Multi-Resolution Geometric Analysis for Data in High Dimensions

Guangliang Chen, Anna V. Little, and Mauro Maggioni

Abstract Large data sets arise in a wide variety of applications and are often modeled as samples from a probability distribution in high-dimensional space. It is sometimes assumed that the support of such probability distribution is well approximated by a set of low intrinsic dimension, perhaps even a low-dimensional smooth manifold. Samples are often corrupted by high-dimensional noise. We are interested in developing tools for studying the geometry of such high-dimensional data sets. In particular, we present here a multiscale transform that maps high-dimensional data as above to a set of multiscale coefficients that are compressible/sparse under suitable assumptions on the data. We think of this as a geometric counterpart to multi-resolution analysis in wavelet theory: whereas wavelets map a signal (typically low dimensional, such as a one-dimensional time series or a two-dimensional image) to a set of multiscale coefficients, the geometric wavelets discussed here map points in a high-dimensional point cloud to a multiscale set of coefficients. The geometric multi-resolution analysis (GMRA) we construct depends on the support of the probability distribution, and in this sense it fits with the paradigm of dictionary learning or data-adaptive representations, albeit the type of representation we construct is in fact mildly nonlinear, as opposed to standard linear representations. Finally, we apply the transform to a set of synthetic and real-world data sets.

Keywords Multiscale analysis • Geometric analysis • High-dimensional data • Covariance matrix estimation

G. Chen • A.V. Little • M. Maggioni (✉)
Mathematics and Computer Science Departments, Duke University,
P.O. Box 90320, Durham, NC 27708, USA
e-mail: mauro.maggioni@duke.edu

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 1,*
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8376-4_13,

1 Introduction

We are interested in developing tools for harmonic analysis and processing of large data set that arise in wide variety of applications, such as sounds, images (RGB or hyperspectral, [16]), gene arrays, EEG signals [9], and manifold-valued data [44], to name a few. These data sets are often modeled as samples from a probability distribution in $\mathbb{R}^D$, but it is sometimes assumed that the support of such probability distribution is in fact a set of low intrinsic dimension, perhaps with some nice geometric properties, for example, those of a smooth manifold.

Approximating and learning functions in high-dimensional spaces is hard because of the curse of high dimensionality, it is natural to try to exploit the intrinsic low dimensionality of the data: this idea has attracted wide interest across different scientific disciplines and various applications. One example of exploitation of low intrinsic dimension is to map the data to low-dimensional space, while preserving salient properties of data [3, 19, 21, 27, 28, 30, 31, 46, 52, 54]. Another example is the construction of dictionaries of functions supported on the data [7, 17, 18, 38–40, 49, 50]. Yet another possibility is modeling the data as a union of low-dimensional subspaces, which is related to the ideas of sparse representations and dictionary learning ([1, 2, 10, 11, 51] and references therein).

When performing dimensionality reduction/manifold learning, the objective is mapping data to a low-dimensional space. The maps used are often nonlinear, and in at least two problems arise: that of extending the map from a training data set to new data points and that of inverting such a map, i.e., going from a low-dimensional representation of a data point back to its higher-dimensional original representation. Both problems seem to be rather hard (depending of course of the map used) and to require some form of high-dimensional interpolation/extrapolation.

We will work directly in the high-dimensional space, but by taking advantage of the assumed low intrinsic dimensionality of the data and its geometry. One advantage of this approach is that while our representations will be low-dimensional, we will not have to produce inverse maps from low dimensions to high dimensions. We construct geometric multi-resolution analysis (GMRA) for analyzing intrinsically low-dimensional data in high-dimensional spaces, modeled as samples from a d-dimensional set $\mathscr{M}$ (in particular, a manifold) embedded in $\mathbb{R}^D$, in the regime $d \ll D$. Data may be sampled from a class of signals of interest; in harmonic analysis, a linear infinite-dimensional function space $\mathscr{F}$ often models the class of signals of interest, and linear representations in the form $f = \sum_i \alpha_i \phi_i$, for $f \in \mathscr{F}$ in terms of a dictionary of atoms $\Phi := \{\phi_i\} \subseteq \mathscr{F}$ are studied. Such dictionaries may be bases or frames and are constructed so that the sequence of coefficients $\{\alpha_i\}_i$ has desirable properties, such as some form of sparsity, or a distribution highly concentrated at zero. Several such dictionaries have been constructed for function classes modeling one- and two-dimensional signals of interest [8, 12, 14, 20, 22, 47] and are proven to provide optimal representations (in a suitably defined sense) for certain function spaces and/or for operators on such spaces. A more recent trend [1, 12, 41–43, 51, 55], motivated by the desire to model classes of signals that are not well modeled by the linear structure of function spaces, has been that of *constructing data-adapted dictionaries*: an algorithm is allowed to see samples from a class of signals $\mathscr{F}$ (not necessarily a linear function space) and constructs a dictionary $\boldsymbol{\Phi} := \{\phi_i\}_i$ that optimizes some functional, such as the sparsity of the coefficients for signals in $\mathscr{F}$.

There are several parameters in this problem: given training data from $\mathscr{F}$, one seeks $\boldsymbol{\Phi}$ with I elements, such that every element in the training set may be represented, up to a certain precision ε, by at most m elements of the dictionary. The smaller I and m are, for a given ε, the better the dictionary.

Several current approaches may be summarized as follows [42]: consider a finite training set of signals $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, which we may represent by a $\mathbb{R}^{D\times n}$ matrix, and optimize the cost function

$$f_n(\boldsymbol{\Phi}) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, \boldsymbol{\Phi}),$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{D\times I}$ is the dictionary, and ℓ a loss function, for example,

$$\ell(x, \boldsymbol{\Phi}) := \min_{\alpha\in\mathbb{R}^I} \frac{1}{2}||x - \boldsymbol{\Phi}\alpha||_{\mathbb{R}^D}^2 + \lambda||\alpha||_1,$$

where λ is a regularization parameter. This is basis pursuit [12] or lasso [53]. One typically adds constraints on the size of the columns of $\boldsymbol{\Phi}$, for example, $||\phi_i||_{\mathbb{R}^D} \leq 1$ for all i, which we can write as $\boldsymbol{\Phi} \in \mathscr{C}$ for some convex set $\mathscr{C}$. The overall problem may then be written as a matrix factorization problem with a sparsity penalty:

$$\min_{\boldsymbol{\Phi}\in\mathscr{C}, \alpha\in\mathbb{R}^{I\times n}} \frac{1}{2}||X_n - \boldsymbol{\Phi}\alpha||_F^2 + \lambda||\alpha||_{1,1},$$

where $||\alpha||_{1,1} := \sum_{i_1,i_2}|\alpha_{i_1,i_2}|$. We refer the reader to [42] and references therein for techniques for attacking this optimization problem.

In this chapter we make additional assumptions on the data, specifically that it is well approximated by a smooth low-dimensional manifold, and we exploit this geometric assumption to construct data-dependent dictionaries. We use a multiscale approach that will lead to a GMRA of the data: this is inspired not only by quantitative geometric analysis techniques in geometric measure theory (see, e.g., [25, 32]) but also from multiscale approximation of functions in high dimensions [5, 6]. These dictionaries are structured in a multiscale fashion and, under suitable assumptions on the data, are computed efficiently; the expansion of a data point on the dictionary elements is guaranteed to have a certain degree of sparsity, m, and may also be computed by fast algorithms; the growth of the number of dictionary elements I as a function of ε is controlled depending on geometric properties of the data. This may be thought of as a wavelet analysis for data sets rather than for functions, where the geometry of a set of points is approximated, rather than a single function.

2 Geometric Multi-resolution Analysis

Let μ be a probability measure in $\mathbb{R}^D$ and $\mathscr{M}$ its support. In this chapter we will consider the case in which $\mathscr{M}$ is endowed with the structure of a Riemannian manifold, but the examples will show that the construction is robust enough to extend and be useful when this assumption is severely violated. In this setting we have a Riemannian metric g and a volume measure $d$vol. The geodesic distance on $\mathscr{M}$ associated with g will be denoted by ρ. We shall assume that $d\mu$ is absolutely continuous with respect to $d$vol, with $d\mu/d$vol bounded above and below. We are interested in the case when the "dimension" d of $\mathscr{M}$ is much smaller than the dimension of the ambient space $\mathbb{R}^D$. While d is typically unknown in practice, efficient (multiscale, geometric) algorithms for its estimation are available (see [37], which also contains many references to previous work on this problem), under additional assumptions on the geometry of $\mathscr{M}$.

2.1 *Dyadic Cubes*

We start by constructing *dyadic cubes* on $\mathscr{M}$. This may be thought of as an analogue of dyadic cubes in Euclidean space. It is a collection of (measurable) subsets $\{Q_{j,k}\}_{k\in\mathscr{K}_j, j\geq J_0}$ of $\mathscr{M}$ with the following properties [13, 23, 24]:

- For every $j \in \mathbb{Z}$, $\mu(\mathscr{M}\setminus\cup_{k\in\mathscr{K}_j}Q_{j,k}) = 0$.
- For $j' \geq j$ and $k' \in \mathscr{K}_{j'}$, either $Q_{j',k'} \subseteq Q_{j,k}$ or $\mu(Q_{j',k'}\cap Q_{j,k}) = 0$.
- For $j < j'$ and $k' \in \mathscr{K}_{j'}$, there exists a unique $k \in \mathscr{K}_j$ such that $Q_{j',k'} \subseteq Q_{j,k}$.
- Each $Q_{j,k}$ contains a point $c_{j,k}$ such that $B^{\mathscr{M}}_{c_1\cdot 2^{-j}}(c_{j,k}) \subseteq Q_{j,k} \subseteq B^{\mathscr{M}}_{2^{-j}}(c_{j,k})$, for a constant c_1 depending on intrinsic geometric properties of $\mathscr{M}$. Here $B^{\mathscr{M}}_r(x)$ is the ρ-ball inside $\mathscr{M}$ of radius $r > 0$ centered at $x \in \mathscr{M}$. In particular, we have $\mu(Q_{j,k}) \sim 2^{-dj}$.

Let $\mathscr{T}$ be the tree structure associated to the decomposition above: for any $j \in \mathbb{Z}$ and $k \in \mathscr{K}_j$, we let $\mathrm{ch}(j,k) = \left\{k' \in \mathscr{K}_{j+1} : Q_{j+1,k'} \subseteq Q_{j,k}\right\}$. We use the notation (j,x) to represent the unique $(j,k(x)), k(x) \in \mathscr{K}_j$ such that $x \in Q_{j,k(x)}$.

