Final Report for *Geometric Analysis for Data Reduction and Structure Discovery* DE-FG02-10ER25983, STRIPES award # DE-SC0004096

Kevin R. Vixie

With contributions by:

William K Allard
Mauro Maggioni
Francois G. Meyer

Washington State University, Duke University, and University of Colorado

Contents

1 Introduction

This final report summarizes the results from the research work supported or partially supported by the DOE grant *Geometric Analysis for Data Reduction and Structure Discovery*. While work that was conducted throughout the project is covered here, we concentrate on the components receiving most of the attention from the participants since the interim report dated Feb 11, 2011.

The work, aimed at exploiting insights from geometric analysis for the unraveling of large scale data challenges, was comprised of three main threads:

1. Exploitation of tools from geometric measure theory for the development of metrics on complicated sets in $\mathbb{R}^2$ and $\mathbb{R}^3$ (see Section 2),
2. development of nonlinear dimension reductions and parameterizations using both graph-based tools and local approximations, as well as an important subthread that looked at the stability of local analysis tools (See Section 3), and
3. development of multiscale nonlinear, geometric decompositions of complicated data sets in high dimensions (see Section 4). A subthread for this work was the cover tree algorithms that were very carefully explored and implemented (See section 5).

2 Tools from Geometric Measure Theory for Data (K.R. Vixie, S. Ibrahim, B. Krishnamoorthy, G. Sandine, T. Asaki, H. Van Dyke, B. Van Dyke)

2.1 Summary of Results

Our work in this area resulted in several, related sub-projects: (1) the multiscale simplicial flat norm, (2) flat norm decompositions, (3) fast computation of flat norm and flat norm surrogate distances, (4) flat norm on graphs, (5) properties of L^1TV minimizers, (6) non-asymptotic densities, and (7) monotonicity in higher dimensions. We give some more details of the results in (1), (2), and (4) in the next subsection. Here are the papers that have resulted from the work:

Simplicial Flat norm and deformation theorem [15] In this paper, the notion of a multiscale flat norm for simplicial complexes is introduced. A deformation theorem is established with better constants than previously established in Sullivan's dissertation in which he proves a deformation theorem for cell complexes. We do that by exploiting the simplicial setting.

Flat Norm Decompositions [16] In this paper we show that there are integral minimizers to the flat norm minimization for general 2 dimensional integral current inputs and that modulo a conjecture in higher dimensions, the same holds for all co-dimension 1 integral current inputs. The conjecture, which seems to be true but tedious to prove, involves the existence of subdivisions with controlled irregularity.

Nonasymptotic Densities [17] In this paper, a non-asymptotic notion of density is used to characterize 2 dimensional sets in the plane, It is shown that for a wide variety of nice shapes, a signature obtained by taking measurements along the boundary of the shape determine the shape.

Monotonicity in higher dimensions [13] Multiple versions of a notion of monotonicity were defined and explored in this paper. After the paper was published it was discovered that some significant pieces were in fact a duplication of earlier work. In spite of this fact, there are new contributions in the paper.

Thin Minimizers for L^1TV [12] We generated some examples of minimizers of the L^1TV functional that are surprisingly thin. This result helps tighten the characterization the global properties of the minimizers. Whereas L^1TV minimizers for convex input sets are the union of balls of radius $\frac{1}{\lambda}$ contained in the input set, the result in this paper shows that the same is **not** true for non-convex inputs to the L^1TV functional.

2.2 Multiscale Distances for shapes and generalized manifolds

Shapes are often thought of as sets in $\mathbb{R}^2$ or $\mathbb{R}^3$ or as boundaries of those sets. Often they are modeled as submanifolds. There are a variety of distances used, for example, to compute distances between parameterized curves in $\mathbb{R}^2$ and $\mathbb{R}^3$. The multiscale flat norm [35] that our work makes more usable and computable, gives us a much more general and powerful distance or metric. Much of the increase in power comes from the fact that multiscale flat norm measures the distance between *currents*, which include not only oriented manifolds of any dimension, but also much, much more general sets and measures in $\mathbb{R}^n$ (or even in manifolds or metric spaces). The results obtained in this work extend the multi-scale flat norm to simplicial complexes and graphs in vector spaces, with suggestions of how to move it to metric spaces. Our results also establish important regularity results that in effect make it much more usable (at least if one cares about rigorously defensible use). Efforts towards fast computation were made and the initial results were encouraging. Those results were reported at the DOE PI meeting in DC in 2011.[1]

Note: Recall that a set is said to be co-dimension k when its dimension is k less than the space it sits in. In $\mathbb{R}^3$, a 1-dimensional curve is therefore co-dimension 2 and a 2-dimensional surface is co-dimension 1. Thinking of k-dimensional currents as unions of oriented k-manifolds will be sufficient for most purposes. While it is important that they are much more general than that, a great deal of the intuitive impact of our results can be obtained thinking of currents as unions of pieces of manifolds.

2.2.1 Simplicial Flat Norm

Simplicial approximations and representations are ubiquitous in science and engineering. It therefore made sense to consider the specialization of the flat norm to this setting. To do this, we proved a new version of the deformation theorem, a theorem that is a fundamental piece of the theory of currents in $\mathbb{R}^n$. In important cases, our result yields better constants than Sullivan's more general deformation theorem for cell complexes. This work also generalizes the integrality result for integral current inputs when everything is simplicial.

2.2.2 Flat Norm Decompositions

The question of when the implicit flat norm computation minimization has integral solutions for integral input currents, was only known in the case of co-dimension 1 bound-

[1] This part of the project ran afoul of some very serious personnel issues and was not developed as far as I would have liked.

aries. In this collaboration between Ibrahim, Krishnamoorthy and Vixie, that result was extended to general 1-dimensional integral currents in $\mathbb{R}^2$ and is established (modulo a conjecture we believe to be true) for n-1 dimensional integral currents in $\mathbb{R}^n$. We also prove that this cannot be expected for currents whose co-dimension is 3 or higher. Only co-dimension 2 is left completely open.

Practically speaking, this work permits us to conclude that in may practical cases, relaxation of the flat norm problem to normal currents (currents corresponding to very general functions and measures of any dimension) still permits us to find integral solutions. For example, if we compute the distance between sets which are oriented manifolds or unions of oriented manifolds (with or without boundaries), we are guaranteed that the implicit minimization problem has minimizers of the same type.

2.2.3 Flat Norm on graphs

In work in 2009, vixie et al. [37] used an approach to computing the flat norm in the form of the L^1TV functional for images in 2 and 3 dimensions. The approach that was used there to compute the weights for the min-cut max-flow method for computing the minimizers suggested an approach for general graphs embedded in $\mathbb{R}^n$. This was tackled by Sandine in his masters project which will be defended soon. The longer term goal for this project is the exploration of the geometry of data sets in metric spaces.

2.3 Students

The following students worked on research supported by this funding or enabled by this funding:

Eric Larson: Masters project, *Lipschitz mappings and doubling measures*. He is currently a PhD student at UCLA in the department of mathematics. His Masters Advisor was Kevin R. Vixie.

