Multiscale Analysis of Data Sets with Diffusion Wavelets *†

Mauro Maggioni ‡ Ronald R Coifman §

Abstract

Analysis of functions of manifolds and graphs is essential in many tasks, such as learning, classification, clustering. The construction of efficient decompositions of functions has till now been quite problematic, and restricted to few choices, such as the eigenfunctions of the Laplacian on a manifold or graph, which has found interesting applications. In this paper we propose a novel paradigm for analysis on manifolds and graphs, based on the recently constructed diffusion wavelets. They allow a coherent and effective multiscale analysis of the space and of functions on the space, and are a promising new tool in classification and learning tasks. In this paper we overview the main motivations behind their introduction, their properties, and sketch an application to multiscale document corpora analysis.

1 Introduction

Efficient approximation and analysis of functions on manifolds and graphs, modeling large data sets (or supports of probability distributions), is essential to many algorithms in learning, classification, clustering and data mining. The problem of learning can be often cast as an approximation and regularization problem on some space. Much recent research focuses on the case when the space is a smooth Riemannian manifold or a graph. Exploiting the geometry of the underlying space has led to interesting results and new techniques for function approximation and learning on such spaces [2, 1, 12, 3].

One the recently used tools in such analysis are the eigenfunctions of a Laplacian operator naturally defined on the graph/manifold [1, 9], which form an orthonormal basis of smooth, global functions on the graph/manifold. Analysis by decomposing a function onto these eigenfunctions is a classical generalization of Fourier analysis with recent successful applications to learning [3, 12]. However, it is well-known even in one dimension that Fourier analysis is a powerful tool for global analysis of functions, but it is not efficient for studying local or transient properties (such as regularity) of functions. This motivated the construction, about 20 years ago, of classical wavelets in Euclidean spaces, and wavelets became mainstream across many different fields, such as mathematics, numerical analysis and scientific computation, physics, engineering, signal and image processing, just to mention a few. One the main differences between wavelets and Fourier modes, is that wavelets are localized in both location and frequency, and are parametrized by location and scale. They allow a very efficient multiscale analysis, like a powerful tunable microscope probing the properties of a function at different locations and scales.

In the paper [8] the authors proposed a generalization the construction of classical wavelets to graphs and manifolds, that were there named diffusion wavelets, because they are associated with a diffusion process that defines the different scales. They allow a multiscale analysis of functions on manifolds and graphs, and hence on data sets that can be modeled as such. Multiscale analysis in general has been an extremely powerful in many different fields, and this is an attempt to coherently transfer some its concepts and tools to the setting of manifolds and graphs. The properties of diffusion wavelets make them good candidates for accomplishing tasks related to function approximation, compression, denoising, and hence learning tasks are natural applications.

We first overview a "diffusion analysis" framework that encompasses both Laplacian eigenfunctions and multiscale diffusion wavelets, and then sketch a few applications of this novel multiscale analysis techniques.

2 Analysis of Diffusion on Graphs and Manifolds

One way of starting the analysis of a data set, modeled as a graph or a manifold, is to consider a natural random walk P on it. The random walk allows to diffuse on the data set, exploring it, discovering clusters and regions separated by bottlenecks. This process can be analyzed at different time scales, and hence we will be interested in the family of operators $\{P^t\}$. For an initial

*This paper appeared in the Data Mining for Biomedical Informatics Workshop which was held in conjunction with the 7th SIAM International Conference on Data Mining on April 28, 2007 in Minneapolis, MN.

†MM is partially supported by NSF DMS-0650413.

‡Department of Mathematics and Computer Science, Duke University, Durham, NC, 27708

§Department of Mathematics, Yale University, New Haven, CT, 06511

condition δ_x, the Markov property $P^t\delta_x(y)$ represents the probability of being at y at time t, having started from x. We can also interpret the matrix P as encoding local similarities between points, and the matrix P^t is diffusing, or integrating, this local information t steps to larger and larger neighborhoods of each point [9].