2.2 *Multiscale SVD and Intrinsic Dimension Estimation*

An introduction to the use of the ideas we present for the estimation of intrinsic dimension of point clouds is in [37] and references therein (see [35, 36] for previous short accounts). These types of constructions are motivated by ideas in both multiscale geometric measure theory [24, 26, 33] and adaptive approximation of functions in high dimensions[5, 6].

In each dyadic cell $Q_{j,k}$ we consider the mean

$$m_{j,k} := \mathbb{E}_\mu[x|x \in Q_{j,x}] = \frac{1}{\mu(Q_{j,k})} \int_{Q_{j,k}} x \, \mathrm{d}\mu(x) \in \mathbb{R}^D \tag{1}$$

and the local covariance

$$\mathrm{cov}_{j,k} = \mathbb{E}_\mu[(x - m_{j,k})(x - m_{j,k})^* | x \in Q_{j,k}] \in \mathbb{R}^{D \times D}, \tag{2}$$

where vectors in $\mathbb{R}^D$ are considered d-dimensional column vectors. Let the rank-d singular value decomposition (SVD) [29] of $\mathrm{cov}_{j,k}$ be

$$\mathrm{cov}_{j,k} \approx \boldsymbol{\Phi}_{j,k} \boldsymbol{\Sigma}_{j,k} \boldsymbol{\Phi}^*_{j,k}, \tag{3}$$

where $\boldsymbol{\Phi}_{j,k}$ is an orthonormal $D \times d$ matrix and $\boldsymbol{\Sigma}$ is a diagonal $d \times d$ matrix. Let

$$\mathbb{V}_{j,k} := V_{j,k} + m_{j,k}, \quad V_{j,k} = \langle \boldsymbol{\Phi}_{j,k} \rangle, \tag{4}$$

where $\langle A \rangle$ denotes the span of the columns of A, so that $\mathbb{V}_{j,k}$ is the affine subspace of dimension d parallel to $V_{j,k}$ and passing through $m_{j,k}$. It is an approximate tangent space to $\mathscr{M}$ at location $m_{j,k}$ and scale 2^{-j}; in fact by the properties of the SVD it provides the best $d_{j,k}$-dimensional planar approximation to $\mathscr{M}$ in the least squares sense:

$$\mathbb{V}_{j,k} = \underset{\Pi}{\mathrm{argmin}} \int_{Q_{j,k}} ||x - \mathbb{P}_\Pi(x)||^2 \, \mathrm{d}\mu(x), \tag{5}$$

where Π is taken on the set of all affine $d_{j,k}$-planes and $\mathbb{P}_\Pi$ is the orthogonal projection onto the affine plane Π. Let $\mathbb{P}_{j,k}$ be the associated affine projection

$$\mathbb{P}_{j,k}(x) := \mathbb{P}_{\mathbb{V}_{j,k}}(x) = \boldsymbol{\Phi}_{j,k} \boldsymbol{\Phi}^*_{j,k}(x - m_{j,k}) + m_{j,k}, \quad x \in Q_{j,k}. \tag{6}$$

The behavior of the singular values in the matrix $\Sigma_{j,x}$ in Eq. (3) as a function of the scale j, for x fixed, contains a lot of useful information about the geometry of the data around x. In particular they may be used to detect the intrinsic dimension of the data in a neighborhood of x. We need to introduce several definition before stating some results. Because of space constraints, we will consider here the case when $\mathscr{M}$ is a manifold of co-dimension one, leaving the discussion of the general case ($\mathscr{M}$ with arbitrary co-dimension and $\mathscr{M}$ not a manifold) to [2, 37]. Let

$$\lambda = \frac{d}{d+2}, \qquad \kappa = \frac{d}{(d+2)^2(d+4)} \left[\frac{d+1}{2} \sum_{i=1}^{d} \kappa_i^2 - \sum_{i<j} \kappa_i \kappa_j \right],$$

where κ_i's are the sectional curvatures of the manifold. We refer the reader to [37] for an extended discussion of these quantities, which arise naturally in the study of multiscale SVD of manifolds. When $\mathscr{M}$ has co-dimension larger than 1 more complicate functions of the curvatures arise [similar to those in Eq. (18)]; in the non-manifold case a notion of L^2 that generalizes the above may be used [37]. Suppose we sample n points $x_1,\dots,x_n$ i.i.d. from the volume measure on the manifold, and each is perturbed by i.i.d. realizations of white Gaussian noise in $\mathbb{R}^D$ with variance $\sigma^2 I_D$. We denote by $\tilde{X}_{n,\tilde{z},r}$ the set of noisy samples $x_i+\eta_i$ that are in $B_{z+\eta_z}(r)$, where η_z is the noise corresponding to the data point z: this is the data being observed, which is sampled and noisy, at disposal of an algorithm. We denote by $X_{z,r}$ a random variable distributed in $B_z(r)\cap\mathscr{M}$ according to volume measure: this is the ideal data, uncorrupted by noise and sampling. Finally, we let $r_{=}^2 := r^2 - 2\sigma^2 D$.

Let $r = 2^{-j}$ and $X_{z,r} = \mathscr{M}\cap B_z(r)$. The behavior of the ideal covariance of $X_{z,r}$ (which is comparable to $\mathrm{cov}_{j,k}$) as a function of r reveals interesting properties of the data, for example, it may be used to measure intrinsic dimension and L^2-curvature of $\mathscr{M}$ around a point z, since the d largest singular values will grow quadratically in r, and the remaining ones will measure L^2-curvatures. In particular for r small the largest gap between these singular values will be the dth gap, leading to an estimator of intrinsic dimension. However, since we do not have access to $X_{z,r}$, we are interested in the behavior of the empirical covariance matrix of the noisy samples $\tilde{X}_{n,\tilde{z},r}$ as a function of r. In particular, we ask how close it is to $\mathrm{cov}(X_{z,r})$ and when is the dth gap of $\mathrm{cov}(\tilde{X}_{n,\tilde{z},r})$ the largest, so that we may use it to estimate the intrinsic dimension of $\mathscr{M}$? Observe that while we would like to choose r small, since then the difference in the behavior of the top d singular values and the remaining ones is largest, we are not allowed to do that anymore: having only n samples forces a lower bound on r, since in small balls we will have too small a number of samples to estimate the covariances. Moreover, the presence of noise also puts a lower bound on the interesting range of r: since the expected length of a noise vector is $\sigma\sqrt{D}$, and the covariance of the noise has norm σ, we expect that r should be larger than a function of these quantities in order for $\mathrm{cov}(\tilde{X}_{n,\tilde{z},r})$ to provide meaningful information about the geometric of $\mathscr{M}$.

Here and in what follows C, C_1, and C_2 will denote numerical constants whose value may change with each occurrence.

Theorem 1 **($n\to\infty$).** *Fix $z\in\mathscr{M}$; assume $D\geq C$, $\sigma\sqrt{D}\leq\frac{\sqrt{d}}{2\sqrt{2}\kappa}$,*

$$r\in\left(R_{\min}+4\sigma\sqrt{D}+\frac{1}{6\kappa}, R_{\max}-\sigma\sqrt{D}-\frac{1}{6\kappa}\right)\cap\left(3\sigma\left(\sqrt{D}\vee d\right),\frac{\sqrt{d}}{\kappa}\right). \tag{7}$$

Then for n large enough, with probability at least $1-Ce^{-C\sqrt{D}}$, we have

$$||\mathrm{cov}(\tilde{X}_{n,\tilde{z},r})-\mathrm{cov}(X_{z,r_{=}})||\leq C\left(\frac{\kappa^2 r_{=}^4}{d}+\sigma^2+\frac{\lambda\kappa r_{=}^3}{d}\left(\frac{\lambda\kappa r_{=}}{\lambda^2-C\kappa^2 r_{=}^2}\wedge 1\right)\right). \tag{8}$$

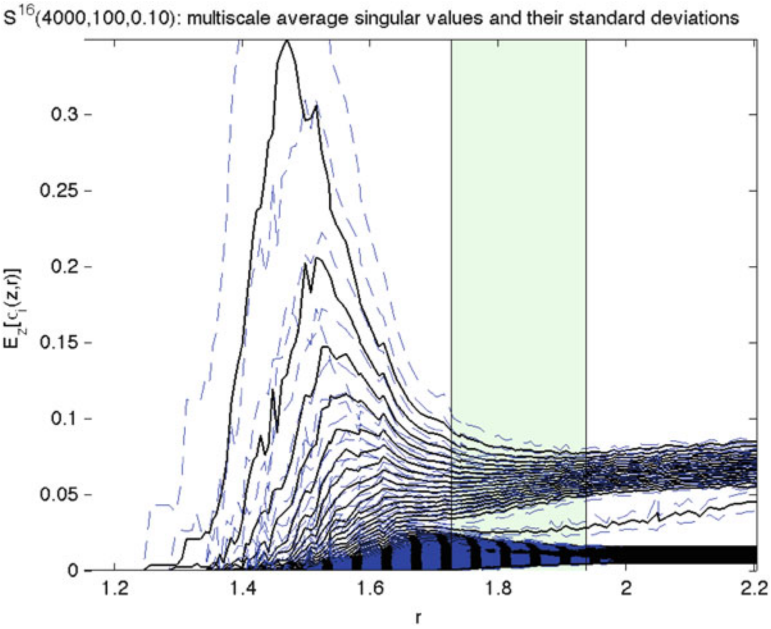

Fig. 1 We consider $4,000$ points uniformly sampled on a 16-dimensional unit sphere, embedded in $\mathbb{R}^{100}$, with $\eta \sim 0.1\mathscr{N}(0, I_{100})$ Gaussian noise added to each point. We plot empirical mean (over data points $\tilde{z}$) of the squared singular values of the empirical covariance matrix $\text{cov}(\tilde{X}_{n,\tilde{z},r})$, as a function of r: in the "reasonable" range of scales, above the size of the noise, we see 16 singular values corresponding to the approximate tangent planes, at 17th squared singular value corresponding to curvature, and all the other squared singular values of size comparable to the energy of the noise 10^{-2}. The algorithm detects a range of scales, above the scale of the noise, where the 16th gap between the squared singular values is largest, i.e., noise is small compared to curvature, which is in turn small compared to elongation along the tangent plane. It is remarkable, albeit predicted by our results, that only $4,000$ points (typically considered a small number if 16 (even more in 100) dimensions), perturbed by large noise (note that $\text{e}[||\eta||] \sim 1$), are enough to obtain accurate geometric information

Moreover, in the range of scales

$$C_1 \frac{\sigma\sqrt{d}}{\sqrt{\lambda^2 - \delta^2}} \leq r_= \leq C_2 \frac{\lambda^2 - \delta^2}{\lambda\kappa}, \tag{9}$$

$\Delta_k(\tilde{X}_{n,\tilde{z},r})$ *is the largest gap, with the probability as above.*

Theorem 1 essentially says that if we have $O(d \log d)$ points in $\mu(B_z(r))$, and the noise variance σ is not too large compared to curvature, then the largest gap in the empirical covariance matrix of the data in $B_z(r)$ is the dth gap, with high probability, for r in the range:

$$C_1\sigma^2 \leq \frac{r^2}{d} \leq C_2\ \frac{\lambda^2}{\kappa^2 d}.$$

The upper bound $\frac{\lambda}{\kappa}$ is dictated by curvature, while the lower bound $\sigma\sqrt{D}$ is forced by the noise level: the lower bound is comparable to the size of the covariance of the noise, the upper bound is comparable to the size of the covariance of the data computed at the largest radius λ/κ where the curvature is not too large, and the term in the middle is comparable to the size of the data along the local approximating plane.