Heather Van Dyke: Dissertation title, *A Study of p-Variation and the p-Laplacian for $0 < p \leq 1$ and Finite Hyperplane Traversal Algorithms for Signal Processing.* She has changed her name to Heather Moon and is now a tenure track assistant professor at St Mary's College in Maryland. Her advisor was T. Asaki.
http://faculty.smcm.edu/hamoon/HVDthesis.pdf

Benjamin Van Dyke Dissertation title, *Directional Direct-Search Optimization Methods with Polling Directions Based on Equal Angle Distributions*. Ben is a visiting assistant Professor at Walla Walla University in College Place, Washington. His Main Advisor was T. Asaki.

Keith Clawson Worked on calculation of the flat norm through the dual formulation as well as the fast computation of surrogates for the flat norm on streaming data. He did not finish and is currently contributing to the Sage project at UW and as a systems administrator in the Seattle area.

Sharif Ibrahim: Dissertation Title, *Data-inspired advances in geometric measure theory: generalized surface and shape metrics*. His Advisor was Kevin R. Vixie, http://arxiv.org/abs/1408.5954.

Gary Sandine: Masters Project, *Extension of the multiscale Flat norm to arbitrary graphs in vectorspaces*. Kevin R. Vixie was a co-advisor on his masters committee and was the faculty member directing his project. His Official Advisor was Marios Patichis in EECS at UNM in Albuquerque, NM.

Altaa Tumurbaatar Masters thesis, *Statistics in Simplicial Shape Spaces*. Bala Krishnamoorthy and Kevin R. Vixie are directing the research, along with Krishnamoorthy Sivakumar in EECS at WSU. While this work was initiated after the funding expired, the work is a direct follow on and was made possible as a direct result of the work carried out under this funding.

2.4 Research Groups

As a follow on to the work funded by this grant, the PI has started a new research group at WSU focused on research at the interface of pure/applied analysis and big data. This effort is a collaboration with 4 other faculty in the WSU mathematics department as well as industrial scientists. The funding from this grant laid the foundation for this new promising effort. A link to the group can be found here: http://analysisplusdata.org.

2.5 2012 Summer School

This grant partially supported the 2012 Summer school on Geometry and Data. The appendix below contains summer school details. There were about 50 participants. The webpage for the summer school can be found here:
http://geometricanalysis.org/Workshops/2012SummerSchool. Summer School Fellows were spread over the areas of mathematics, electrical engineering and computer science. For a complete list of Summer School Fellows, please consult the webpage.

Instead of giving a breakdown of what the school covered (that can be found on the website), here is what the participants had to say about the summer school:

> *Thanks for organizing such a great event to promote geometry education and research. Geometry is well used in almost every area in computer science, path planning in robotics, manifold learning, geometric modeling, etc. But still the education is not coherent and prerequisite courses are rare. I found such a summer school will well motivate students to study geometry. Thanks for the opportunities for me and my student to get involved in the summer school. Personally I also enjoyed very much meeting these friendly scholars and the nice volleyball game.* Yalin Wang, Arizona State University

Thanks again for organizing the summer school–it was very educational and beneficial specially for people with an engineering background (like myself). I really appreciate the efforts that you put into this event and wish you more success in the future ones. Armin Eftekhari, Colorado School of Mines

I found this to be a very fruitful time in Moscow. I spent many evenings writing notes about what I learned - I feel like a huge area of potential research work is now possible now that I've learned more about the topics discussed (particularly intrigued by David H. And Bala's work, as well as the patch graphs and revisiting our old paper on the minimum spanning tree based stuff). Also, less applied to my day to day work, I think I can actually make a bit more sense of Frank Morgan's book now too - that has been on my "to understand" list for a number of years since you put it in my hands at LANL. Matthew Sottile, Galois (Portland, Oregon)

I wanted to let you know that I really appreciate the opportunity to participate in the summer school this past year. As one of the younger participants and also as a new graduate student in mathematics, I felt like it was an outstanding way to experience some of the directions and practices of modern mathematical research. In addition it was a fantastic way to meet and connect with people, both those known such as yourself and Dr. Allard and those newer to the field, for both professional and intellectual future development. In particular, I think the balance struck between funding for interesting and significant speakers and for those of us just there to listen was excellent. I know I would not have been able to attend without some form of funding, and I appreciate that once I did arrive my time was well spent attending the talks and other events. Andrew Farrar, Oregon State University

I thought the summer school was thorough and comprehensive, yet accessible. I increased in my mathematical maturity through not only during the many informative lectures, but especially through the insightful and friendly conversations with mathematicians from varied levels, backgrounds, and perspectives. This atmosphere was in no way an accident, and I appreciate the obvious effort that you put into making the school a success. Josh Cruz, Washington State University

The summer school provided me with the opportunity to learn about geometry and data analysis from several different perspectives, many of which were new to me. I now have an awareness of the relationship between many topics in geometry and data and can approach research problems from new angles. Such multifaceted approaches will prove valuable as problems become very complex in high dimensions. I also made many contacts and greatly expanded my professional network - a crucial element for success in my early career stage. Daniel Kaslovsky, University of Colorado

> *The choice of the papers presented was very pleasing - a nice combination of recent "hot" papers and less recent "classical" papers. Also the topics are of interest for people of different backgrounds - mathematics, statistics, electrical and computer engineering, which was proved by the diversity of the participants: from theoretical math students to data gurus, to both. Also the level of the students was variable - from undergrads to masters to thesis writers: it is never to early or too late to learn!*
>
> *It is impressing how all the selected presenters were talented speakers, which made the lectures quite engaging and truly interactive, an aspect crucial when bringing researchers from different fields.*
>
> *I found the lab sessions extremely useful and enjoyable. I have been to many summer schools and workshops where the format has been purely lecture oriented. This is definitely not the best way we learn. Trying to implement the ideas which were discussed is really testing our understanding. Also the little challenges were simple enough so that nobody gets discouraged but also open enough to allow deeper thinking and tinkering. The initial installation time was worth it since it allowed everybody to take part in the exercises and use the same platform. I think increasing the length of these sessions and probably adding some mathematical problem-solving sessions can make the summer school even better.*
>
> *Overall the atmosphere was home-like and there was a chance to interact with everybody during the breaks, lunch, in the hotel, and around town.*
>
> *I personally appreciate this chance to get caught up on the recent advances in these research topics and meet such a unique group of people. I am glad I stayed longer and I wish I could have stayed all the three weeks.* Valentina Staneva, Johns Hopkins University

2.6 Current Directions

This funding enabled the establishment of a research program that is now beginning to look at the intersection of (1) analysis and GMT in metric spaces and (2) data science. Data often has a natural distance associated with it, but no easy or natural isometric embedding of the data in some vector space. Thus there is also a practical motivation for the research. One can always generate a graph on which a metric is defined by shortest paths, and that is the starting point for some of our work in this area.

Existing work at the interface between (1) stochastic tools, (2) geometric analysis and (3) uncertainty quantification is minimal *in comparison* to the potential presented by this interface. We are beginning to work at that interface. Statistics in shape spaces, or more generally, in spaces of currents, is the subject of one project that is now underway using the computational tools generated in our work on the simplicial flat norm.

