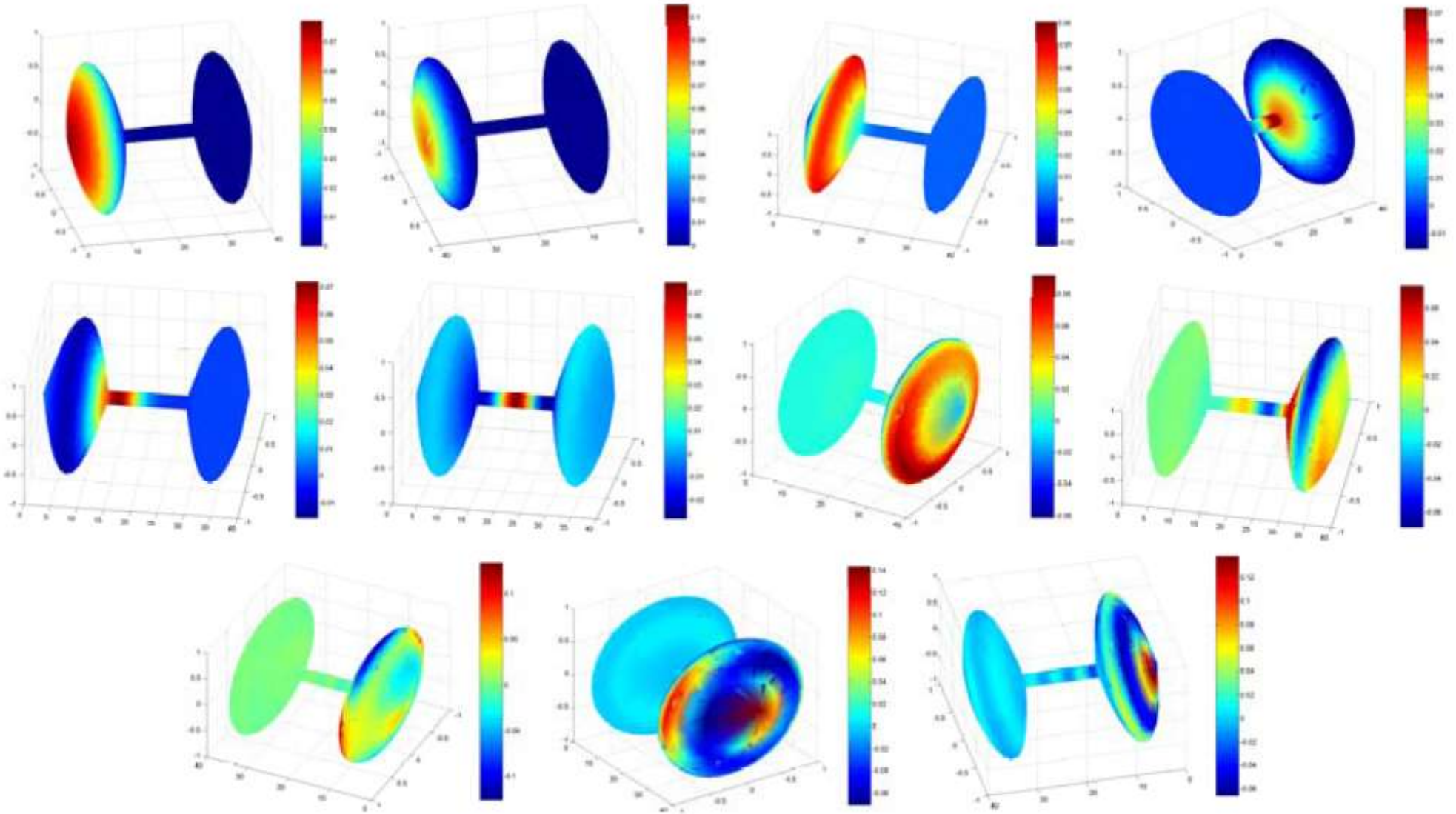


Fig. 8. Some diffusion scaling functions and wavelets at different scales on a dumbbell-shaped manifold sampled at 1400 points.

Connections...

Wavelets:
- Lifting (Sweldens, Daubechies, ...)
- Continuous wavelets from group actions

Classical Harmonic Analysis:
- Littlewood-Paley on semigroups (Stein), Markov diffusion semigroups
- Martingales associated to the above, Brownian motion
- Generalized Heisenberg principles (Nahmod)
- Harmonic Analysis of eigenfunctions of the Laplacian on domains/manifolds
- Atomic decompositions

Numerics:
- Algebraic Multigrid (Brandt, ...)
- Kind of "inverse" FMM (Rohklin, Belkyin-Coifman-Rohklin, ...)
- Multiscale matrix compression techniques (Gu-Eisenstat, Gimbutas-Martinsson-Rohklin,...)
- FFTs!
- Randomizable

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.