Our second theorem explores the regime where the ambient dimension D goes to infinity, but the number of samples n is fixed, dependent on the intrinsic dimension. While of course n samples certainly lie in an n-dimensional affine subspace, because of the ambient noise such subspace is unreliable at small scales, and this regime captures the range of scales where we have independence from the ambient dimension and the essentially linear dependence on d for the minimal needed number of points in $B_z(r_=)$.

Theorem 2 **($D \to \infty$, $\sigma\sqrt{D} = O(1)$).** *Fix $z \in \mathcal{M}$. Let the assumptions of Theorem 1 and the restriction (7) hold. Fix $t \in (C, Cd)$ and assume $\varepsilon := \varepsilon_{r_=,n,t} \leq \frac{1}{2}$. Then for $D \geq C$ and $m \leq D$, and σ_0 constant, for r in the range of scales (7) intersected with*

$$r \in \left(\frac{4\sigma_0 \left(1 \vee \frac{d}{\sqrt{D}} \vee \lambda_{\max}\varepsilon\right)}{\lambda_{\min}^2 - \delta^2 \lambda_{\max}\varepsilon - \frac{\varepsilon^2}{\lambda_{\min}^2}\left(\frac{C\sigma_0 d}{r} \vee \frac{1}{m}\right) - \frac{\sigma_0 \kappa}{t}}, \frac{\frac{\lambda_{\max}}{4} \wedge \sqrt{d}}{\kappa} \right),$$

the following hold, with probability at least $1 - Ce^{-Ct^2}$:

(i) $\Delta_k(\mathrm{cov}(\tilde{X}_{n,\tilde{z},r}))$ *is the largest gap of* $\mathrm{cov}(\tilde{X}_{n,\tilde{z},r})$.

(ii) $||\mathrm{cov}(\tilde{X}_{n,\tilde{z},r}) - \mathrm{cov}(X_{z,r_=}) - \sigma^2 I_D|| \leq \left(\sigma_0^2\varepsilon + \lambda_{\max}\sigma_0 r + \left(\lambda_{\max} + 2\sigma_0\kappa + \frac{\varepsilon}{m}\right) r^2 + O\left(\frac{r^3}{\varepsilon}\right)\right)\frac{\varepsilon}{d}$.

These bounds, and in fact the finer bounds of [37], may of course be trivially used to obtain perturbation bounds for the empirical noisy tangent spaces estimated by looking at the top d singular vectors of the empirical covariance matrix of the data in $B_{\tilde{z}}(r)$ (a slightly better approach, taken in [37], uses Wieland's lemma before applying the usual sine theorems). It turns out that, since the noise essentially "to first order" adds only a multiple to the identity matrix, the approximate tangent space computed in this fashion is very stable, even in the regime of Theorem 2 [37].

This is the subject of [37], where it is shown that under rather general conditions on the geometry of the data (much more general than the manifold case) and under sampling and ambient noise, one may use these multiscale singular values to estimate the intrinsic dimension of the data. Moreover, under suitable assumptions, the number of samples in a ball around x required in order to do so is linear in the intrinsic dimension and independent of the ambient dimension. We refer the reader to [10, 35–37]. We now proceed by using not only the information in the singular values but also in the singular vectors in the SVD decomposition in Eq. (3).

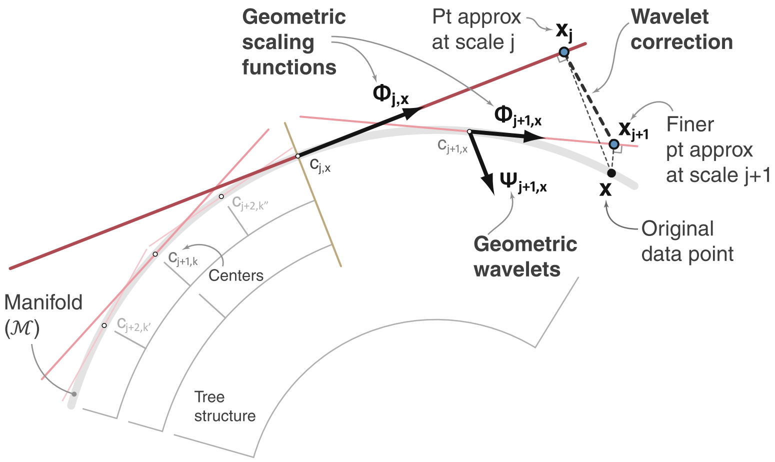

Fig. 2 An illustration of the geometric wavelet decomposition. The centers $m_{j,x}$'s are represented as lying on $\mathscr{M}$ while in fact they are only close to $\mathscr{M}$, and the corresponding planes $\mathbb{V}_{j,x}$ are represented as tangent planes, albeit they are only an approximation to them. Art courtesy of E. Monson

2.3 Geometric Scaling Functions

Then $\mathbb{P}_{j,k}(Q_{j,k})$ is the projection of $Q_{j,k}$ onto the local linear approximation given by the affine subspace in Eq. (4). The fact that this linear subspaces are affine will have various implications in our construction, creating mild nonlinearities and forcing us to construct a different transform and data representation which is not simply in the form of linear combination of certain atoms. On the other hand it seems an extremely natural construction, and not only the nonlinearities involved will not cause conceptual or computational overheads, but in fact we shall obtain algorithms which are faster than those needed to compute sparse linear representations in the standard dictionary learning setting. $\{\Phi_{j,k}\}_{k\in\mathscr{K}_j}$ are the geometric analogue of a family of scaling functions at scale j, and therefore we call them *geometric scaling functions*. They "span" an approximate piecewise linear manifold at scale j

$$\mathscr{M}_j := \{\mathbb{P}_{j,k}(Q_{j,k})\}_{k\in\mathscr{K}_j} \tag{10}$$

Under general conditions, $\mathscr{M}_j \to \mathscr{M}$ in the Hausdorff distance, as $j \to +\infty$. It is natural to define the nonlinear projection of $\mathscr{M}$ onto $\mathscr{M}_j$ by

$$x_j \equiv P_{\mathscr{M}_j}(x) := \mathbb{P}_{j,k}(x), \qquad x \in Q_{j,k}. \tag{11}$$

Note that in general $\mathscr{M}_j$ is not contained in $\mathscr{M}_{j+1}$, due to the nonlinearity of the underlying manifold $\mathscr{M}$. This is important as we move into the next section when we will encode "the difference" between $\mathscr{M}_j$ and $\mathscr{M}_{j+1}$.

2.4 Geometric Wavelets

In wavelet analysis, wavelets span the difference between scaling function spaces and are contained in the finer scale scaling function space. In our setting that would correspond to encoding the difference needed to go from $\mathscr{M}_j$ to $\mathscr{M}_{j+1}$: for a fixed $x \in \mathscr{M}$, $x_{j+1} - x_j \in \mathbb{R}^D$, but in general not contained in $\mathscr{M}_{j+1}$, due to the nonlinearity of $\mathscr{M}_j$ and $\mathscr{M}_{j+1}$. The main observation is that nevertheless the collection of vectors $x_{j+1} - x_j$ for x varying in $Q_{j+1,x}$ is in fact contained in a low-dimensional subspace and may be therefore encoded efficiently in terms of a basis of that subspace. We proceed as follows: for $j \leq J-1$ we let

$$\begin{aligned} Q_{\mathscr{M}_{j+1}}(x) :&= x_{j+1} - x_j = x_{j+1} - \mathbb{P}_{j,x}(x_{j+1}) + \mathbb{P}_{j,x}(x_{j+1}) - \mathbb{P}_{j,x}(x) \\ &= (I - P_{j,x})(x_{j+1} - c_{j,x}) + P_{j,x}(x_{j+1} - x) \\ &= (I - P_{j,x})(\underbrace{x_{j+1} - c_{j+1,x}}_{\in V_{j+1,x}} + c_{j+1,x} - c_{j,x}) - P_{j,x}(x - x_{j+1}). \end{aligned} \tag{12}$$

Let $W_{j+1,x} := (I - P_{j,x})V_{j+1,x}$, $Q_{j+1,x}$ be the orthogonal projection onto $W_{j+1,x}$, and let $\Psi_{j+1,x}$ be an orthonormal basis for $W_{j+1,x}$, which we will call a *geometric wavelet basis*. Observe $\dim W_{j+1,x} \leq \dim V_{j+1,x} = d_{j+1,x}$. We define several quantities below:

$$\begin{aligned} t_{j+1,x} :&= c_{j+1,x} - c_{j,x}, w_{j+1,x} := (I - P_{j,x})t_{j+1,x}; \\ &\mathbb{Q}_{j+1,x}(x) := Q_{j+1,x}(x - c_{j+1,x}) + w_{j+1,x}. \end{aligned}$$