In spite of some very difficult personnel issues, we have regrouped and now have a collaboration that involves five faculty members in mathematics, several industry scientists and about 15 graduate and undergraduate students. We are also building connections to faculty members in EECS and Economics.

3 Low Dimensional, Nonlinear Sets in High Dimensions (F. Meyer, D. Kaslovsky, K. Taylor, N. Monnig, N. Bertrand and J. Ramirez)

This project had several major goals. The first goal was the investigation of novel methods to parameterize low-dimensional datasets using nonlinear techniques. The project was specifically interested in studying the low-dimensional structure of datasets formed by collecting patches from signals and images. The second goal of the project was the development of interactive tools to facilitate the analysis of a complex high-dimensional dataset by an analyst using nonlinear dimension reduction.

3.1 Significant Results

3.1.1 Searching for the anomalies in massive datasets of time series

In many areas of science and engineering the only method to study a complex system entails making indirect observations of the state of the system. For instance, in geophysics, recordings of the earth ground motion made with a seismogram provide indirect measurements of the complicated physical process that gives rise to a seismic event and the associated seismic waves. Unfortunately, the complexity of the physical processes involved at all scales in these systems currently prevents the derivation from first principles of a precise model that predicts the configuration of the system, given the measurements.

An indirect approach involves replacing the unknown space of all possible configurations (phase space) of the system with an equivalent phase space estimated from the measurements. Indeed, for each time series, one can slide a window in time, which we call a temporal patch, or *patch*. For each patch, we collect the d values of the time series within the patch, and stack them into a d-dimensional vector. The trajectory of the patch in d dimensions characterizes the dynamics of the measurements. This process is known as *time-delay embedding*, and there exists a rich literature on the equivalence between the trajectory of a vector of d consecutive measurements from a dynamical system, and the properties of the corresponding dynamical system.

In the analysis of seismograms, we proposed a novel perspective on the concept of time-delay embedding by combining the experimental phase spaces (patch-trajectories) collected from several seismograms. We proposed to compute a nonlinear parametrization of the combined patch trajectories. This nonlinear parametrization assigns to each patch a small number of coordinates that uniquely characterizes the state of the dynamical system at the corresponding instant. The parametrization relies on the assumption that the union of trajectories lies along a smooth set in $\mathbb{R}^d$, which can be parametrized using a nonlinear method based on the eigenvectors of the graph Laplacian.

The problem of detecting seismic wave packets becomes then the problem of characterizing the regions of the *patch space* associated with seismic activity. In collaboration with graduate student Kye Taylor, we implemented this approach and developed an algorithm to detect seismic wave packets and compute their arrival time. The method led to the development of a software package that was carefully evaluated by scientists in the

Next Generation Monitoring Systems group at Sandia National Labs. Our algorithm outperformed the existing gold standards [36]. Our work concluded that the set of patches contained nonlinear structures that could not be well approximated by linear methods, such as a principal component analysis, or a wavelet transform. Furthermore, our study confirmed that the combined phase spaces associated with regional seismic waves were remarkably low-dimensional: we needed only 25 dimensions (instead of the $d = 1024$ dimensions of each patch) to optimally detect seismic waves.

3.1.2 Random Graph Models for Datasets of Image Patches

Inspired by our work on temporal patches, we looked for an explanation for the success of algorithms that organize image patches according to graph-based metrics. Indeed, algorithms that analyze patches extracted from images have led to state-of-the art image processing methods for denoising, inpainting, and super resolution. In collaboration with graduate student Kye Taylor, we provided a theoretical explanation for these experimental observations [36]. Our approach relied on a detailed analysis of the commute time metric (a notion of distance on a graph) on prototypical graph models that epitomize the geometry observed in general image-patch graphs. We proved that a parametrization of the graph based on commute time shrinks the mutual distances between patches that contain textures and edges, while the distances between patches that contain uniform (or slowly varying) intensity expand. In effect, our results explain why the parametrization of the set of patches based on the eigenfunctions of the Laplacian can concentrate patches that correspond to rapid local changes of the intensity, which would otherwise be scattered in the space of patches.

Of course, in practice noisy images result in noisy patches, and the graph of patches extracted from a noisy image is a perturbed version of the graph constructed using the "clean" patches. The question becomes: what is the influence of the perturbation of the graph on the eigenvectors of the graph Laplacian, which are used to parametrize the set of patches. We studied this problem using an experimental approach [31, 29]. It turns out that the low frequency eigenfunctions of the graph Laplacian are remarkably stable to perturbation of the graph.

3.1.3 Non-Asymptotic Analysis of Tangent Space Perturbation

Many datasets collected from physical or biological systems are high-dimensional. In reality, many of the internal variables of the system being measured are coupled, resulting in a potentially dramatic reduction of the true degrees of freedom for the measurements. Our ability to efficiently re-parametrize a dataset to take advantage of its intrinsic low-dimensional structure is therefore fundamental. While it is very rare that the datapoints lie exactly in a linear subspace, it is often the case that the points organize themselves along a smooth low-dimensional manifold. A fundamental problem in processing such datasets is the construction of an efficient parametrization that allows for the data to be well-represented in fewer dimensions. Such a parametrization may be realized by exploiting the inherent manifold structure of the data. However, discovering the geometry of an underlying manifold from only noisy samples remains an open topic of research.

One approach consists in recovering a local parametrization using the local geometric information provided by the tangent planes. Principal component analysis (PCA) is often the tool of choice, as it returns an optimal basis in the case of noise-free samples from a linear subspace. To process noisy data, PCA must be applied locally, at a scale small enough such that the manifold is approximately linear, but at a scale large enough such that structure may be discerned from noise. With graduate student Daniel Kaslovsky (now an NSF postdoctoral fellow), we used eigenspace perturbation theory to analyze the stability of the subspace estimated by PCA as a function of scale, and bound (with high probability) the angle it forms with the true tangent space. By adaptively selecting the scale that minimizes this bound, our analysis reveals the existence of an optimal scale for local tangent plane recovery. This is a fundamental problem that has been a subject of interest for decades in dynamical systems theory, and has now very practical application, as more and more geometrically-inspired algorithms rely on a notion of locality to select nearest neighbors. This work was published in Information and Inference: a Journal of the IMA [25]. Some initial versions of this work were presented in several conferences [19, 20, 22, 24, 21, 23].

3.1.4 Interactive exploration of manifold

Meyer and graduate student Monnig addressed the problem of computing the inverse of a general smooth bi-Lipschitz nonlinear dimensionality reduction mapping over all points in the image of the forward map.The approach relies on a scale free radial basis functions (RBFs) to interpolate the inverse map everywhere on the low-dimensional range of the forward map. Our algorithm provides a computational solution to a longstanding problem in the machine learning and data science community. This work was published in Applied and Computational Harmonic Analysis [32], and presented at several conferences [30, 33].

3.2 Key outcomes

PI Meyer and his graduate students published five journal papers. Two papers were published in the proceedings of international conferences. During the award, Meyer graduated two Ph.D. students: Kaslovsky is a Data Scientist at Seagate, Taylor is a faculty at Tufts University. Two graduate student working on the project graduated with M.S. They are both working toward a Ph.D.: Ramirez atd Duke University, Bertrand at the Georgia Institute of Technology.