For very large times, the random walk can be analyzed through its top eigenvectors, which are related to those of a Laplacian on the graph/manifold, which we discuss in the next session. The analysis for large times leads to *Fourier analysis* of the very large-scale structures of the graph/manifold, and to the identification of useful structures, such as large scale clusters.

For small and medium times, the random walk in general cannot be studied effectively with eigenfunctions, which are global and not suited for analyzing local behavior. On the other hand many interesting features of the data and of functions on the data can be expected to exist at small and medium time scales: a remarkable example are complex (computer, biological, information, social) networks. The task of analyzing P^t for all times and locations seems tantalizing. We show that it is not so, for large interesting classes of random walks P: the redundancy in time and space of the information in the family $P^t(x, y)$ can be compressed out, and an efficient multiscale encoding is possible. This leads to (diffusion) *wavelet analysis* [8].

3 Multiscale Analysis of Document Corpora

We consider the following cloud of digital data. We are given 1047 articles from Science News, from which we collected 2036 words chosen as being relevant for this body of documents. A document-word matrix whose entry (i, j) is the frequency of the j^{th} word in the dictionary in the i^{th} document was constructed. Each document is categorized as belonging to one of the following fields: Anthropology, Astronomy, Social Sciences, Earth Sciences, Biology, Mathematics, Medicine or Physics. Let $\mathcal{C}$ denotes the set of these categories. This information can be assimilated to the function $\text{cat} : X \to \mathcal{C}$, defined by $\text{cat}(x) = \{\text{category of the point } x\}$. A similarity $W_{xy} = W_{yx}$ is computed between a each document pair by correlation. We let the degree matrix D to be the diagonal matrix defined by $D_{xx} = \sum_y W_{xy}$, and let the normalized Laplacian to be $\mathcal{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$.

In the kind of analysis that follows, nothing will be specific to the "documents" being text documents. The initial data matrix might have been a matrix from a gene micro-array experiment. The similarity between rows (or columns) would have been different, but given any such definition the analysis below can be carried through, leading to coarser representations of the relationships among genes (or patients), as well as allowing to regress indicator functions such as a positive or negative biopsy result.

Scaling Fcn	**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

Table 1: In this table we present some example of scaling functions on the documents, with some of the documents in their support, and some of the words most frequent in the documents.

We compute the eigenvalues $\{\lambda_i\}$ of the normalized Laplacian $\mathcal{L}$ and the corresponding eigenvectors $\{\xi_i\}$, and use them to embed the high-dimensional graph into Euclidean space as described [9]: see Figure 1. This embedding seems to be particularly meaningful since the different categories appear rather well-separated. A simple K-means or hierarchical clustering algorithm ran on $\Xi_n^{(t)}(X)$, yields clusters which match quite closely the given labels. This would correspond to a particular choice of kernel K-means (or hierarchical clustering), motivated by diffusion distance. We do not have space here to discuss the details here, the point of this discussion being that much more information can be extracted from the multiscale construction: the kernel is iterated over the set, induces a natural multiscale structure, that gets the data organized coherently, in

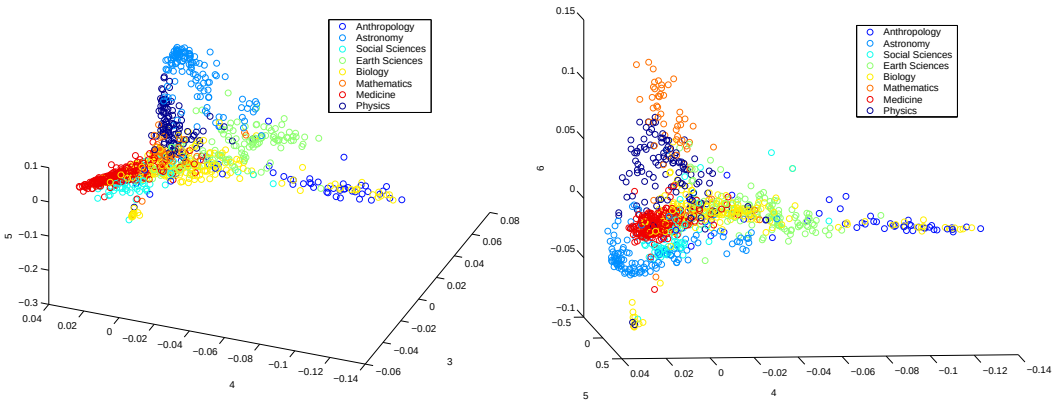

Figure 1: Embedding $\Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x))$: on the left coordinates $3, 4, 5$, and on the right coordinates $4, 5, 6$.