Then we may rewrite Eq. (12) as

$$\begin{aligned} Q_{\mathscr{M}_{j+1}}(x) &= \underbrace{Q_{j+1,x}(x_{j+1} - c_{j+1,x})}_{\in W_{j+1,x}} + w_{j+1,x} - P_{j,x}\left(x - x_J + \sum_{l=j+1}^{J-1}(x_{l+1} - x_l)\right) \\ &= \mathbb{Q}_{j+1,x}(x_{j+1}) - P_{j,x}\sum_{l=j+1}^{J-1}(x_{l+1} - x_l) - P_{j,x}(x - x_J) \\ &= \mathbb{Q}_{j+1,x}(x_{j+1}) - P_{j,x}\sum_{l=j+1}^{J-1}Q_{\mathscr{M}_{l+1}}(x) - P_{j,x}(x - x_J), \end{aligned} \tag{13}$$

where $J \geq j+1$ is the index of the finest scale (and the last term vanishes as $J \to +\infty$, under general conditions). Note that this multiscale expansion contains terms that involve not only the current scale $j+1$ and the previous scale j but terms from finer scales as well, all the way to the finest scale J. This is once again due to the nonlinearity of $\mathscr{M}$ and of the whole construction: knowing $P_{\mathscr{M}_{j+1}}(x)$ is not enough to construct $P_{\mathscr{M}_j}(x)$, since the whole local nonlinear structure of $\mathscr{M}$ determines the

	J	$J-1$	$J-2$	$\dots$	$j+2$	$j+1$	j	$\dots$
$P_{\mathcal{M}_{J-1}}(x)$	$\mathbb{Q}_{J,x}(x_J)$	$P_{\mathcal{M}_{J-1}}(x)$						
$P_{\mathcal{M}_{J-2}}(x)$	$P_{J-2}(x-x_J)$	$\mathbb{Q}_{J-1,x}(x_{J-1})$	$P_{\mathcal{M}_{J-2}}(x)$					
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
$P_{\mathcal{M}_{j+1}}(x)$	$P_j(x-x_J)$	$P_jQ_{\mathcal{M}_J}(x)$	$P_jQ_{\mathcal{M}_{J+1}}(x)$	$\cdots$	$P_jQ_{\mathcal{M}_{j+2}}(x)$	$\mathbb{Q}_{j+1,x}(x_{j+1})$	$P_{\mathcal{M}_j}(x)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Fig. 3 We represent in this table the triangular array summarizing the geometric wavelet expansion of a term in the first column in terms of geometric wavelets, according to the multiscale relations (15) and the equalities in Eq. (13)

locally optimal projection $P_{\mathcal{M}_j}(x)$. In [2] we describe a variation of the transform where optimality is relaxed and a "two-scale equation" is obtained.

In terms of the geometric scaling functions and wavelets, the above may be written as

$$x_{j+1} - x_j = \Psi_{j+1,x}\Psi^*_{j+1,x}\left(x_{j+1} - m_{j+1,x}\right) + w_{j+1,x} - \Phi_{j,x}\Phi^*_{j,x}\sum_{l=j+1}^{J-1} Q_{\mathcal{M}_{l+1}}(x)$$
$$- \Phi_{j,x}\Phi^*_{j,x}\left(x - x_J\right). \tag{14}$$

This shows that the difference $x_{j+1} - x_j$ can be expressed as the sum of a component in $W_{j+1,x}$, a second component that only depends on the cell $(j+1,x)$ (but not on the point x itself) which accounts for the translation of centers and lying in $V^{\perp}_{j,x}$ (but not necessarily in $W_{j+1,x}$), and a sum of projections on $V_{j,x}$ of differences $x_{l+1} - x_l$ at finer scales. By construction we have the two-scale equation

$$P_{\mathcal{M}_{j+1}}(x) = P_{\mathcal{M}_j}(x) + Q_{\mathcal{M}_{j+1}}(x), \quad x \in \mathcal{M} \tag{15}$$

which can be iterated across scales, leading to a multiscale decomposition along low-dimensional subspaces, with efficient encoding and algorithms. We think of $P_{j,k}$ as being attached to the node (j,k) of $\mathcal{T}$ and the $Q_{j+1,k'}$ as being attached to the edge connecting the node $(j+1,k')$ to its parent.

We say that the set of multiscale piecewise affine operators $\{P_{\mathcal{M}_j}\}$ and $\{Q_{\mathcal{M}_{j+1}}\}$ form a **geometric multi-resolution analysis** or GMRA for short.

2.5 Approximation for Manifolds

We analyze the error of approximation to a d-dimensional manifold in $\mathbb{R}^D$ by using geometric wavelets representation. Our analysis gives a full explanation of the examples in Sect. 4.1. We have the following theorem from [2]:

Theorem 3. *Let $(\mathscr{M},\rho,\mu)$ be a compact $\mathscr{C}^{1+\alpha}$ Riemannian manifold of dimension d isometrically embedded in $\mathbb{R}^D$, with $\alpha\in(0,1]$, and μ absolutely continuous with respect to the volume measure on $\mathscr{M}$. Let $\{P_{\mathscr{M}_j},Q_{\mathscr{M}_{j+1}}\}$ be a GMRA for $(\mathscr{M},\rho,\mu)$. For any $x\in\mathscr{M}$, there exists a scale $j_0=j_0(x)$ such that for any $j\geq j_0$ and any $p>0$, if we let $d\mu_{j,x}:=\mu(Q_{j,x})^{-1}d\mu$,*

$$\left|\left|\left|\left|z-P_{\mathscr{M}_j}(z)\right|\right|_{\mathbb{R}^D}\right|\right|_{L^p(Q_{j,x},d\mu_{j,x}(z))}=\left|\left|\left|\left|z-P_{\mathscr{M}_{j_0}}(z)-\sum_{l=j_0}^{j-1}Q_{\mathscr{M}_{l+1}}(z)\right|\right|_{\mathbb{R}^D}\right|\right|_{L^p(Q_{j,x},d\mu_{j,x}(z))}$$
$$\leq||\kappa||_{L^\infty(Q_{j,x})}2^{-(1+\alpha)j}+o(2^{-(1+\alpha)j}). \tag{16}$$

If $\alpha<1$, $\kappa(x)$ depends on the $\mathscr{C}^{1+\alpha}$ norm of a coordinate chart from $T_x(\mathscr{M})$ to $Q_{j,x}\subseteq\mathscr{M}$ and on $\left|\left|\frac{d\mu}{d\mathrm{vol}}\right|\right|_{L^\infty(Q_{j,x})}$.

If $\alpha=1$,

$$\kappa(x)=\left|\left|\frac{d\mu}{d\mathrm{vol}}\right|\right|_{L^\infty(Q_{j,x})}\min(\kappa_1(x),\kappa_2(x)), \tag{17}$$

with

$$\kappa_1(x):=\frac{1}{2}\max_{i\in\{1,\dots,D-d\}}||H_i(x)||;$$
$$\kappa_2^2(x):=\max_{w\in\mathbb{S}^{D-d}}\frac{d(d+1)}{4(d+2)(d+4)}\left[\left|\left|\sum_{l=1}^{D-d}w_lH_l(x)\right|\right|_F^2-\frac{1}{d+2}\left(\sum_{l=1}^{D-d}w_l\mathrm{Tr}(H_l(x))\right)^2\right], \tag{18}$$

and the $D-d$ matrices $H_l(x)$ are the d-dimensional Hessians of $\mathscr{M}$ at x.

Observe that κ_2 can be smaller than κ_1 (by a constant factor) or larger (by factors depending on d^2), depending on the spectral properties and commutativity relations between the Hessians H_l. κ_2^2 may be unexpectedly small, in the sense that it may scale as $d^{-2}r^4$ as a function of d and r, as observed in [37]. For the proof we refer the reader to [2].

3 Algorithms

We present in this section algorithms implementing the construction of the GMRA and the corresponding geometric wavelet transform (GWT).

3.1 Construction of Geometric Multi-resolution Analysis

The first step in the construction of the geometric wavelets is to perform a geometric nested partition of the data set, forming a tree structure. For this end, one may consider various methods listed below:

- Use of METIS [34]: a multiscale variation of iterative spectral partitioning. We construct a weighted graph as done for the construction of diffusion maps [15, 21]: we add an edge between each data point and its k nearest neighbors and assign to any such edge between x_i and x_j the weight $e^{-||x_i-x_j||^2/\sigma}$. Here k and σ are parameters whose selection we do not discuss here (but see [45] for a discussion in the context of molecular dynamics data). In practice, we choose k between 10 and 50 and choose σ adaptively at each point x_i as the distance between x_i and its $\lfloor k/2 \rfloor$ nearest neighbor.
- Use of cover trees [4].
- Use of iterated PCA: at scale 1, compute the top d principal components of data and partition the data based on the sign of the $(d+1)$-st singular vector. Repeat on each of the two partitions.
- Iterated k-means: at scale 1 partition the data based on k-means clustering, then iterate on each of the elements of the partition.

Each construction has pros and cons, in terms of performance and guarantees. For (I) we refer the reader to [34], for (II) to [4] (which also discussed several other constructions), and for (III) and (IV) to [48]. Only (II) provides the needed properties for the cells $Q_{j,k}$. However constructed, we denote by $\{Q_{j,k}\}$ the family of resulting dyadic cells and let $\mathscr{T}$ be the associated tree structure, as in Section 2.1.

In Fig. 4 we display pseudo-code for the GMRA of a data set X_n given a precision $\varepsilon > 0$ and a method τ_0 for choosing local dimensions (e.g., using thresholds or a fixed dimension). The code first constructs a family of multiscale dyadic cells (with local centers $c_{j,k}$ and bases $\Phi_{j,k}$) and then computes the geometric wavelets $\Psi_{j,k}$ and translations $w_{j,k}$ at all scales. In practice, we use METIS [34] to construct a dyadic (not 2^d-adic) tree $\mathscr{T}$ and the associated cells $Q_{j,k}$.

3.2 The Fast Geometric Wavelet Transform and Its Inverse

For simplicity of presentation, we shall assume $x = x_J$; otherwise, we may first project x onto the local linear approximation of the cell $Q_{J,x}$ and use x_J instead of x from now on. That is, we will define $x_{j;J} = P_{\mathscr{M}_j}(x_J)$, for all $j < J$, and encode the differences $x_{j+1;J} - x_{j;J}$ using the geometric wavelets. Note also that $\|x_{j;J} - x_j\| \leq \|x - x_J\|$ at all scales.