3.3 Products

3.3.1 Journal papers

1. Kaslovsky D.N., and **Meyer F.G.**, "Non-Asymptotic Analysis of Tangent Space Perturbation", *Information and Inference: A Journal of the IMA*, 3 (2), pp 134–187, http://dx.doi.org/10.1093/imaiai/iau004.

2. Monnig N., Fornberg B., **Meyer F.G.**, "Inverting Non-Linear Dimensionality Reduction with Scale-Free Radial Basis Interpolation", *Applied and Computational Harmonic Analysis*, 37(1), pp 162-170, 2014, http://dx.doi.org/10.1016/j.acha.2013.10.004.

3. **Meyer F.G.**, and Shen X., "Perturbation of the Eigenvectors of the Graph Laplacian: Application to Image Denoising"; *Applied and Computational Harmonic Analysis*, 36(2), pp 326–334, 2014, http://dx.doi.org/10.1016/j.acha.2013.06.004

4. K.M. Taylor and **F.G. Meyer**,"A random walk on image patches", *SIAM Journal on Imaging Sciences*, 5(2), pp 688-725, 2012, http://dx.doi.org/10.1137/110839370.

5. KM Taylor, MJ Procopio CJ Young and **F.G. Meyer**, "Estimation of arrival times from seismic waves: a manifold-based approach", *Geophysical Journal International*, 185 (1), pp 435–452, 2011, http://dx.doi.org/10.1111/j.1365-246X.2011.04947.x.

3.3.2 Conference papers

1. Kaslovsky D.N., **Meyer F.G.**, "Overcoming noise, avoiding curvature: optimal scale selection for tangent plane recovery", *Proc. IEEE Statistical Signal Processing Workshop*, pp. 904-907, 2012.

2. Ramirez J., **Meyer F.G.**,"Machine Learning for Seismic Signal Processing: Seismic Phase Classification on a Manifold", *in Proc. IEEE International Conference on Machine Learning and Applications*, pp 382-388, 2011. [Acceptance rate: 27%]

3. **Meyer F.G.**, Taylor KM, Kaslovsky D., Procopio MJ, and Young CJ "Evaluation of Empirical Mode Decomposition and Chirplet Transform for Regional Seismic Phase Detection and Identification", Seismological Society of America 2009 Annual Meeting, Seismological Research Letters, Volume 80, No. 2 p 347, 2009.

3.3.3 Invited Conference Presentations with no Proceedings

1. **Meyer F.G.**, Invited Speaker, *5th International Conference on Computational Harmonic Analysis*, Vanderbilt University, May 2014.

2. N.D. Monnig and B. Fornberg and **Meyer F.G.**, "Nonlinear Dimensionality Reduction: The Inverse Map", *Workshop on Electrical Flows, Graph Laplacians, and Algorithms: Spectral Graph Theory and Beyond'*, ICERM, Brown University, 2014.

3. **Meyer F.G.**, Kaslovsky D.N., and B. Wohlberg, "Analysis of image patches: a unified geometric perspective", *SIAM Conference on Imaging Science*, 2012.

4. Ramirez J, **Meyer F.G.**, "Machine Learning for Seismic Signal Processing: Phase classification of seismic events on a manifold"; *Society for the Advancement of Chicanos and Native Americans in Science National Conference*, 2011. Juan Ramirez received the Student Research Presentation Award (category: Applied Mathematics).

5. Kaslovsky D.N. and **Meyer F.G.**, "Image Manifolds: Processing Along the Tangent Plane", *International Congress on Industrial and Applied Mathematics*, 2011.

6. **Meyer F.G.**, "Image de-noising on the manifold of patches: a spectral approach", *SIAM Conference on Imaging Science*, 2010.

7. **Meyer F.G.**, "Exploring the Manifold of Seismic Waves: Application to Phase Detection", *Random Shapes Reunion Conference II*, Institute for Pure and Applied Mathematics, UCLA, December 6-11, 2009

3.3.4 Invited lectures

1. "Nonlinear Dimensionality Reduction: The Inverse Map", *Department of Mathematics Colloquium, Washington University in St. Louis*, November 7, 2013.

2. "Low-Dimensional Representations of High-Dimensional Datasets: A Geometric Perspective", *Department of Electrical Engineering Seminar, University of Colorado at Boulder*, Nov. 2012.

3. "A Random Walk on Image Patches", *Electrical & Computer Engineering Seminar, Colorado State University*, April 2, 2012.

4. "A Random Walk on Image Patches", *Applied Mathematics Seminar, Yale University*, February 22, 2012.

5. "A Random Walk on Image Patches", *PACM/Applied Mathematics Colloquium, Princeton University*, February 20, 2012.

6. "A Random Walk on Image Patches", *Department of Mathematics Colloquium, Washington University in St. Louis*, February 16, 2012.

7. "Image Manifolds: Processing Along the Tangent Plane", Department of Mathematics Seminar, Washington State University, January 2011

8. "A Random Walk on Image Patches", Applied Mathematics Colloquium, University of Colorado at Boulder, October 2011

9. "Exploring the manifold of patches: a spectral approach", Bigroup Seminar, University of Colorado at Boulder and JILA, March 2011.

10. "Exploring the Manifold of Seismic Waves: Application to Phase Detection", *Applied Mathematics Seminar, Yale University*, March 2010.

11. "Exploring the Manifold of Seismic Waves", *CARDI seminar, Colorado School of Mines*, Feb 2010.

12. "Image de-noising on the manifold of patches: a spectral approach", *COSI Seminar Series, University of Colorado at Boulder,* March 2009.

3.3.5 Software

We provide MATLAB code to implement the algorithms described in the different papers at the permanent link: http://ecee.colorado.edu/~fmeyer/software.html

3.3.6 Thesis/Dissertations

1. Daniel Kaslowsky, Ph.D. Applied Mathematics, 2012;*Geometric Sparsity in High Dimension*.
2. Juan Ramirez. M.S. Electrical Engineering; 2012; *Learning from Manifold-Valued Data: An Application to Seismic Signal Processing*.
3. Kye Taylor, Ph.D. Applied Mathematics, 2011; *The geometry of signal and image patch-sets*.

3.4 Education

Meyer developed a graduate level course on the analysis of high-dimensional datasets, http://ecee.colorado.edu/~fmeyer/class/ecen5322/. The course is centered around the concept of concentration of measure in high-dimensions and its consequences, such as the Johnson-Lindenstrauss theorem. Meyer was awarded the *2010 Holland Teaching Excellence Award in Electrical and Computer Engineering* for developing this new course.

3.5 Impacts

This project had several major impacts on problems related to the analysis of complex datasets that depend on a small number of parameters.

The first major impact involves a novel theoretical understanding of algorithms that organize patches extracted from images and signals using a similarity graph. Indeed, its has become a common practice to use graphs to model dependencies between points in large datasets. For instance, algorithms that analyze patches extracted from images provide state-of-the art image processing methods. This project provided an unprecedented understanding of the success of such algorithms. Our approach led us to the construction, and subsequent analysis, of novel graph models to describe functional dependencies between signal and image patches. In a larger context, this project has shown that our results enable scientific discovery in scientific fields such as geosciences, neuroscience, etc. by providing novel tools to detect interesting patterns and anomalies and make inferences from noisy data.