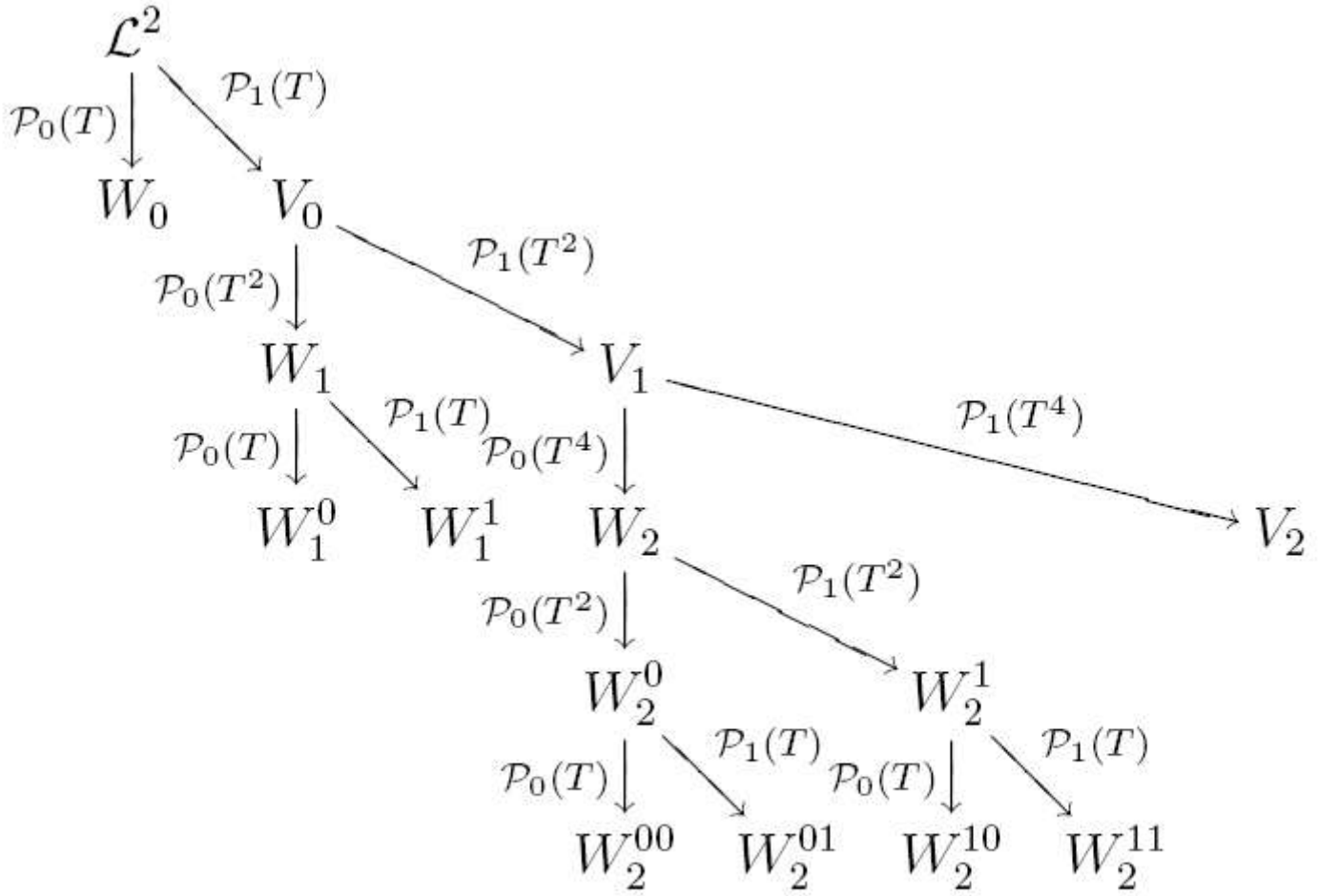


Fig. 2. Diagram for wavelet packet construction

Wavelet packets, best basis & compression

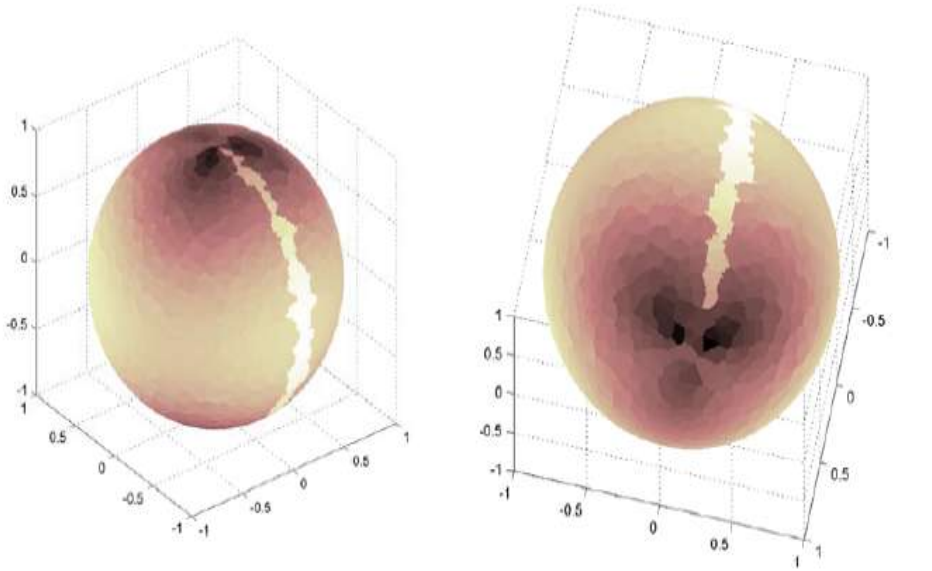


Fig. 9. Two different views of the function F on the sphere.

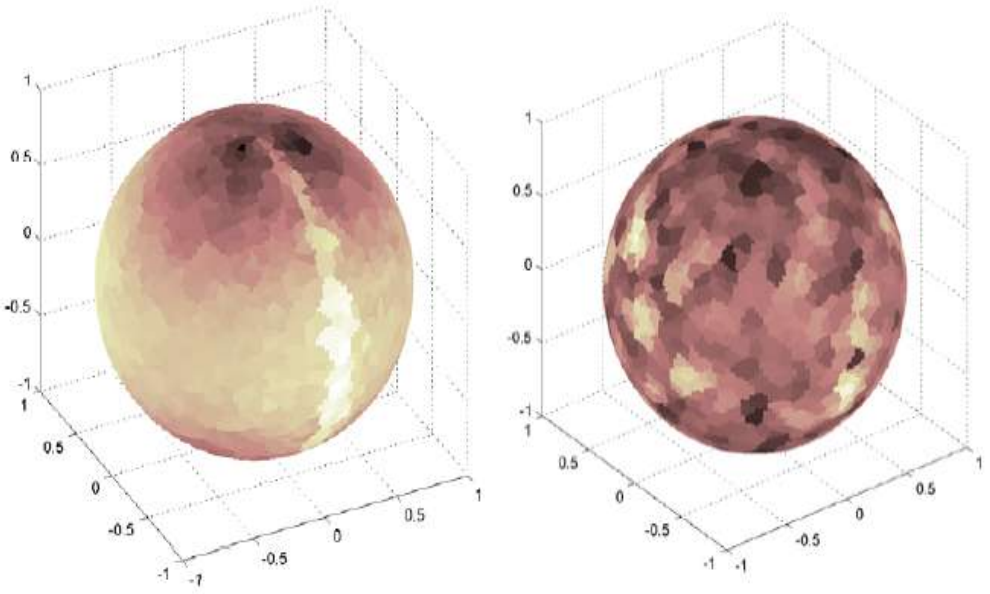


Fig. 11. Left: reconstruction of the function F with top 50 best basis packets. Right: reconstruction with top 200 eigenfunctions of the Beltrami Laplacian operator.