space and scale. The selection of the number of clusters is automatic at each scale: the number of scaling functions is determined by the constraint of their independence up to a certain precision.

We construct the diffusion scaling functions and wavelets on this cloud of points, see Figure 2. Scaling functions at different scales represent (from coarse to fine) categories at different levels of specialization: the (essential) support of scaling functions consist of a set of documents which are related by diffusion at a certain time (=scale), and have a common, well-distinguished topic. For example, $\phi_{3,4}$ 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 supports of the scaling functions grow with the scale, and so do the corresponding topics. For example $\phi_{4,3}$ now corresponds to a larger portion of Biology and Anthropology, compared to $\phi_{3,5}$, and includes articles on fossils in general, not just dinosaurs. $\phi_{4,2}$ corresponds to a larger portion of Astronomy, compared to $\phi_{3,10}$, and includes articles on comets, asteroids and space travel.

One can expand the function χ_i which is 1 on documents of class i and 0 elsewhere on diffusion wavelets. All classes have very compact representations, meaning that the diffusion wavelets do capture the structure of the classes. The exception is the function χ_i corresponding to Biology, which is not well-approximated by wavelets: we conjecture this is due to the complexity of the class and its many interconnections with other field, that make is function highly complex.

We also conducted an experiment in semi-supervised learning: we constructed the diffusion wavelets on the whole set of documents, and labeled $\tilde{X}$, a random subset of 10% of the points. We expanded $\chi_i|_{\tilde{X}}$ on the diffusion wavelets: $\chi_i|_{\tilde{X}}(x) = \sum \alpha^i_{j,k}\psi_{j,k}(x)$, $x \in \tilde{X}$, with a penalty cost for high frequency, small scale wavelets. We then evaluated the sum for $x \in X$ and labeled x by $\mathrm{argmax}_i \sum \alpha^i_{j,k}\psi_{j,k}(x)$. The misclassification rate was less than half of the best k-nearest neighbor classifier.

4 Conclusions and Future Work

Work is in progress on further applications of diffusion wavelets to the analysis of document corpora and information extraction, to semi-supervised learning, classification and clustering. Further applications for Markov decision processes are discussed in [11, 10]. A generalization of Support Vector Machine algorithms and Regularized Least Squares using diffusion wavelets is currently being studied, with the goal of introducing a coherent framework for automatic local scale selection for the kernel, a long-standing complex problem in the area.

References

[1] M Belkin and P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 6(15):1373–1396, June 2003.

[2] M Belkin and P Niyogi. Using manifold structure for partially labelled classification. *Advances in NIPS*, 15, 2003.

[3] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(Invited Special Issue on Clustering):209–239, 2004. TR-2001-30, Univ. Chicago, CS Dept., 2001.

[4] M Brand. Continuous nonlinear dimensionality reduction by kernel eigenmaps. *Proc. IJCAI*, 2003.

[5] J.C. Bremer, R.R Coifman, M. Maggioni, and A.D. Szlam. Diffusion wavelet packets. *Tech. Rep.*