The second major impact involves the analysis of high-dimensional datasets. Such data are usually presented to the practitioner as measurements that depend on many coordinates, or parameters. Fortunately, many of the internal variables of the system being measured are usually coupled, resulting in a potentially dramatic reduction of the true degrees of freedom. Our ability to efficiently re-parametrize a dataset to take advantage of its intrinsic low-dimensional structure is therefore fundamental. This project studied the problem of constructing a local chart of a noisy dataset that lies close to a smooth low-dimensional structure. Specifically, we studied the distortion of the map as a function of the amount of noise and how fast if the dataset departing from a local plane. This analysis provide a fundamental answer to the question of choosing the right scale to estimate a low-dimensional representation of a noisy and curved dataset. This work has immediate application to the analysis of large and complex dataset.

4 Multiscale Geometric Analysis: Geometric Wavelets (M. Maggioni, W.K. Allard, G. Chen)

4.1 Main Results

The main goal of this project was development of novel data structures for the analysis of high-dimensional data sets, and extraction of information from them. We developed a novel multi-scale data structure, called Geometric-Multi-Resolution Analysis, that maps a whole data set to a novel representation which is both hierarchically organized and yields sparse (or compressible) representations of data sets.

On the one hand, this generalized standard linear dimension reduction procedures such as Principal Component Analysis, and fits into the large body of research on nonlinear dimensionality reduction and manifold learning that was developed in the last several years. On the other hand, the data representation we introduced has completely novel aspects: it is multi-scale, with each scale having its own accuracy in the approximation of the data, it comes with guarantees of such accuracy, it is achieved with fast algorithms (more on this later), that scale linearly in the size of the data or, even better, that may be run in an online fashion at a cost that is independent of the number of points and dependent only on the requested accuracy. Finally, both the computational cost and the sample complexity of this procedure depend fundamentally on a suitable, robust notion of intrinsic dimensionality of the data, therefore defying the curse of dimensionality.

We have explored generalizations and applications of GMRA in a wide variety of problems in machine and statistical learning for high-dimensional data problems, such as high-dimensional classification and regression, and anomaly detection. We also made a connection with dictionary learning, which is a well-studied problem in statistical signal processing and machine learning, where given data one wishes to construct a dictionary that yields sparse representations of such data: while existing techniques rely on algorithms that are expensive and at the same time lack performance guarantees, the GMRA yields a novel algorithm for constructing such dictionaries, together with a full set of guarantees. Indeed we performed a carefully finite sample analysis of the construction, showing that GMRA dictionaries yield the best known dictionary learning algorithms from every aspect (guarantees, accuracy, sparsity, running time), under suitable geometric conditions on the data (a robust notion of low-intrinsic dimension, that accommodates noise).

4.2 Geometric Multi-Resolution Analysis (GMRA)

We continued the work on Geometric Multi-Resolution Analysis (GMRA), pushing in several directions. First of all, we recall that a GMRA consists in a hierarchically organized family of portions of low-dimensional planes that approximate a manifold or a set of data points sampled from a manifold, together with an efficient of encoding of such family of portions of planes, and fast transforms mapping points on (or near to) the manifold to sets of sparse coefficients, and conversely a set of coefficients to a point on a manifold.

We recall that the construction of a GMRA proceeds in 3 stages:

(i) Construct **multiscale partitions** $\{\{C_{j,k}\}_{k\in\Gamma_j}\}_{j=0}^{J}$ of the data: for each j, $\mathcal{M}=\cup_{k\in\Gamma_j}C_{j,k}$, and $C_{j,k}$ is a nice "cube" at scale 2^{-j}. In practice we obtain $C_{j,k}$ by recursive spectral partitioning or cover trees.

(ii) Compute **low-rank** SVD of the local covariance: $\mathrm{cov}_{j,k}=\Phi_{j,k}\Sigma_{j,k}\Phi_{j,k}^T$. Let $P_{j,k}$ be the affine projection $\mathbb{R}^D\to\mathbb{V}_{j,k}:=\langle\Phi_{j,k}\rangle$ (local approximate tangent space): $P_{j,k}(x)=\Phi_{j,k}\Phi_{j,k}^*(x-c_{j,k})+c_{j,k}$. These pieces of planes $P_{j,k}(C_{j,k})$ form an approximation $\mathcal{M}_j$ to the original data $\mathcal{M}$; let $P_{\mathcal{M}_j}(x):=P_{j,k}(x)$ for $x\in C_{j,k}$.

(iii) We efficiently **encode the difference** $Q_{\mathcal{M}_{j+1}}$ between $P_{\mathcal{M}_{j+1}}(x)$ and $P_{\mathcal{M}_j}(x)$, by constructing affine "detail" operators analogous to the wavelet projections in wavelet theory.

We obtain a multiscale nonlinear transform mapping data to a multiscale family of pieces of planes. Fast algorithms and multiscale organization allow for fast pruning and optimization algorithms to be run on this multiscale structure.

4.3 Extensions & applications of GMRA

4.3.1 Online construction

The original construction could not be adapted to an online learning setting, where the points are seen incrementally rather than all at once. One of the main reasons was that the multiscale partitions used to construct the local planes where computed by using static tree structures used for nearest neighbor searches. W.K.Allard, as detailed in his portion of the report, has been adapting cover trees, a data structure for fast proximity searches in low-dimensional metric spaces, to the setting of interest, and the construction of this data structure, which yields multiscale partitions, is perfectly suited for online updates. It also theoretical guarantees in terms of geometric properties of the partitions, which are useful theoretically, as well as guarantees in terms of computational complexity. We are re-adapting the code in order to use these data structures.

4.3.2 Density Estimation

In Figure 1 and 2 we show some experiments were we performed density estimation as mentioned above on a simple manifold $\mathcal{M}$, and then distort the manifold as a function