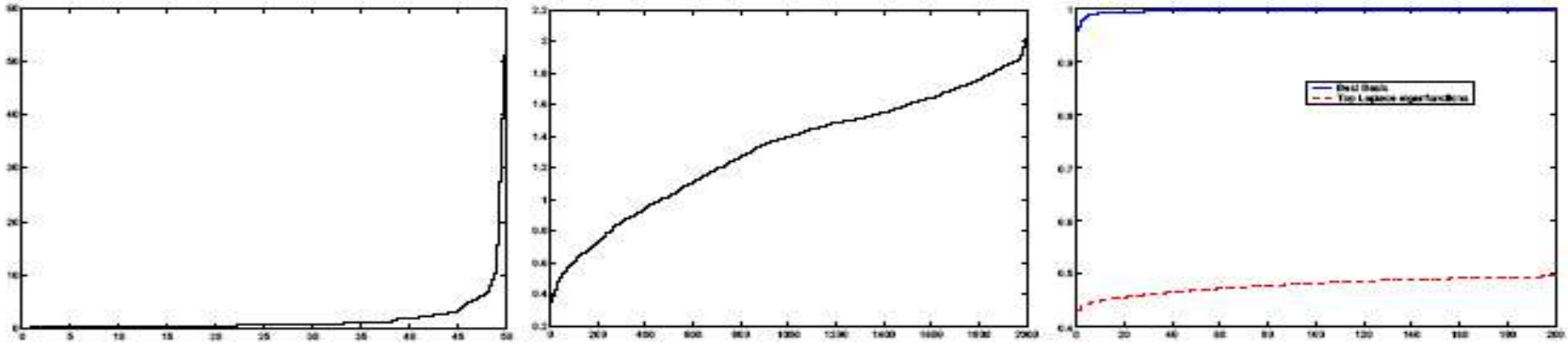


Fig. 10. Left to right: 50 top coefficients of F in its best diffusion wavelet basis, distribution coefficients$F$ in the delta basis, first 200 coefficients of F in the best basis and in the basis of eigenfunctions.

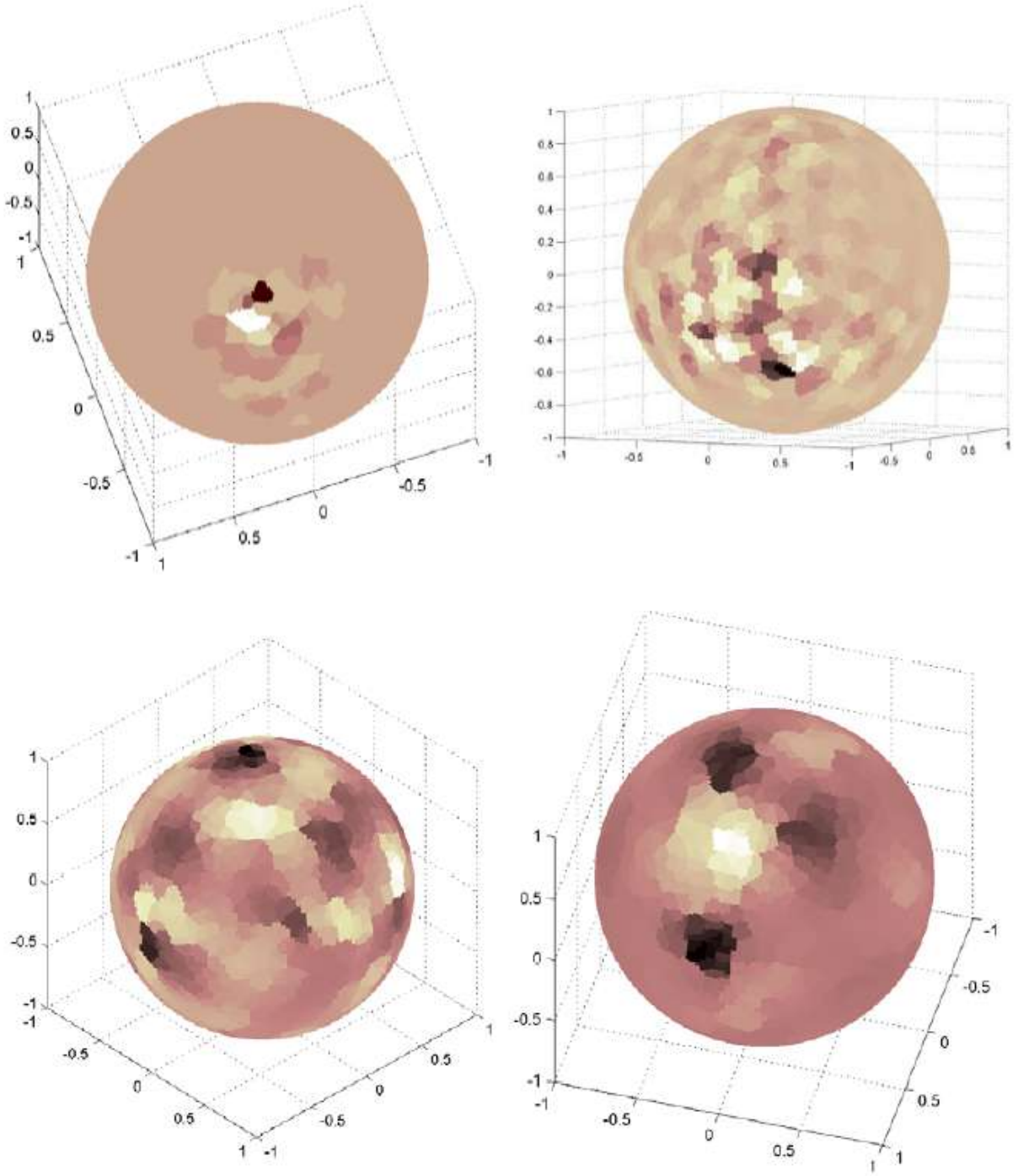


Fig. 6. Some diffusion wavelets and wavelet packets on the sphere, sampled randomly uniformly at 2000 points.