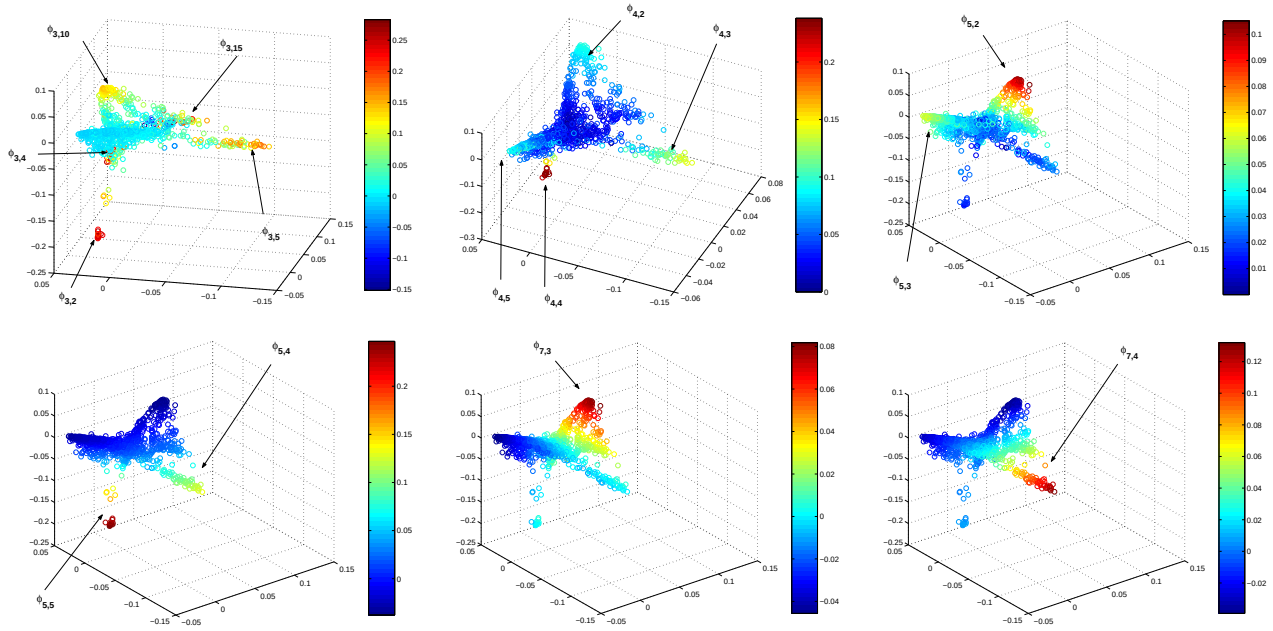

Figure 2: Scaling functions at different scales represented on the set embedded in R^3 via $(\xi_3(x), \xi_4(x), \xi_5(x))$. $\phi_{3,4}$ is about Mathematics, but in particular applications to networks, encryption and number theory; $\phi_{3,10}$ is about Astronomy, but in particular papers in X-ray 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.

YALE/DCS/TR-1304, Yale Univ. 2004, Appl. Comp. Harm. Anal., 21(1):95–112, July 2006. (Tech. Rep. YALE/DCS/TR-1304, Yale Univ., Sep. 2004).

[6] F. Chung. *Spectral Graph Theory.* American Mathematical Society, 1997.

[7] R.R. Coifman and S. Lafon. Diffusion maps. *Appl. Comp. Harm. Anal.*, 21(1):5–30, 2006.

[8] R.R Coifman and M. Maggioni. Diffusion wavelets. *Appl. Comp. Harm. Anal.*, 21(1):53–94, July 2006. (Tech. Rep. YALE/DCS/TR-1303, Yale Univ., Sep. 2004).

[9] Stephane Lafon. *Diffusion maps and geometric harmonics.* PhD thesis, Yale University, Dept of Mathematics & Applied Mathematics, 2004.

[10] M. Maggioni and S. Mahadevan. Fast direct policy evaluation using multiscale analysis of markov diffusion processes. In *University of Massachusetts, Department of Computer Science Technical Report TR-2005-39; accepted at ICML 2006*, 2005.

[11] S. Mahadevan and M. Maggioni. Value function approximation with diffusion wavelets and laplacian eigenfunctions. In *University of Massachusetts, Department of Computer Science Technical Report TR-2005-38; Proc. NIPS 2005*, 2005.

[12] P. Niyogi, I. Matveeva, and M. Belkin. Regression and regularization on large graphs. Technical report, University of Chicago, Nov. 2003.