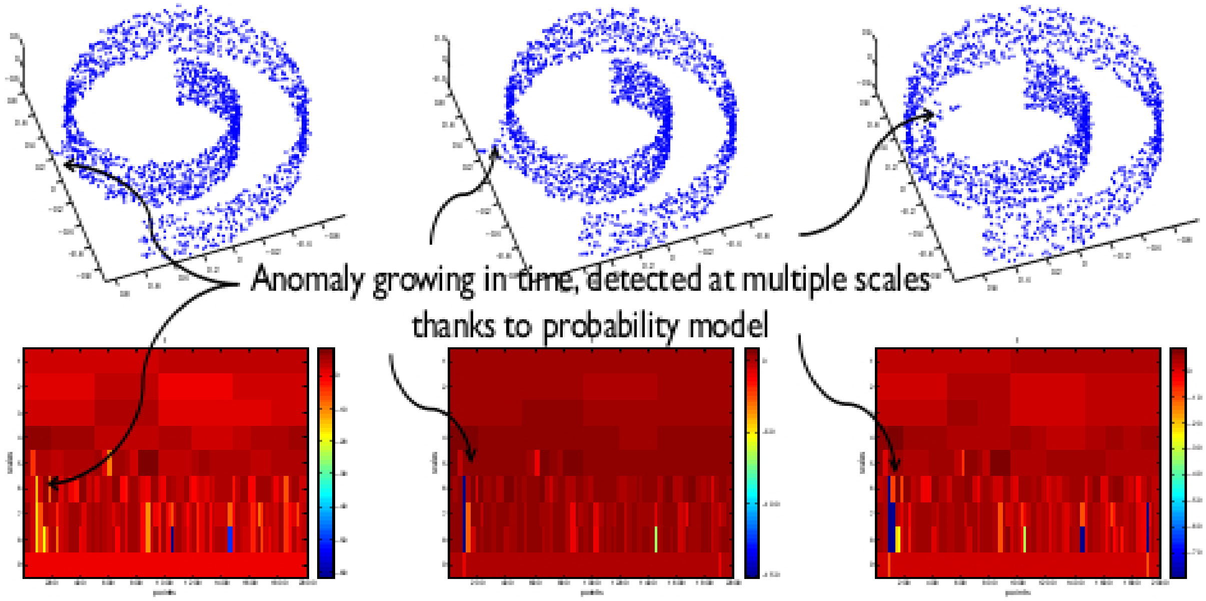
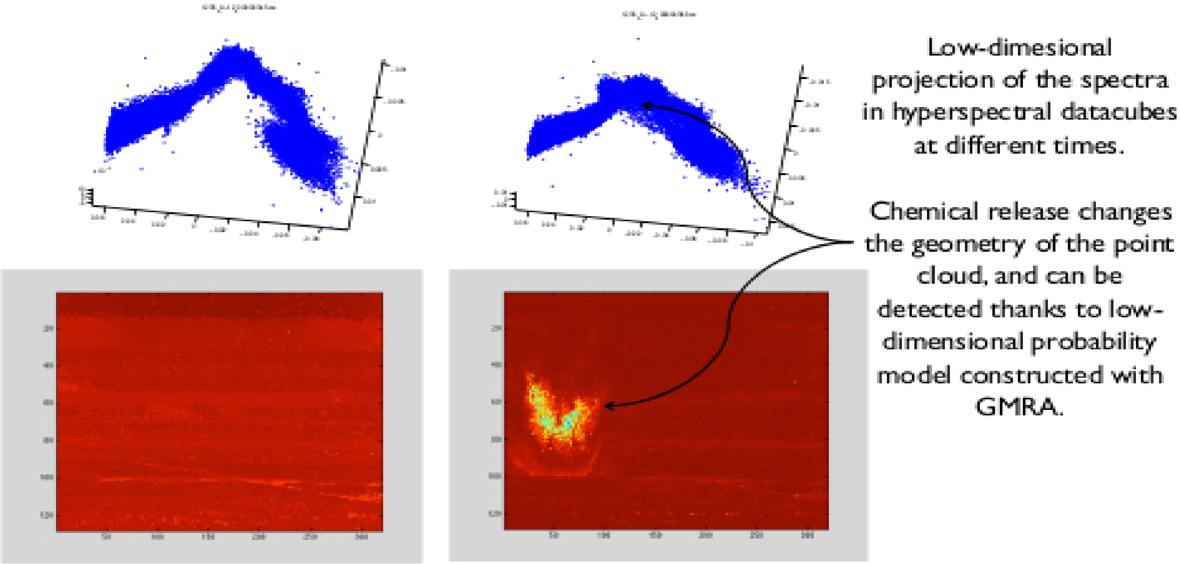


Figure 1: We superimpose to the geometry encoded by GMRA density estimators.This can be done in different ways, but we discovered a general framework for which we believe very strong guarantees can be proved, and still very fast algorithms exist. Once a density estimator (i.e. a model for the data) is constructed, a new point cloud can be tested against the model. Since these models have scale and location information, fit vs. anomaly is assessed at multiple scales and locations. A cusp growing in time on a simple manifold (top) affects the goodness of fit at different scales and locations (bottom, color in log scale represents multiscale measure of fit of the model to the data; columns correspond to points, and the regions of poor fit do correspond to the cusp.).

of time, obtaining $\mathcal{M}_t$, and measure the likelihood of seeing the points in $\mathcal{M}_t$ according to our multiscale density estimator. We see that the "anomaly" is being detected at the appropriate scales and locations. In Figure 2 we have real data, in the form of hyper spectral data cubes changing in time, and with the GMRA density estimator we are able to detect a chemical release occurring at time t, that affects the shape of the "manifold" of spectra $\mathcal{M}_t$ at a later time t.

While the original GMRA produces geometric approximations, it is of interest in many applications to also approximate the actual density of points, for example in order to attach likelihoods to changes in the density, for example because of data changing in time, or anomalies developing. Instead of approximating a manifold $\mathcal{M}$ the goal is to estimate the probability distribution μ of the data. Very general framework for *geometric density estimation*:

- start with a space of local models $\mathcal{F}_{C_{j,k}}$, a subset of probability measures that are supported in $C_{j,k}$, a region at scale j in the GMRA. For example it could be the uniform normalized Hausdorff measure on sets in the form $\pi \cap C_{j,k}$, where π is a plane of dimension d, of an appropriately truncated and normalized version of $\mathcal{N}(m, C)$, a Gaussian distribution of mean m and covariance C, restricted to $C_{j,k}$.



Low-dimesional projection of the spectra in hyperspectral datacubes at different times.

Chemical release changes the geometry of the point cloud, and can be detected thanks to low-dimensional probability model constructed with GMRA.

Figure 2: Cube: 320x256x120 every 8sec. Spectra from 8μm to 11.7μm. Each data set contains 5-10 min. 4 distinct chemical types, 3 chemical release mechanisms, 2 different locations. Currently developing anomaly detection algorithms with GMRA. In this case we construct a model of normal spectra on the first cube in the movie. We then test the model on the spectra at time t.Anomalies are detected as points with very small probability.

- While each $\mathcal{F}_{C_{j,k}}$ is "small", we consider the (typically) much larger space obtained by picking a convex combination of measures in the"local" spaces associated with a partition. It is within $\mathcal{F}$ that we look for an approximation $\hat{\mu}_n$, dependent on n observed samples, to the probability measure μ, aiming at being close to the optimal $\hat{\mu} := argmin_{\nu\in\mathcal{F}} W_2(\nu, \mu)$.
- We develop *fast multiscale adaptive algorithms*, with finite sample guarantees on the distance between $\hat{\mu}_n$ and μ, depending only on intrinsic dimension, complexity of the local model class, and approximability of μ. We can even perform these *online*.

4.3.3 Manifold compressed sensing

We are developing a compressive sensing framework for GMRA: given a GMRA for $\mathcal{M}$, can we measure a small number of linear projections of a new point x, on or near the manifold $\mathcal{M}$, and reconstruct x? We have results showing that the answer is yes as soon as the number of linear measurements is of the order of the intrinsic dimension of $\mathcal{M}$, and the measurements satisfy certain conditions. Such conditions are indeed satisfied with high probability by random projections, for example, but we are considering the problem of adaptively designing a measurement matrix, which may yield even better reconstruction. All of this generalizes compressive sensing and compressive reconstruction algorithms from the standard sparsity setting to the case of nonlinear low-dimensional manifolds.