Compression example II

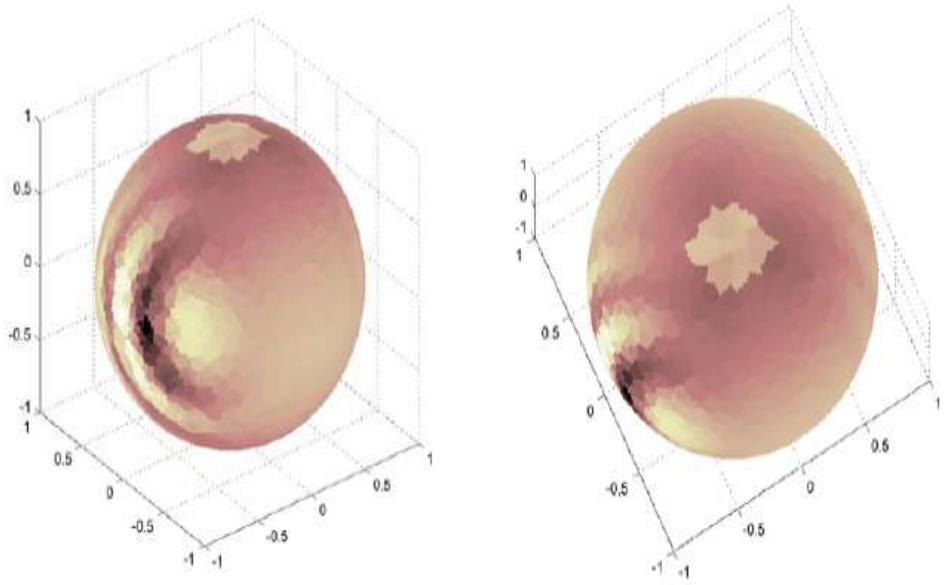


Fig. 12. Two different views of the function F on the sphere.

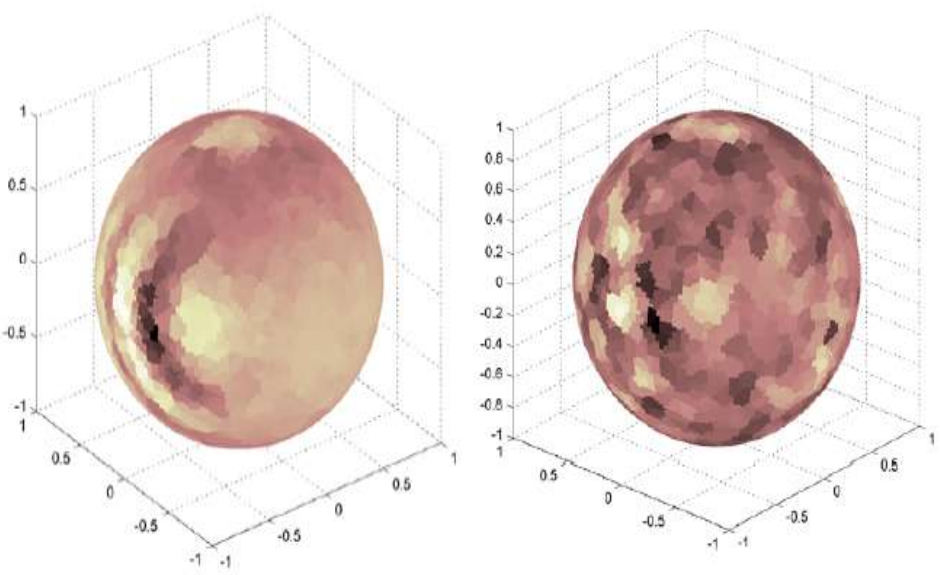


Fig. 14. Left: reconstruction of the function F from 200 best basis diffusion wavelet packet coefficients. Right: reconstruction from top 200 eigenfunctions.

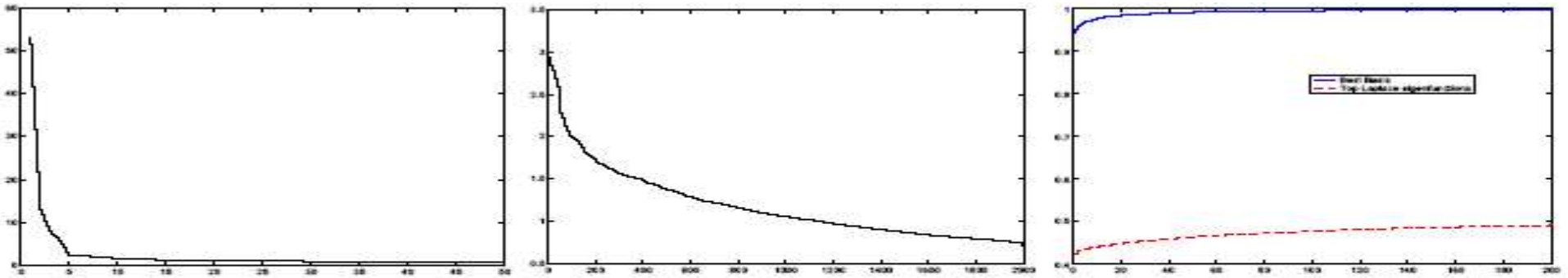


Fig. 13. Left to right: 50 top coefficients of F in its best diffusion wavelet basis, distribution coefficients$F$ in the delta basis, first 200 coefficients of F in the best basis and in the basis of eigenfunctions.

Denoising

[a la Coifman-Wickerhauser & Donoho-Johnstone]

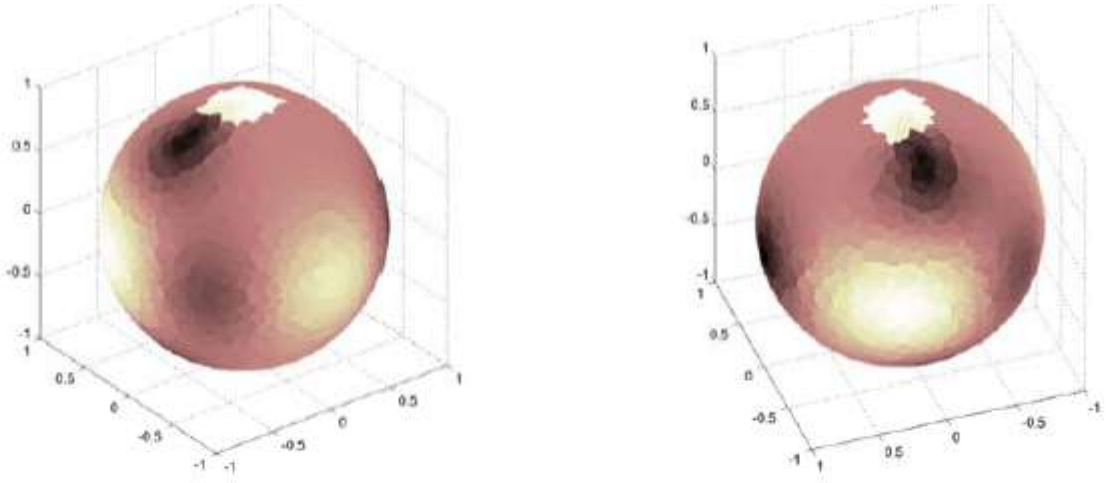


Fig. 15. Two views of G.