The geometric scaling and wavelet coefficients $\{p_{j,x}\}, \{q_{j+1,x}\}$, for $j \geq 0$, of a point $x \in \mathscr{M}$ are chosen to satisfy the equations

```
GMRA = GeometricMultiResolutionAnalysis $(X_n, \tau_0, \varepsilon)$

// Input:
// $X_n$: a set of $n$ samples from $\mathcal{M}$
// $\tau_0$: some method for choosing local dimensions
// $\varepsilon$: precision

// Output:
// A tree $\mathcal{T}$ of dyadic cells $\{Q_{j,k}\}$, their local means $\{m_{j,k}\}$ and bases $\{\Phi_{j,k}\}$, together with a
family of geometric wavelets $\{\Psi_{j,k}\}, \{w_{j,k}\}$

Construct the dyadic cells $Q_{j,k}$ with centers $\{m_{j,k}\}$ and form a tree $\mathcal{T}$.
$J \leftarrow$ finest scale with the $\varepsilon$-approximation property.
Let $\mathrm{cov}_{J,k} = |C_{J,k}|^{-1} \sum_{x \in C_{J,k}} (x - m_{J,k})(x - m_{J,k})^*$, for $k \in \mathcal{K}_J$, and compute $\mathrm{SVD}(\mathrm{cov}_{J,k}) =$
$\Phi_{J,k} \Sigma_{J,k} \Phi_{J,k}$ (where the dimension of $\Phi_{J,k}$ is determined by $\tau_0$).
for $j = J-1$ down to $0$

   for $k \in \mathcal{K}_j$

      Compute $\mathrm{cov}_{j,k}$ and $\Phi_{j,k}$ as above.
      For each $k' \in \mathrm{ch}(j,k)$, construct the wavelet bases $\Psi_{j+1,k'}$ and translations $w_{j+1,k'}$.

   end

end
For convenience, set $\Psi_{0,k} := \Phi_{0,k}$ and $w_{0,k} := m_{0,k}$ for $k \in \mathcal{K}_0$.
```

Fig. 4 Pseudo-code for the construction of geometric wavelets

$$P_{\mathcal{M}_j}(x) = \Phi_{j,x} p_{j,x} + m_{j,x}; \tag{19}$$

$$Q_{\mathcal{M}_{j+1}}(x) = \Psi_{j+1,x} q_{j+1,x} + w_{j+1,x} - P_{j,x} \sum_{l=j+1}^{J-1} Q_{\mathcal{M}_{l+1}}(x). \tag{20}$$

The computation of the coefficients, from fine to coarse, is simple and fast: since we assume $x = x_J$, we have

$$\begin{aligned} p_{j,x} &= \Phi_{j,x}^*(x_J - c_{j,x}) = \Phi_{j,x}^*(\Phi_{J,x} p_{J,x} + c_{J,x} - c_{j,x}) \\ &= \left(\Phi_{j,x}^* \Phi_{J,x}\right) p_{J,x} + \Phi_{j,x}^*(c_{J,x} - c_{j,x}). \end{aligned} \tag{21}$$

Moreover the wavelet coefficients $q_{j+1,x}$ [defined in Eq. (20)] are obtained from Eq. (14):

$$q_{j+1,x} = \Psi_{j+1,x}^*(x_{j+1} - c_{j+1,x}) = \left(\Psi_{j+1,x}^* \Phi_{j+1,x}\right) p_{j+1,x}. \tag{22}$$

Note that $\Phi_{j,x}^* \Phi_{J,x}$ and $\Psi_{j+1,x}^* \Phi_{j+1,x}$ are both small matrices (at most $d_{j,x} \times d_{j,x}$) and are the only matrices we need to compute and store (once for all, and only up to a specified precision) in order to compute all the wavelet coefficients $q_{j+1,x}$ and the scaling coefficients $p_{j,x}$, given $p_{J,x}$ at the finest scale.

```
$\{q_{j,x}\}$ =FGWT(GMRA,$x$)

// Input: GMRA structure, $x \in \mathscr{M}$
// Output: A sequence $\{q_{j,x}\}$ of wavelet coefficients

$p_{J,x} = \Phi^*_{J,x}(x - m_{J,x})$
for $j = J$ down to 1

    $q_{j,x} = (\Psi^*_{j,x}\Phi_{j,x})\, p_{j,x}$
    $p_{j-1,x} = (\Phi^*_{j-1,x}\Phi_{J,x})\, p_{J,x} + \Phi^*_{j-1,x}(m_{J,x} - m_{j-1,x})$

end
$q_{0,x} = p_{0,x}$ (for convenience)
```

Fig. 5 Pseudo-code for the forward geometric wavelet transform

```
$\hat{x}$ =IGWT(GMRA,$\{q_{j,x}\}$)

// Input: GMRA structure, wavelet coefficients $\{q_{j,x}\}$
// Output: Approximation $\hat{x}$ at scale $J$

$Q_{J,x} = \Psi_{J,x} q_{J,x} + w_{J,x}$
for $j = J - 1$ down to 1

    $Q_j(x) = \Psi_{j,x} q_{j,x} + w_{j,x} + \Phi_{j-1,x}\Phi^*_{j-1,x}\ \sum_{\ell > j} Q_\ell(x)$

end
$\hat{x} = \Psi_{0,x} q_{0,x} + w_{0,x} + \sum_{j>0} Q_j(x)$
```

Fig. 6 Pseudo-code for the inverse geometric wavelet transform

In Figs. 5 and 6 we display pseudo-codes for the computation of the forward and inverse geometric wavelet transforms (F/IGWT). The input to FGWT is a GMRA object, as returned by `GeometricMultiResolutionAnalysis`, and a point $x \in \mathscr{M}$. Its output is the wavelet coefficients of the point x at all scales, which are then used by IGWT for reconstruction of the point at all scales.

For any $x \in \mathscr{M}_J$, the set of coefficients

$$q_x = (q_{J,x}; q_{J-1,x}; \ldots; q_{1,x}; p_{0,x}) \tag{23}$$

is called the discrete *GWT* of x. Letting $d^w_{j,x} = \operatorname{rank}(\Psi_{j+1,x})$, the length of the transform is $d + \sum_{j>0} d^w_{j,x}$, which is bounded by $(J+1)d$ in the case of samples from a d-dimensional manifold (due to $d^w_{j,x} \leq d$).

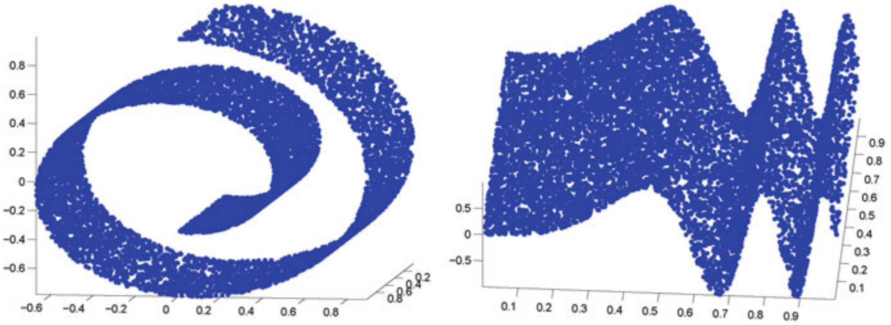

Fig. 7 Toy data sets for the following examples of GMRA

4 Examples

We conduct numerical experiments in this section to demonstrate the performance of the algorithm (i.e., Figs. 4–6).

4.1 *Low-Dimensional Smooth Manifolds*

To illustrate the construction presented so far, we consider simple synthetic data sets: a *SwissRoll*, an *S-Manifold*, and an *Oscillating2DWave*, all two-dimensional manifolds but embedded in $\mathbb{R}^{50}$ (see Fig. 7). We apply the algorithm to construct the GMRA and obtain the FGWT of the sampled data (10,000 points, without noise) in Fig. 8. We use the manifold dimension $d_{j,k} = d = 2$ at each node of the tree when constructing scaling functions and choose the smallest finest scale for achieving an absolute precision .001 in each case. We compute the average magnitude of the wavelet coefficients at each scale and plot it as a function of scale in Fig. 8. The reconstructed manifolds obtained by the inverse geometric wavelets transform (at selected scales) are shown in Fig. 9, together with a plot of relative approximation errors,

$$\mathscr{E}_{j,2}^{\mathrm{rel}} = \frac{1}{\sqrt{\mathrm{Var}(X_n)}} \sqrt{\frac{1}{n} \sum_{x \in X_n} \left(\frac{||x - P_{j,x}(x)||}{||x||} \right)^2}, \tag{24}$$

where X_n is the training data of n samples. Both the approximation error and the magnitude of the wavelet coefficients decrease quadratically with respect to scale as expected.

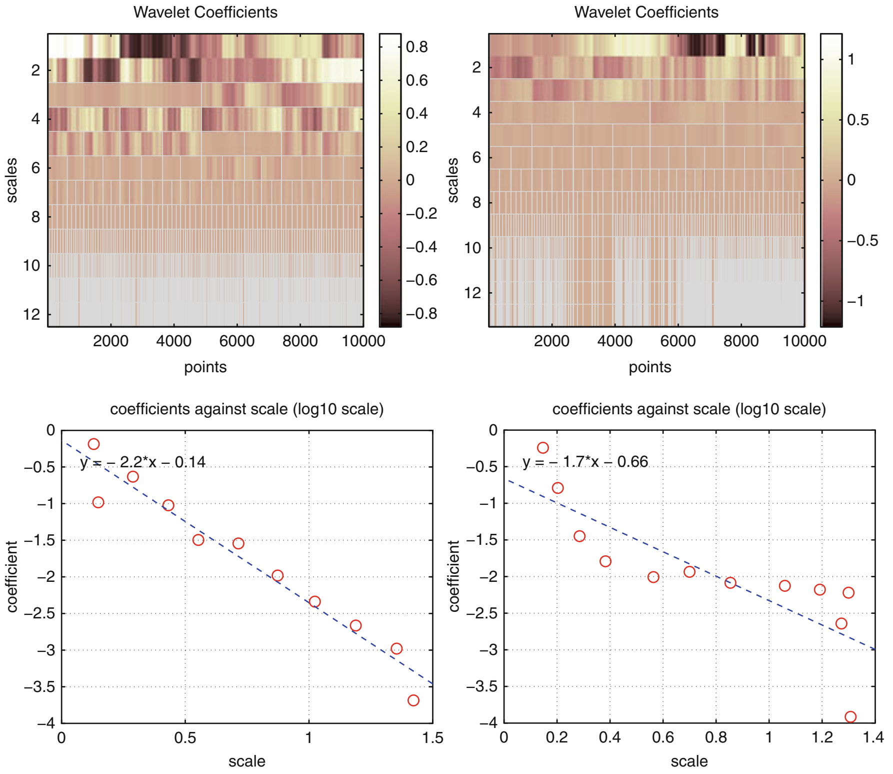

Fig. 8 *Top row*: wavelet coefficients obtained by the algorithm for the three data sets in Fig. 7. The horizontal axis indexes the points (arranged according to the tree), and the vertical axis multi-indexes the wavelet coefficients, from coarse (*top*) to fine (*bottom*) scales: the block of entries $(x, j), x \in Q_{j,k}$ displays $\log_{10} |q_{j,x}|$, where $q_{j,x}$ is the vector of geometric wavelet coefficients of x at scale j (see Sect. 3). In particular, each row indexes multiple wavelet elements, one for each $k \in \mathcal{K}_j$. *Bottom row*: magnitude of wavelet coefficients decreasing quadratically as a function of scale

We threshold the wavelet coefficients to study the compressibility of the wavelet coefficients and the rate of change of the approximation errors (using compressed wavelet coefficients). For this end, we use a smaller precision 10^{-5} so that the algorithm can examine a larger interval of thresholds. We threshold the wavelet coefficients of the *Oscillating2DWave* data at the level .01 and plot in Fig. 10 the reduced matrix of wavelet coefficients and the corresponding best reconstruction of the manifold (i.e., at the finest scale).