This work lead to the paper [18]. A corresponding Matlab toolbox is published on the M. Maggioni's webpage, together with the paper, and it is also available together with the paper on the publisher's website, in the spirit of reproducible research.

4.3.4 Robust Dictionary Learning

In the paper [27] we use GMRA to construct dictionaries from data, with guarantees on the quality of the dictionary, and fast algorithms for both the construction of the dictionaries and fast transforms from data to dictionary coefficients and back. The problem may be stated as follows: we are given i.i.d. samples $x_1, \ldots, x_n$ from an unknown probability measure Π in $\mathbb{R}^D$ whose support is concentrated (in a suitable technical sense) near a d-dimensional manifold $\mathcal{M}$ in $\mathbb{R}^D$, with $d \ll D$ being the interesting case. The goal is to construct a dictionary $\Phi = \{\varphi_1, \ldots, \varphi_m\} \subset \mathbb{R}^D$ such that if x is a sample from Π, there exists $\alpha \in \mathbb{R}^m$ such that $||x - \Phi\alpha|| < \epsilon$ and $||\alpha||_0 \leq k$. We would like m, ϵ, k to be all small, but clearly there are tradeoffs between them (for example, the large m is, the smaller we should be able to make ϵ and k), as well as tradeoffs with n, the number of samples at hand. In this paper we show that a GMRA-based construction leads to the following: if ϵ is fixed and given as a desired accuracy, if $n \gtrsim \epsilon^{-1+\frac{d}{2}}$, then with high probability we obtain a dictionary Φ with $O(\epsilon^{-\frac{d}{2}})$ elements, that achieves accuracy ϵ on all samples x from Π, and any sample x from Π will have a $(d+1)$-sparse representations in terms of Φ. The algorithm for constructing Φ is fast, scaling linearly in the size of the input data matrix, and exponentially in the intrinsic dimension d of the data, and the map from x to the set of coefficients α that yields a sparse representation for x is also fast. All the rates depend crucially on the intrinsic dimension d and not on the ambient dimension of the space. The (rather long) paper has been submitted for publication in Journal of Machine Learning Research and is currently available on the ArXiv. A toolbox for the paper is published on M.Maggioni's webpage.

4.3.5 Publications & Future Work

We have so far published the papers [2, 9, 34, 8, 7, 18, 26, 5, 10], posters [34], and large data visualization techniques. We have started writing a long paper on GMRA density estimation (together with other conference papers, e.g. [28]), and we extended these techniques to the study of high-dimensional systems, in particular to model reduction and homogenization problems [11], as well as to the fast construction of multi-resolution (in the spatial domain) dictionaries for images [14]. We submitted a long paper on multi-scale techniques for Markov Decision Processes, also inspired by the multi-scale constructions above [6] together with short papers on multi-scale methods for control of high-dimensional systems [4, 5]. We are also developing fast multi-scale techniques, inspired by the GMRA construction, for the calculation of optimal transportation distances and plans for high-dimensional point clouds (with S. Gerber).

4.3.6 Students and Postdocs supported by the award

G. Chen, J. Bouvrie, S. Gerber were Visiting Assistant Professors in the Mathematics Department at Duke University that were partially supported by this award. G. Chen is an author of [2] and several other papers under this award, developed an initial version of the GMRA code. J. Bouvrie worked on the generalization of GMRA-type ideas to control systems and Markov Decision Processes. S. Gerber used GMRA-type ideas to create new algorithms for the fast calculation of optimal transportation plans, as well as developing a fast version for dictionary learning on multi-scale patches of large databases of images. Eric Monson is a research scientist in Computer Science Department at Duke University who collaborated with the PI on the development of novel user interfaces and visualization techniques for high-dimensional data. GMRA-based visualization are enabling the visualizations of very large data sets in an interactive fashion, with minimal communication requirements with the server on which the data is stored, making it possible to use on mobile devices. Co-PI W.K. Allard besides co-authoring [2], was heavily involved in the design and development of the algorithms for GMRA, in particular for the fast, online hierarchical data structures underlying the construction.

5 Cover Trees (W. K. Allard)

5.1 Introduction

My work on this grant was concerned with the theory and application of a fundamental data structure called a **cover tree** which is used to provide a useful and efficient multiscale decomposition of a subset S of a metric space X. This structure was introduced in the paper ([3]) *Cover Trees for Nearest Neighbor* by Alina Beygelzimer, Sham Kakade and John Langford. In the preprint [1], supported by this grant, I refined and extended the work done in [3] in ways I will indicate below.

5.2 The algorithm

Whereas many algorithms designed for usefully decomposing S as above require the computation of all $\binom{N}{2}$ pairwise distances between the N points of S, the cover tree algorithm requires only $N \log N$ such distance calculations under hypotheses that are satisfied for many data sets of interest. Moreover, of all such decompositions, in a sense that can be made precise, that provided by the cover tree algorithm has the highest quality. In addition, and perhaps more importantly, the constant in front of the complexity mentioned a small and simple universal constant raised to the power of the **intrinsic dimension** of S. A precise an appealing definition of the intrinsic dimension, though certainly not the only reasonable one, is given in [1]. If one's goal is to build a decomposition of S which has the desirable qualities of a cover tree decomposition it is seems one cannot do better.

It is not at all obvious how to parallelize the cover tree algorithm. In [1] I provide what I believe to be a very efficient way to do this on a shared memory machine. The user must set one parameter. I have found that setting this parameter to 100 times the number of cores works pretty well. Very often the speedup is close to the number of cores.

5.3 The code

All told, I wrote around 7000 lines of C^{++} code to efficiently implement the construction of a cover tree as well as to perform a number of associated tasks, such as a k-nearest neighbor search. This code is written for shared memory machines and, most of the time, provides good speedup.

In addition, I have written Matlab wrappers using MEX so that a user can call the code from within Matlab.

I have outlined a distributed memory version of the code but have yet to implement it; the algorithm I have outlined is more naive than the one I wrote for shared memory. Having built such a program it is not difficult to build one that that works on a distributed shared memory machine.

5.4 GMRA.

Mauro Maggioni and his many collaborators are using this code to implement his beautiful Geometric Multiscale Resolution Analysis (GMRA) suite of constructions. The first step in any GMRA algorithm is the multiscale decomposition of a point cloud, a task for which the cover tree construction is ideally suited.

6 Summary

In spite of some very difficult personnel issues, the group as a whole made significant progress in pushing forward the research agenda focused on (1) exploiting insights and techniques in geometric analysis for data and conversely, (2) exploiting motivation from data analysis for further developments in geometric analysis.

References

[1] William K. Allard. *θ-covers*. Technical report, Mathematics Department, Duke University, 2013.

[2] William K. Allard, Guangliang Chen, and Mauro Maggioni. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32(3):435–462, 2012. (submitted:5/2011).

[3] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the Twenty-Third International Conference (ICML 2006)*, ACM International Conference Proceeding Series 148 ACM 2006, ISBN 1-59593-383-2, Pittsburgh, Pennsylvania, USA, June 25-29 2006.