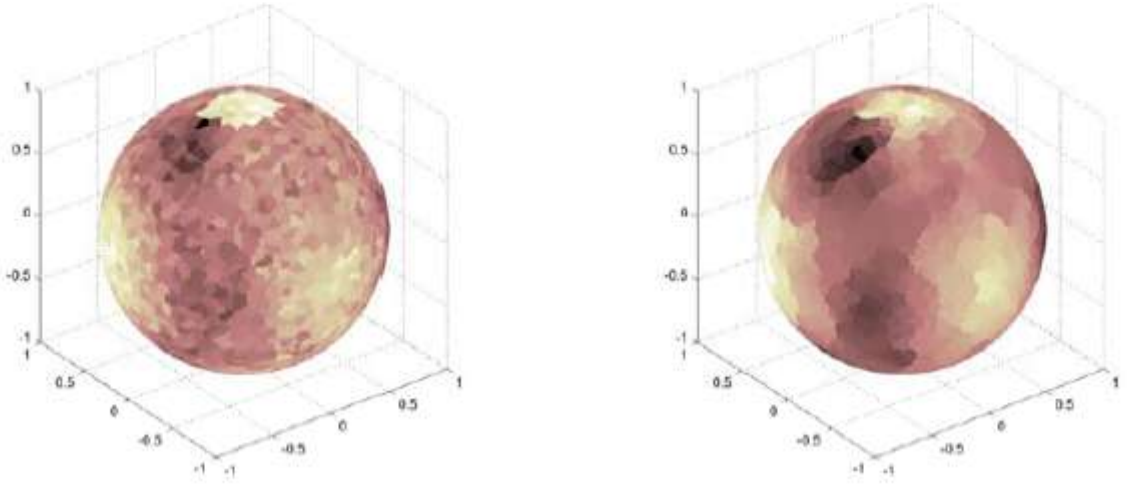


Fig. 16. Left: G with noise; right: G denoised

Compression on nonuniformly anisotropic function spaces

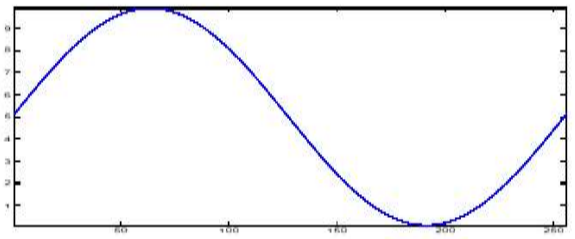


Fig. 3. Impedance of the anisotropic diffusion operator T: large on one part of the circle and almost 0 on another.

Consider circle with non-uniform impedance as before. The measure of smoothness, if defined according to these wavelets, is non-uniform on the circle.

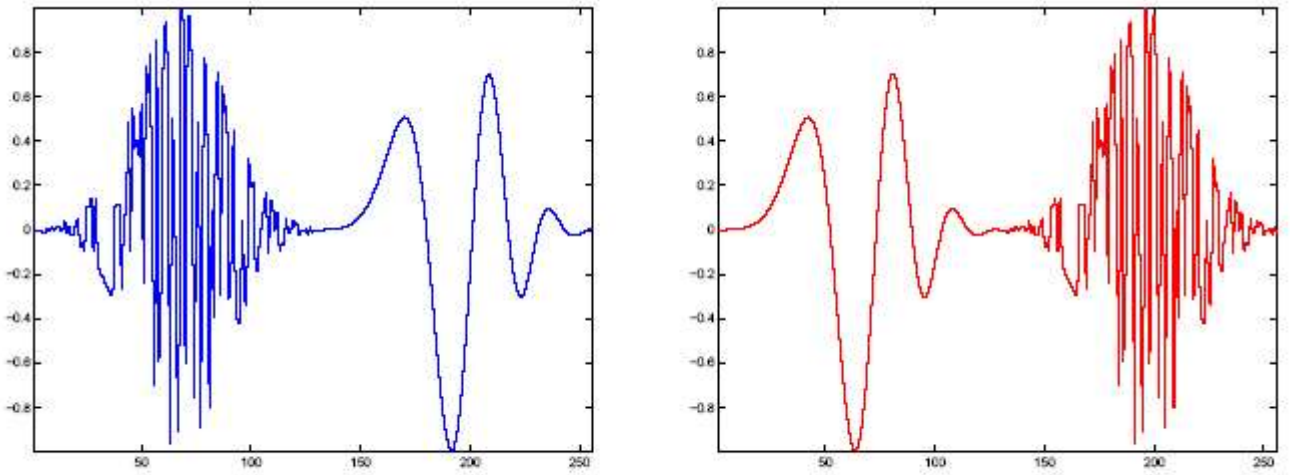


Fig. 7. The function F (on the left) and its reflection.

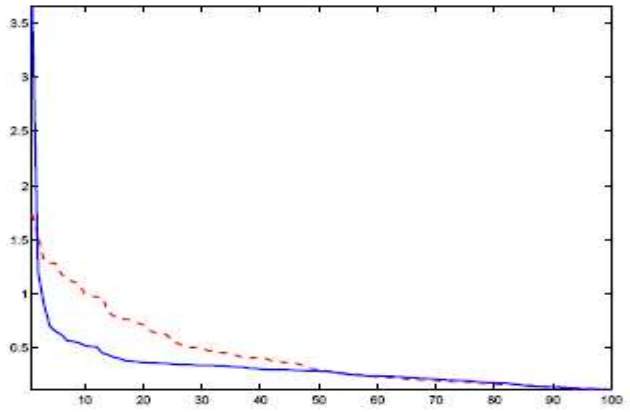


Fig. 8. Comparison of the magnitude first 50 coefficients of F and its reflection in their best wavelet packet bases.