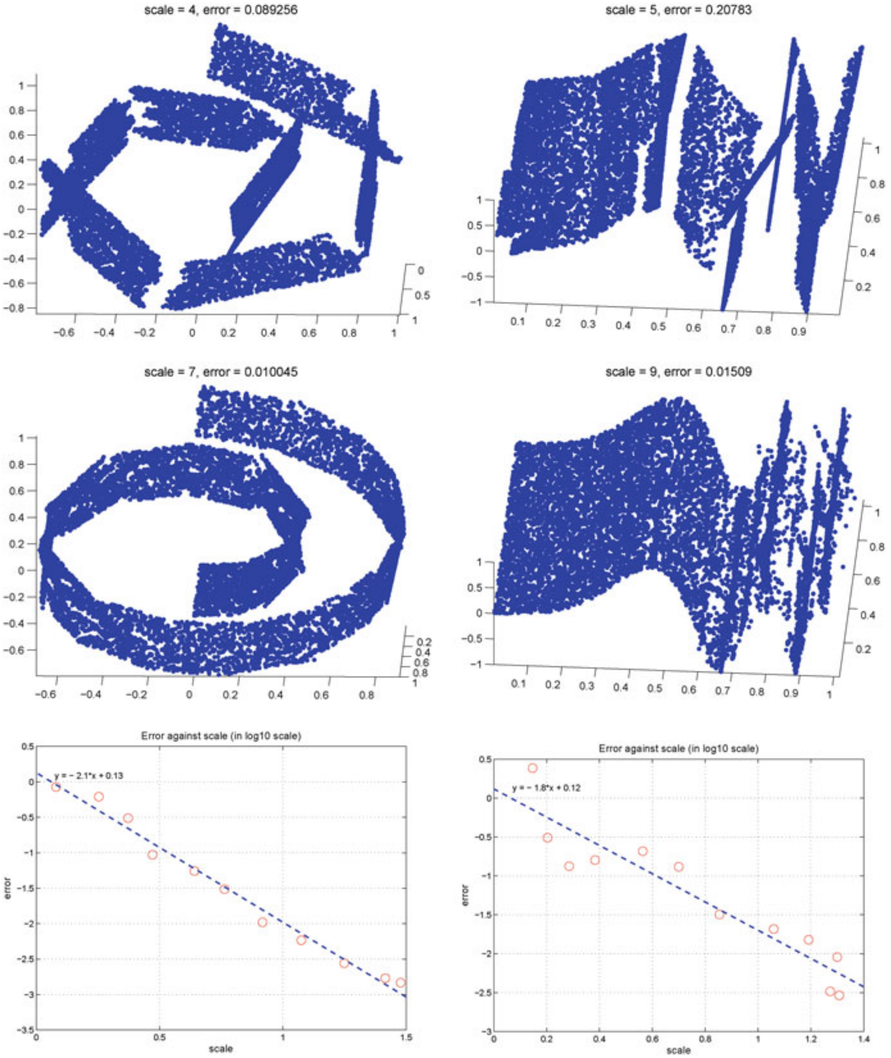

Fig. 9 *Top* and *middle*: reconstructions by the algorithm of the three toy data sets in Fig. 7 at two selected scales. *Bottom*: reconstruction errors as a function of scale

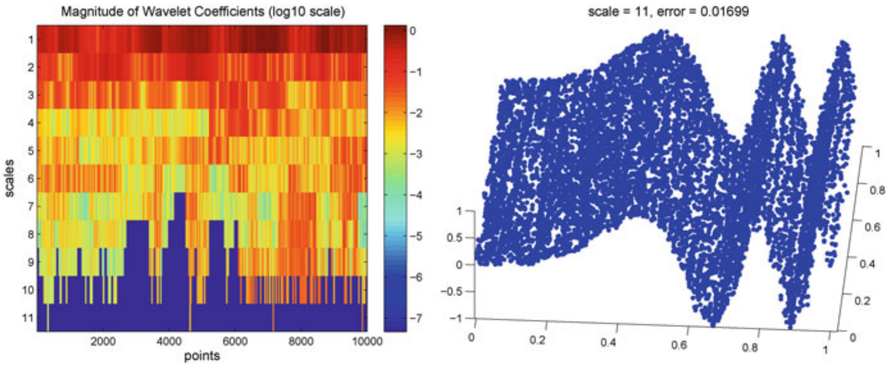

Fig. 10 The wavelet coefficients of the *Oscillating2DWave* data may be thresholded leading to adaptive approximation. *Left*: after sorting the points so that the x-axis orders them as going from *left* to *right* on the manifold, we see that when the manifold oscillates more, larger wavelet coefficients arise at fine scales. By threshold at the level of .01 and prune the dyadic tree accordingly, we reconstruct the manifold at the corresponding precision (*right*)

4.2 Data Sets

4.2.1 MNIST Handwritten Digits

We first consider the MNIST data set of images of handwritten digits,[1] each of size 28×28. We use the digits 0 and 1 and randomly sample for each digit 3,000 images from the database. We apply the algorithm to construct the geometric wavelets and show the wavelet coefficients and the reconstruction errors at all scales in Fig. 11. We select local dimensions for scaling functions by keeping 50 % and 95 % of the variance, respectively, at the nonleaf and leaf nodes. We observe that the magnitudes of the coefficients stop decaying after a certain scale. This indicates that the data is not on a smooth manifold. We expect optimization of the tree and of the wavelet dimensions in future work to lead to a more efficient representation in this case.

We then fix a data point (or equivalently an image), for each digit, and show in Fig. 12 its reconstructed coordinates at all scales and the corresponding dictionary elements (all of which are also images). We see that at every scale we have a handwritten digit, which is an approximation to the fixed image, and those digits are refined successively to approximate the original data point. The elements of the dictionary quickly fix the orientation and the thickness, and then they add other distinguishing features of the image being approximated.

[1] Available at http://yann.lecun.com/exdb/mnist/.

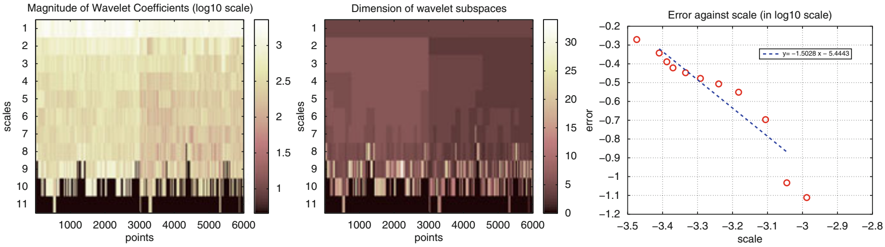

Fig. 11 From *left* to *right*: geometric wavelet representation of the MNIST digits data set for 1 and 0. As usual, the vertical axis multi-indexes the wavelet coefficients, from coarse (*top*) to fine (*bottom*) scales: the block of entries at $(x, j), x \in Q_{j,k}$ is $\log_{10}|q_{j,x}|$, where $q_{j,x}$ is the vector of geometric wavelet coefficients of x at scale j (see Sect. 3). In particular, each row indexes multiple wavelet elements, one for each $k \in \mathscr{K}_j$. *Top right*: dimensions of the wavelet subspaces (with the same convention as in the previous plot): even if the data lies in 784 dimensions, the approximating planes used have mostly dimension 1–6, except for some planes at the leaf nodes. *Rightmost inset*: reconstruction error as functions of scale. The decay is nonlinear and not what we would expect from a manifold structure

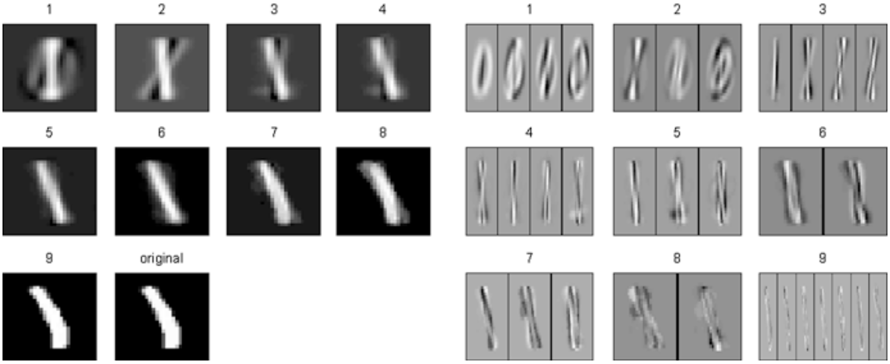

Fig. 12 *Left*: in each figure we plot coarse-to-fine geometric wavelet approximations of the original data point (represented in the last image). *Right*: elements of the wavelet dictionary (ordered from coarsest to finest scales) used in the expansion of the data point on the *left*