[4] Jake Bouvrie and Mauro Maggioni. Efficient solution of markov decision problems with multiscale representations. In *Proc. 50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[5] Jake Bouvrie and Mauro Maggioni. Geometric multiscale reduction for autonomous and controlled nonlinear systems. In *IEEE Conference on Decision and Control (CDC)*, 2012.

[6] Jake Bouvrie and Mauro Maggioni. Multiscale markov decision problems: Compression, solution, and transfer learning. 2012.

[7] G. Chen, A.V. Little, M. Maggioni, and L. Rosasco. *Wavelets and Multiscale Analysis: Theory and Applications*. Springer Verlag, 2011. submitted March 12th, 2010.

[8] G. Chen and M. Maggioni. Multiscale geometric dictionaries for point-cloud data. In *Proc. SampTA*, 2011.

[9] G. Chen and M.Maggioni. Multiscale geometric wavelets for the analysis of point clouds. *Proc. CISS 2010*, 2010.

[10] Guangliang Chen, M. Iwen, Sang Chin, and M. Maggioni. A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements. In *Visual Communications and Image Processing (VCIP), 2012 IEEE*, pages 1–6, 2012.

[11] M.C. Crosskey and M. Maggioni. Learning of intrinsically low-dimensional stochastic systems in high-dimensions, i. Technical report, 2013. in preparation.

[12] Benjamin Van Dyke and Kevin R. Vixie. There are thin minimizers of the l1tv functional. *Abstract and Applied Analysis*, 2012(Article ID 930978), 2012. http://www.hindawi.com/journals/aaa/2012/930978/.

[13] Heather A. Van Dyke, Kevin R. Vixie, and Thomas J. Asaki. Cone monotonicity: Structure theorem, properties, and comparisons to other notions of monotonicity. *Abstract and Applied Analysis*, 2013(Article ID 134751), 2013. http://www.hindawi.com/journals/aaa/2013/134751/.

[14] S. Gerber and M. Maggioni. Multiscale dictionaries, transforms, and learning in high-dimensions. In *Proc. SPIE conference Optics and Photonics*, 2013.

[15] Sharif Ibrahim, Bala Krishnamoorthy, and Kevin R. Vixie. Simplicial flat norm with scale. *Journal of Computational Geometry*, 4(1):133–159, 2013. arxiv:1105.5104.

[16] Sharif Ibrahim, Bala Krishnamoorthy, and Kevin R. Vixie. Flat norm decomposition of integral currents. *arXiv*, 2014. http://arxiv.org/abs/1411.0882.

[17] Sharif Ibrahim, Kevin Sonnanburg, Thomas J. Asaki, and Kevin R. Vixie. Nonasymptotic densities for shape reconstruction. *Abstract and Applied Analysis*, 2014(Article ID 341910,), 2014. http://www.hindawi.com/journals/aaa/2012/930978/.

[18] Mark A. Iwen and Mauro Maggioni. Approximation of points on low-dimensional manifolds via random linear projections. *Inference & Information*, 2(1):1–31, 2013. arXiv:1204.3337v1, 2012.

[19] D.N. Kaslovsky. The deluge of images and videos: Understanding the manifold of image patches with randomized techniques. Colorado Photonics Industry Association Annual Meeting, Boulder, CO, 2010.

[20] D.N. Kaslovsky. Understanding the manifold of image patches with randomized techniques. NSF IGERT Project Meeting, Washington, DC, 2010.

[21] D.N. Kaslovsky. Geometric image processing: A local approach. SIAM Front Range Applied Mathematics Student Conference., 2011.

[22] D.N. Kaslovsky and F.G. Meyer. Image manifolds: Processing along the tangent plane. In *7th International Congress on Industrial and Applied Mathematics - ICIAM 2011*, 2011.

[23] D.N. Kaslovsky and F.G. Meyer. Overcoming noise, avoiding curvature: Optimal scale selection for tangent plane recovery. In *Proc. IEEE Workshop on Statistical Signal Processing*, pages 904–907, 2012. `http://dx.doi.org/10.1109/SSP.2012.6319851`.

[24] D.N. Kaslovsky and F.G. Meyer. Non-asymptotic analysis of tangent space perturbation. revised, re-submitted to *Information and Inference: A Journal of the IMA*, 53 pages, `http://arxiv.org/abs/1111.4601v4`, 2013.

[25] D.N. Kaslovsky and F.G. Meyer. Non-asymptotic analysis of tangent space perturbation. *Information and Inference: A Journal of the IMA*, 3(2):134–187, 2014.

[26] Anna V. Little, Mauro Maggioni, and Lorenzo Rosasco. Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. Technical report, MIT-CSAIL-TR-2012-029/CBCL-310, MIT, Cambridge, MA, September 2012.

[27] M. Maggioni, S. Minsker, and N. Strawn. Dictionary learning and non-asymptotic bounds for the Geometric Multi-Resolution Analysis. *arXiv*, 2014.

[28] Mauro Maggioni. Geometric estimation of probability measures in high dimensions. In *IEEE Asilomar Conference on Signals, Systems and Computers*, 2013.

[29] F.G. Meyer. Image de-noising on the manifold of patches: a spectral approach. (Invited Paper), SIAM Conference on Imaging Science (IS10), 2010.

[30] F.G. Meyer. Nonlinear dimensionality reduction: The inverse map. (Invited Speaker), 5th International Conference on Computational Harmonic Analysis, Vanderbilt University, 2014.

[31] Francois G. Meyer and Xilin Shen. Perturbation of the eigenvectors of the graph laplacian: Application to image denoising. *Applied and Computational Harmonic Analysis*, 36(2):326 – 334, 2014.

[32] Nathan D. Monnig, Bengt Fornberg, and Francois G. Meyer. Inverting nonlinear dimensionality reduction with scale-free radial basis function interpolation. *Applied and Computational Harmonic Analysis*, 37(1):162–170, 2014.

[33] N.D. Monnig, B. Fornberg, and F.G. Meyer. Nonlinear dimensionality reduction: The inverse map. Workshop on Electrical Flows, Graph Laplacians, and Algorithms: Spectral Graph Theory and Beyond, ICERM, Brown University, 2014.

[34] E.E. Monson, G. Chen, R. Brady, and M. Maggioni. Data representation and exploration with geometric wavelets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium*, pages 243–244, Dec 2010.

[35] Simon P. Morgan and Kevin R. Vixie. L^1TV computes the flat norm for boundaries. *Abstract and Applied Analysis*, 2007:Article ID 45153, 14 pages, 2007. doi:10.1155/2007/45153.

[36] K.M. Taylor, , M. Procopio, C. Young, and F.G. Meyer. Estimation of arrival times from seismic waves: a manifold-based approach. *Geophysical Journal International*, 185(1):435–452, 2011. http://dx.doi.org/10.1111/j.1365-246X.2011.04947.x.

[37] Kevin R. Vixie, Keith Clawson, Thomas J. Asaki, Gary Sandine, Simon P. Morgan, and Brandon Price. Multiscale flat norm signatures for shapes and images. *Applied Mathematical Sciences*, 4(14):667–680, 2010.