Local Discriminant Bases

[Saito-Coifman]

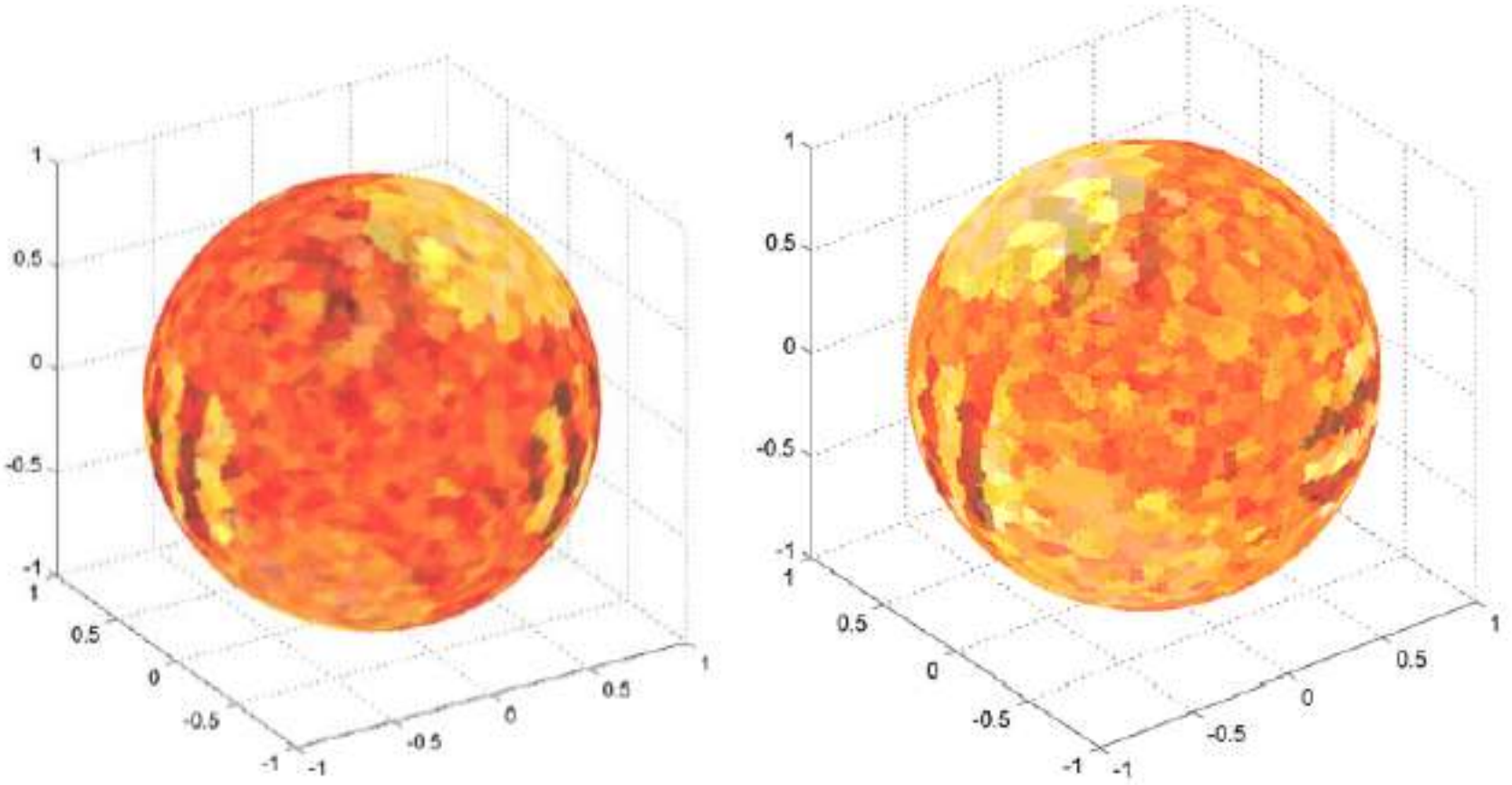
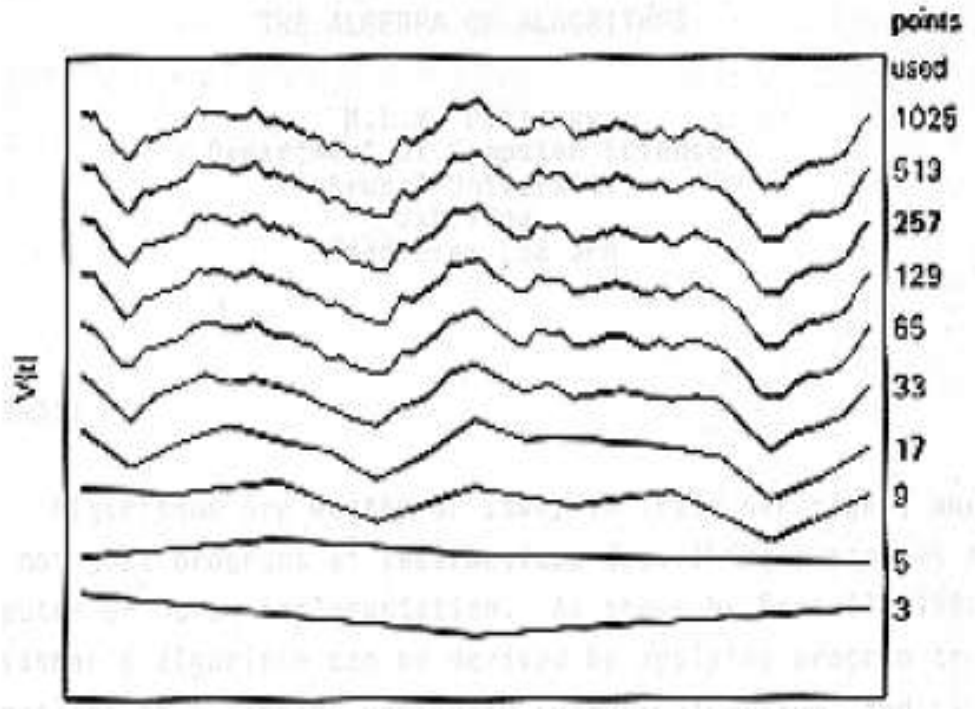


Fig. 19. Left to right, a realization of a function from class 1 and 2 respectively. Note that the third smooth texture patch is on the back side of the sphere, and can be viewed in semitransparency. The other two smooth patches are decoys in random non-overlapping positions.