Example: A Connection to Fourier Analysis and FFT

Consider band-limited functions of band B:

$$BF_B = \{f : \text{supp.}\,\hat{f} \subseteq [-B\pi, B\pi]\}.$$

Classes of smooth functions (e.g., $W^{k,2}$) are essentially characterized by their L^2-energy in dyadic spectral bands of the form $[-2^{j+1}\pi, -2^j\pi] \cup [2^j\pi, 2^{j+1}\pi]$, i.e., by the L^2-size of their projection onto $BF_{2^{j+1}} \ominus BF_{2^j}$ (some care is of course needed in that smooth frequency cutoff, but this issue is not relevant for our purposes here). We generate random smooth (band-limited!) functions as follows:

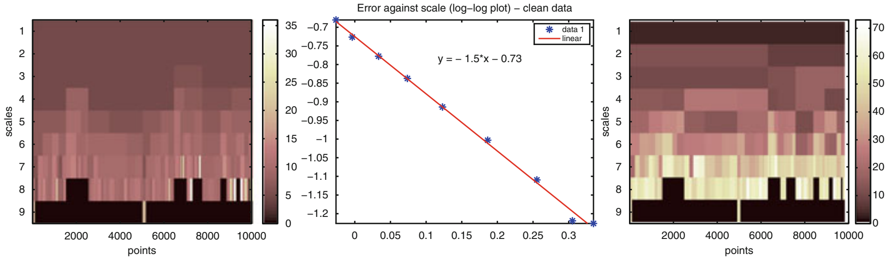

Fig. 13 We construct an orthogonal geometric multi-resolution analysis (see [2]) on a random sample of 10, 000 band-limited functions. *Left*: dimension of the GMRA wavelet subspaces. *Center*: approximation error as a function of scale. *Right*: dominant frequency in each GMRA subspace, showing that frequencies are sorted from low (*top*, coarse GMRA scales) to high (*bottom*, fine GMRA scales). This implies that the geometric scaling function subspaces roughly correspond to a Littlewood–Paley decomposition, and the GWT of a function f corresponds to a rough standard wavelet transform

$$f(x) = \sum_{j=0}^{J} a_j(\omega)\cos(jx)$$

with a_j random Gaussian (or bounded) with mean $2^{-\lfloor \frac{j}{J} \rfloor \alpha}$ and standard deviation $2^{-\lfloor \frac{j}{J} \rfloor \alpha} \cdot \frac{1}{5}$. The GMRA associated with a random sample of this family of functions takes advantage of the multiscale nature of the data and organizes this family of functions in a Littlewood–Paley type of decomposition: the scaling function subspace at scale j roughly corresponds to $BF_{2^{j+1}} \ominus BF_{2^j}$, and the GMRA of a point is essentially a block Fourier transform, where coefficients in the same dyadic band are grouped together. Observe that the cost of the GMRA of a point f is comparable to the cost of the fast Fourier transform.

5 Data Representation, Compression, and Computational Considerations

A set of n points in $\mathbb{R}^D$ can trivially be stored in space Dn; if it lies, up to a least squares error ε in a linear subspace of dimension $d_\varepsilon \ll D$, we could encode n points in space $d_\varepsilon(D+n)$ (cost of encoding a basis for the linear subspace, plus encoding of the coefficients of the points on that basis). This is much less than the trivial encoding for $d_\varepsilon \ll D$. It can be shown [2] that the cost of encoding with a GMRA a $\mathscr{C}^2$ manifold $\mathscr{M}$ of dimension d sampled at n points, for a fixed precision $\varepsilon > 0$ and n large, is $O(\varepsilon^{-\frac{d}{2}} dD + nd\log_2 \varepsilon^{-\frac{1}{2}})$.

Also, the cost of the algorithm is

$$O(nD(\log(n)+d^2)) + O_{d,D}(n\log n),$$

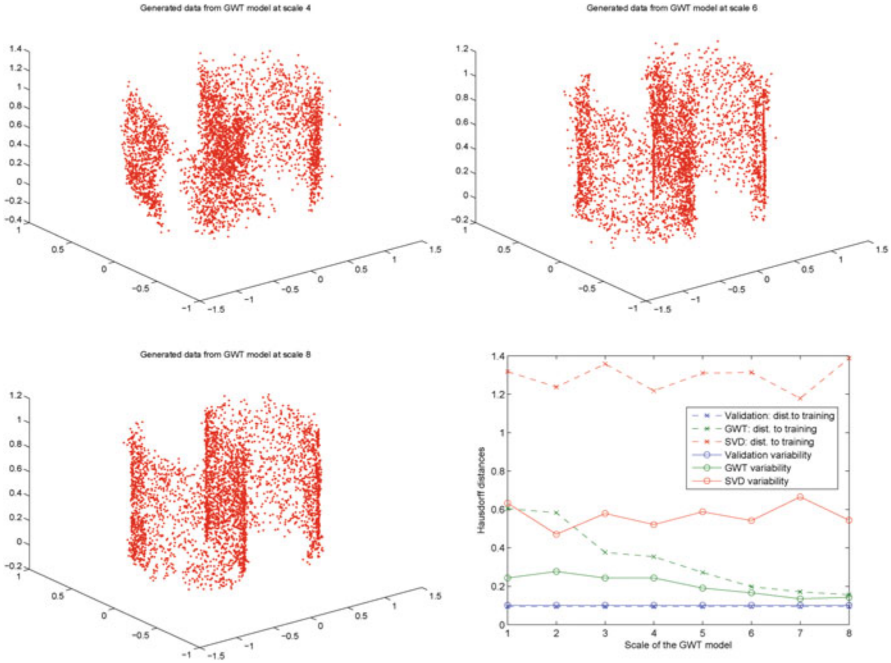

Fig. 14 Approximations of the probability distribution concentrated on a S-shaped 2-dimensional manifold within the GMRA framework. From *left* to *right*, *top* to *bottom*: $4,000$ samples drawn from our approximate distribution, constructed at scale 4, 6, and 8, respectively, from $2,000$ training samples. *Bottom right*: as a function of scale, the Hausdorff distance between points generated by the SVD model and GWT models and the training data, as well as the Hausdorff distance variability of the generated data and true data. We see that $p_{\mathscr{M}_j}$ has small distance to the training set and decreasingly so for models constructed at finer scales, while p_{SVD_j}, being a model in the ambient space, generates points farther from the distribution. Looking at the plots of the in-model Hausdorff distance variability, we see that such measure increases for $p_{\mathscr{M}_j}$ as a function of j (reflecting the increasing expression power of the model). Samples from the SVD model look like a Gaussian point cloud, as the kernel density estimator did not have enough training samples to adapt to the low-dimensional manifold structure

while the cost of performing the FGWT of a point is

$$O(2^d D \log n + dD + d^2 \log \varepsilon^{-\frac{1}{2}}).$$

The cost of the IGWT is similar but without the first term.

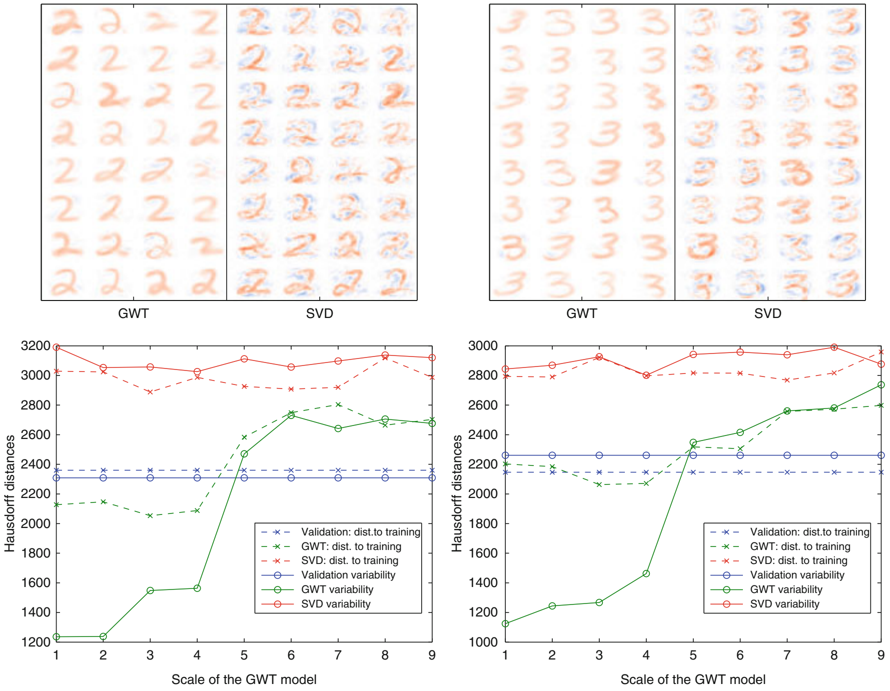

Fig. 15 *Left* and *right columns*: a training set of 2,000 digits 2 (respectively, 3) from the MNIST data set are used to train probability models with GMRA ($p_{\mathscr{M}_j}$, one for each scale j in the GMRA of the training set) and SVD (p_{SVD_j}, one for each GMRA scale, see text). *Left:* 32 digits drawn from $p_{\mathscr{M}_5}$, p_{SVD_5}: the quality of $p_{\mathscr{M}_5}$ is qualitatively better than that of p_{SVD_5}. *Center*: plots of the Hausdorff distance to training set and in-model Hausdorff distance variability. *Right*: a similar experiment with a training set of 2,000 points from a SwissRoll-shaped manifold with no noise: the finest scale GMRA-based models perform best (in terms of both approximation and variability, the SVD-based models are once again unable to take advantage of the low intrinsic dimension)

6 Multiscale Models of Densities

We present a simple example of how our techniques may be used to model measures supported on low-dimensional sets which are well approximated by the multiscale planes we constructed; results from more extensive investigations will be reported in an upcoming publication.

We sample n training points from a point cloud $\mathscr{M}$ and, for a fixed scale j, we consider the coarse approximation $\mathscr{M}_j$ [defined in Eq. (10)], and on each local linear approximating plane $V_{j,k}$ we use the training set to construct a multifactor Gaussian model on $Q_{j,k}$: let $\pi_{j,k}$ be the estimated distribution. We also estimate from the training data the probability $\pi_j(k)$ that a given point in $\mathscr{M}$ belongs to $Q_{j,k}$ (recall that j is fixed, so this is a probability distribution over the $|\mathscr{K}_j|$ labels of the planes at scale j). We may then generate new data points by drawing a $k \in \mathscr{K}_j$ according to π_j

and then drawing a point in $V_{j,k}$ from the distribution $\pi_{j,k}$: this defines a probability distribution supported on $\mathscr{M}_j$ that we denote by $p_{\mathscr{M}_j}$.

In this way we may generate new data points which are consistent with both the geometry of the approximating planes $V_{j,k}$ and with the distribution of the data on each such plane. In Fig. 14 we display the result of such modeling on a simple manifold. In Fig. 15 we construct $p_{\mathscr{M}_j}$ by training on $2,000$ handwritten 2s and 3s from the MNIST database, and on the same training set we train two other algorithms: the first one is based on projecting the data on the first a_j principal components, where a_j is chosen so that the cost of encoding the projection and the projected data is the same as the cost of encoding the GMRA up to scale j and the GMRA of the data and then running the same multifactor Gaussian model used above for generating $\pi_{j,k}$. This leads to a probability distribution we denote by p_{SVD_j}. In order to test the quality of these models, we consider the following two measures. The first measure is simply the Hausdorff distance between $2,000$ randomly chosen samples according to each model and the training set: this is measuring how close the generated samples are to the training set. The second measure quantifies if the model captures the variability of the true data and is computed by generating multiple point clouds of $2,000$ points for a fixed model and looking at the pairwise Hausdorff distances between such point clouds, called the within-model Hausdorff distance variability.

The bias–variance trade-off in the models $p_{\mathscr{M}_j}$ is the following: as j increases the planes better model the geometry of the data (under our usual assumptions), so that the bias of the model (and the approximation error) decreases as j increases; on the other hand the sampling requirements for correctly estimating the density of $Q_{j,k}$ projected on $V_{j,k}$ increases with j as less and less training points fall in $Q_{j,k}$. A pruning greedy algorithm that selects, in each region of the data, the correct scale for obtaining the correct bias–variance trade-off, depending on the samples and the geometry of the data, similar in spirit to the what has been studied in the case of multiscale approximation of functions, will be presented in a forthcoming publication. It should be remarked that such a model would be very complicated in the wavelet domain, as one would need to model very complex dependencies among wavelet coefficients, in both space and scale.

Acknowledgments The authors thank E. Monson for useful discussions. AVL was partially supported by NSF and ONR. GC was partially supported by DARPA, ONR, NSF CCF, and NSF/DHS FODAVA program. MM is grateful for partial support from DARPA, NSF, ONR, and the Sloan Foundation.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: Design of dictionaries for sparse representation. In: Proceedings of SPARS 05', pp. 9–12 (2005)
2. Allard, W.K., Chen, G., Maggioni, M.: Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. Appl. Computat. Harmonic Analysis **32**, 435–462 (2012)

3. Belkin, M., Niyogi, P.: Using manifold structure for partially labelled classification. Advances in NIPS, vol. 15. MIT Press, Cambridge (2003)
4. Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbor. In: ICML, pp. 97–104 (2006)
5. Binev, P., Cohen, A., Dahmen, W., Devore, R., Temlyakov, V.: Universal algorithms for learning theory part i: Piecewise constant functions. J. Mach. Learn. **6**, 1297–1321 (2005)
6. Binev, P., Devore, R.: Fast computation in adaptive tree approximation. Numer. Math. **97**, 193–217 (2004)
7. Bremer, J., Coifman, R., Maggioni, M., Szlam, A.: Diffusion wavelet packets. Appl. Comp. Harm. Anal. **21**, 95–112 (2006) (Tech. Rep. YALE/DCS/TR-1304, 2004)
8. Candès, E., Donoho, D.L.: Curvelets: A surprisingly effective nonadaptive representation of objects with edges. In: Schumaker, L.L., et al. (eds.) Curves and Surfaces. Vanderbilt University Press, Nashville (1999)
9. Causevic, E., Coifman, R., Isenhart, R., Jacquin, A., John, E., Maggioni, M., Prichep, L., Warner, F.: QEEG-based classification with wavelet packets and microstate features for triage applications in the ER, vol. 3. ICASSP Proc., May 2006 10.1109/ICASSP.2006.1660859
10. Chen, G., Little, A., Maggioni, M., Rosasco, L.: Wavelets and Multiscale Analysis: Theory and Applications. Springer (2011) submitted March 12th, 2010
11. Chen, G., Maggioni, M.: Multiscale geometric wavelets for the analysis of point clouds. Information Sciences and Systems (CISS), 2010 44th Annual Conference on. IEEE, 2010.
12. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
13. Christ, M.: A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral. Colloq. Math. **60–61**, 601–628 (1990)
14. Christensen, O.: An introduction to frames and Riesz bases. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2003)
15. Coifman, R., Lafon, S.: Diffusion maps. Appl. Comp. Harm. Anal. **21**, 5–30 (2006)
16. Coifman, R., Lafon, S., Maggioni, M., Keller, Y., Szlam, A., Warner, F., Zucker, S.: Geometries of sensor outputs, inference, and information processing. In: Athale, R.A. (ed.) Proc. SPIE, J. C. Z. E. Intelligent Integrated Microsystems, vol. 6232, p. 623209, May 2006
17. Coifman, R., Maggioni, M.: Diffusion wavelets. Appl. Comp. Harm. Anal. **21**, 53–94 (2006) (Tech. Rep. YALE/DCS/TR-1303, Yale Univ., Sep. 2004).
18. Coifman, R., Maggioni, M.: Multiscale data analysis with diffusion wavelets. In: Proc. SIAM Bioinf. Workshop, Minneapolis (2007)
19. Coifman, R., Maggioni, M.: Geometry analysis and signal processing on digital data, emergent structures, and knowledge building. SIAM News, November 2008
20. Coifman, R., Meyer, Y., Quake, S., Wickerhauser, M.V.: Signal processing and compression with wavelet packets. In: Progress in Wavelet Analysis and Applications (Toulouse, 1992), pp. 77–93. Frontières, Gif (1993)
21. Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. PNAS **102**, 7426–7431 (2005)
22. Daubechies, I.: Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (1992) ISBN: 0-89871-274-2.
23. David, G.: Wavelets and singular integrals on curves and surfaces. In: Lecture Notes in Mathematics, vol. 1465. Springer, Berlin (1991)
24. David, G.: Wavelets and Singular Integrals on Curves and Surfaces. Springer, Berlin (1991)
25. David, G., Semmes, S.: Analysis of and on uniformly rectifiable sets. Mathematical Surveys and Monographs, vol. 38. American Mathematical Society, Providence (1993)
26. David, G., Semmes, S.: Uniform Rectifiability and Quasiminimizing Sets of Arbitrary Codimension. American Mathematical Society, Providence (2000)
27. Donoho, D.L., Grimes, C.: When does isomap recover natural parameterization of families of articulated images? Tech. Rep. 2002–2027, Department of Statistics, Stanford University, August 2002

28. Donoho, D.L., Grimes, C.: Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Proc. Nat. Acad. Sciences **100**, 5591–5596 (2003)
29. Golub, G., Loan, C.V.: Matrix Computations. Johns Hopkins University Press, Baltimore (1989)
30. Jones, P., Maggioni, M., Schul, R.: Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. Proc. Nat. Acad. Sci. **105**, 1803–1808 (2008)
31. Jones, P., Maggioni, M., Schul, R.: Universal local manifold parametrizations via heat kernels and eigenfunctions of the Laplacian. Ann. Acad. Scient. Fen. **35**, 1–44 (2010) http://arxiv.org/abs/0709.1975.
32. Jones, P.W.: Rectifiable sets and the traveling salesman problem. Invent. Math. **102**, 1–15 (1990)
33. Jones, P.W.: The traveling salesman problem and harmonic analysis. Publ. Mat. **35**, 259–267 (1991) Conference on Mathematical Analysis (El Escorial, 1989)
34. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. **20**, 359–392 (1999)
35. Little, A., Jung, Y.-M., Maggioni, M.: Multiscale estimation of intrinsic dimensionality of data sets. In: Proc. A.A.A.I. (2009)
36. Little, A., Lee, J., Jung, Y.-M., Maggioni, M.: Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In: Proc. S.S.P. (2009)
37. Little, A., Maggioni, M., Rosasco, L.: Multiscale geometric methods for data sets I: Estimation of intrinsic dimension, submitted (2010)
38. Maggioni, M., Bremer, J. Jr., Coifman, R., Szlam, A.: Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs. SPIE, vol. 5914, p. 59141M (2005)
39. Maggioni, M., Mahadevan, S.: Fast direct policy evaluation using multiscale analysis of markov diffusion processes. In: ICML 2006, pp. 601–608 (2006)
40. Mahadevan, S., Maggioni, M.: Proto-value functions: A spectral framework for solving markov decision processes. JMLR **8**, 2169–2231 (2007)
41. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML, p. 87 (2009)
42. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11**, 19–60 (2010)
43. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res. **37**, 3311–3325 (1997)
44. Rahman, I.U., Drori, I., Stodden, V.C., Donoho, D.L.: Multiscale representations for manifold-valued data. SIAM J. Multiscale Model. Simul. **4**, 1201–1232 (2005).
45. Rohrdanz, M.A., Zheng, W., Maggioni, M., Clementi, C.: Determination of reaction coordinates via locally scaled diffusion map. J. Chem. Phys. **134**, 124116 (2011)
46. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
47. Starck, J.L., Elad, M., Donoho, D.: Image decomposition via the combination of sparse representations and a variational approach. IEEE T. Image Process. **14**, 1570–1582 (2004)
48. Szlam, A.: Asymptotic regularity of subdivisions of euclidean domains by iterated PCA and iterated 2-means. Appl. Comp. Harm. Anal. **27**, 342–350 (2009)
49. Szlam, A., Maggioni, M., Coifman, R., Bremer, J. Jr.: Diffusion-driven multiscale analysis on manifolds and graphs: top-down and bottom-up constructions. SPIE, vol. 5914(1), p. 59141D (2005)
50. Szlam, A., Maggioni, M., Coifman, R.: Regularization on graphs with function-adapted diffusion processes. J. Mach. Learn. Res. **9**, 1711–1739 (2008) (YALE/DCS/TR1365, Yale Univ, July 2006)
51. Szlam, A., Sapiro, G.: Discriminative *k*-metrics. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1009–1016 (2009)
52. Tenenbaum, J.B., Silva, V.D., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)

53. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc. B **58**, 267–288 (1996)
54. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J. Sci. Comput. **26**, 313–338 (2002)
55. Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., Carin, L.: Non-parametric Bayesian dictionary learning for sparse image representations. In: Neural and Information Processing Systems (NIPS) (2009)