Brownian Motion

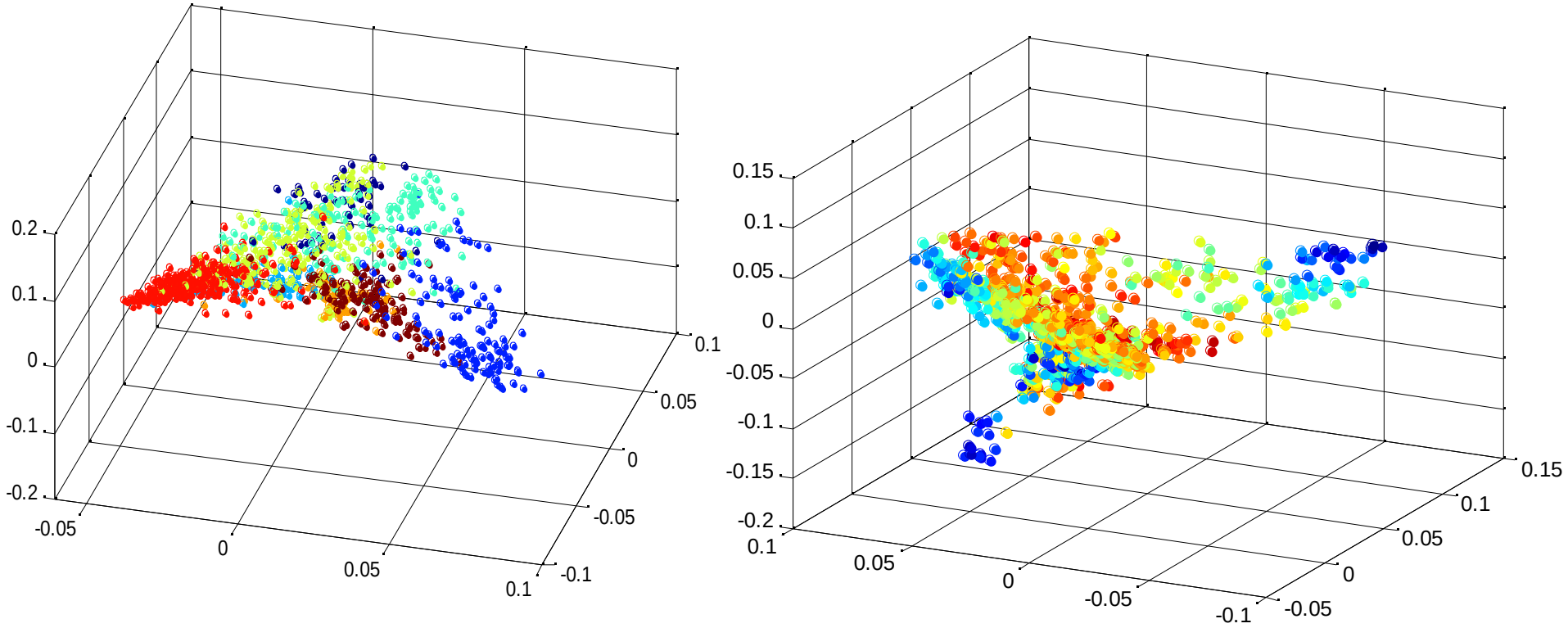
Diffusion wavelets allow for a natural generalization of the wavelet expansion of Brownian motion, for the efficient computation of long Brownian paths, evaluated at few points.

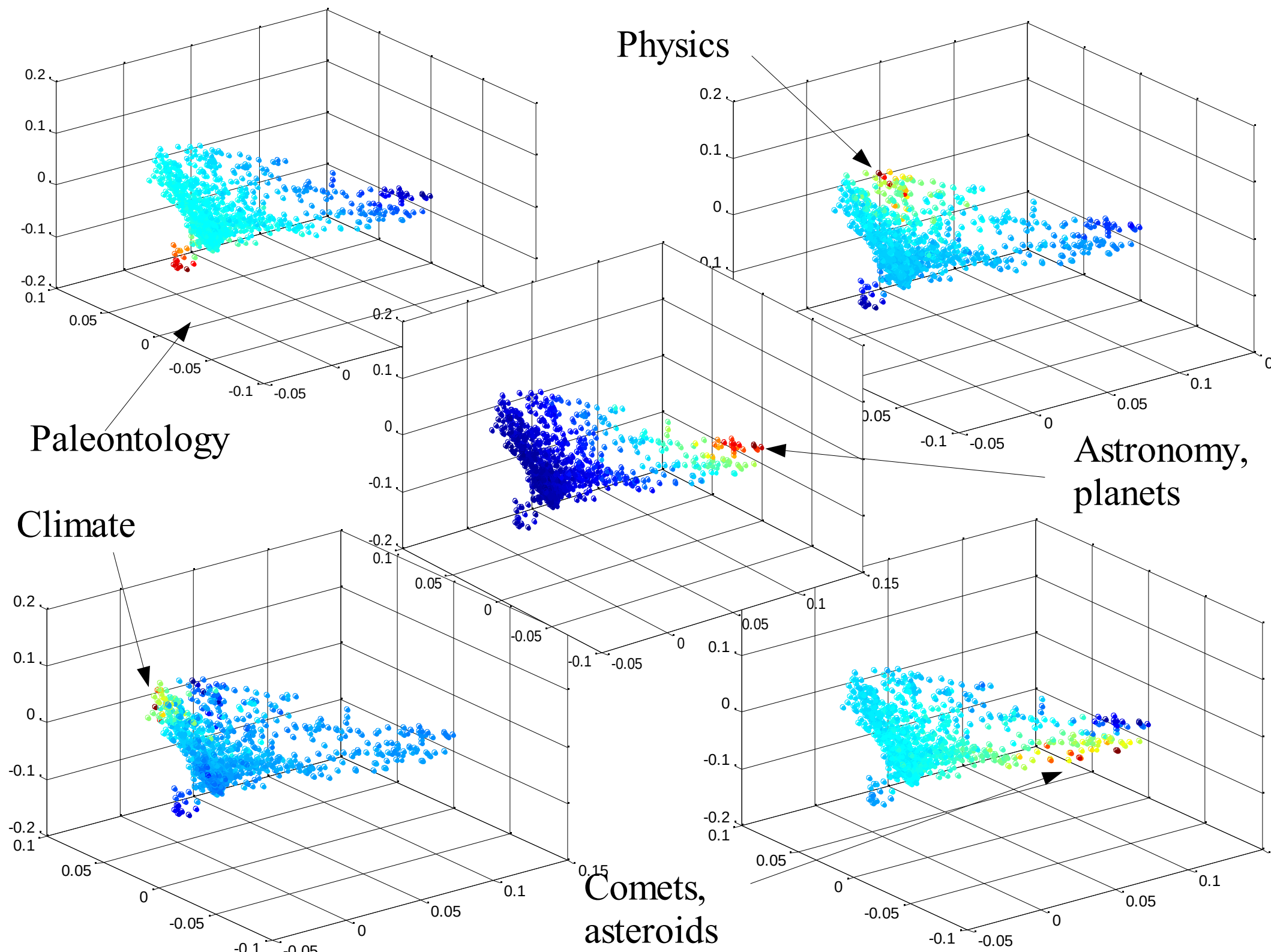


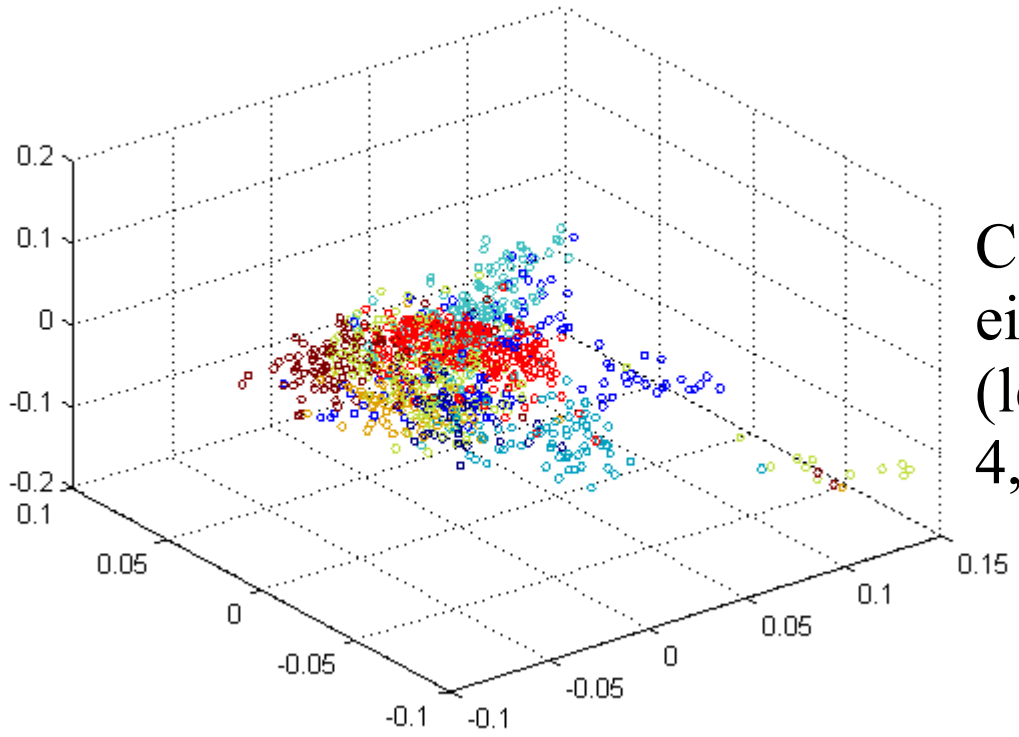
[PW Jones, MM, M Mohlenkamp, R Schul]

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.

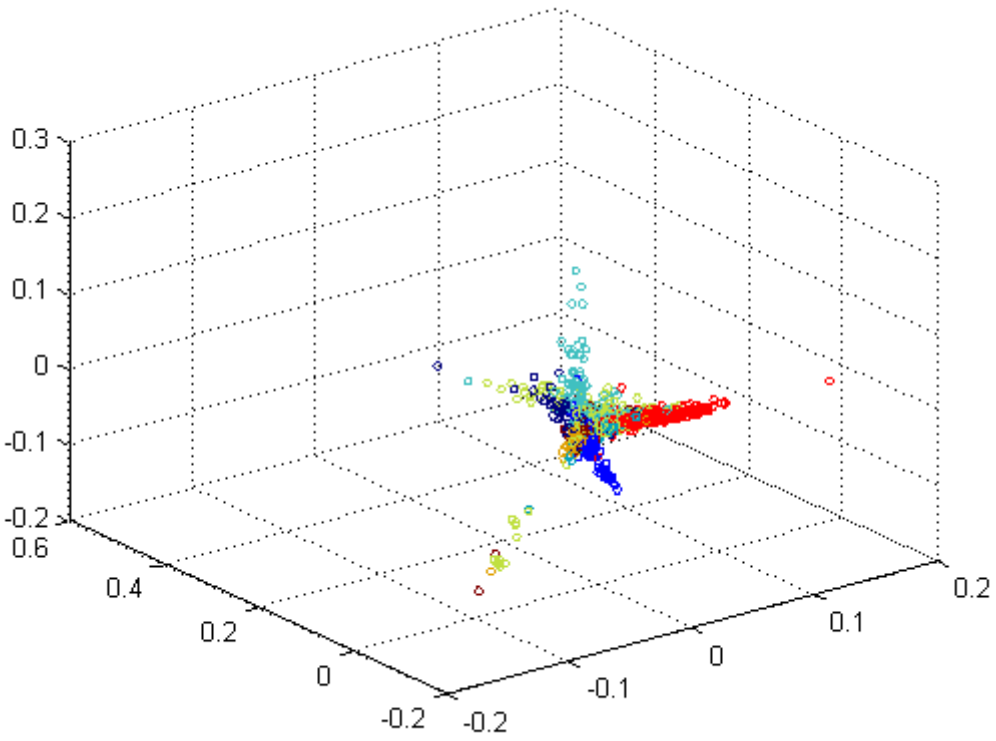



Physics
Paleontology
Astronomy, planets
Climate
Comets, asteroids



Classes get quickly garbled up in the eigenfunction coordinates (left:eigenmap with eigenfunctions 4,5,6)...

...but can stay separated with diffusion scaling functions (right: the scaling function embedding at scale 3).



Comments, Applications, etc...

- This is a wavelet analysis on manifolds (and more, e.g. fractals), graphs, markov chains, while Laplacian eigenfunctions do Fourier Analysis on manifolds (and fractals, etc...).
- We are "compressing" powers of the operator, functions of the operators, subspaces of the function subspaces on which its powers act (Heisenberg principle...), and the space itself (sampling theorems, quadrature formulas...)
- We are constructing a biorthogonal version of the transform (better adapted to studying Markov chains) and wavelet packets: this will allow efficient denoising, compression, discrimination on all the spaces mentioned above.
- Does not require the diffusion to be self-adjoint, nor eigenvectors.
- The multiscale spaces are a natural scale of complexity spaces for learning empirical functions on the data set.
- Diffusion wavelets extend outside the set, in a natural multiscale fashion.
- To be tied with measure-geometric considerations used to embed metric spaces in Euclidean spaces with small distortion.
- Study and compression of dynamical systems.

Current & Future Work

- Multiscale embeddings for graphs, measure-geometric implications
- Martingale aspects and Brownian motion
- 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!