Some recent advances in multiscale geometric analysis of point clouds

Guangliang Chen, Anna V. Little, Mauro Maggioni, and Lorenzo Rosasco

Abstract We discuss recent work based on multiscale geometric analysis for the study of large data sets that lie in high-dimensional spaces but have low-dimensional structure. We present three applications: the first one to the estimation of intrinsic dimension of sampled manifolds, the second one to the construction of multiscale dictionaries, called geometric wavelets, for the analysis of point clouds, and the third one to the inference of point clouds modeled as unions of multiple planes of varying dimension.

1 Introduction

Data sets that arise in a variety of settings - from images and movies to web pages, customer transaction records, gene microarrays, etc... - are being collected at ever increasing speeds and level of detail. The increase in the amount of data has not always been matched by our understanding of how to efficiently extract information, and to search, organize, and derive useful predictions from such data sets. The analysis of such data sets, modeled as point clouds in high-dimensional spaces, is an emerging and challenging area in applied mathematics, at the boundary with other disciplines such as computer science, engineering, signal processing, biology, and more. There are many applications, including organization of large libraries of

Guangliang Chen
Duke University, Durham NC, e-mail: glchen@math.duke.edu

Anna V. Little
Duke University, Durham NC, e-mail: avl@math.duke.edu

Mauro Maggioni
Duke University, Durham NC, e-mail: mauro@math.duke.edu

Lorenzo Rosasco
Massachusetts Institute of Technology, Cambridge MA, e-mail: lrosasco@mit.edu, and DISI, Universitá di Genova, Italy

documents, face recognition [52], semi–supervised learning [5, 73, 83], nonlinear image denoising and segmentation [80, 83], clustering [72, 3], machine learning [5, 73, 83, 68, 67, 84], processing of articulated images [44], cataloguing of galaxies [45], pattern analysis of brain potentials [66], the study of brain tumors [21], document classication and web searching, hyperspectral imaging, and many others. The analysis and estimation of geometric (intended in the widest sense) properties of the data include problems such as dimension estimation (e.g. [31, 15, 13, 14] and references therein), nonlinear dimension reduction [10, 86, 43, 76, 4, 6, 60, 29, 44, 78, 79, 90, 87, 88] and metric space embeddings [12, 10, 59]. Oftentimes one is interested in studying functions on the data, for the purpose of denoising, fitting, and prediction. These questions can be studied through approximation theory (e.g. [8, 9] and references therein), machine learning [4], and signal processing [70, 34], at least in low-dimensional Euclidean spaces. The combination of the study of geometric properties with the study of functions defined on the data is quite a recent and promising trend [83, 4, 6, 92, 60, 30, 29].

We will for the moment restrict our attention to data sets represented as discrete sets in $\mathbb{R}^D$. A feature common to many data sets is their high-dimension D, which may range from 10 to 10^6. This implies that classical statistics, by which we mean the analysis in the case where the dimension D is fixed and the number of points n goes to infinity (or, at least $n >> 2^D$), is not applicable. Typical situations that we consider have n of the same order as D, and oftentimes $n < D$. In this regime, more appropriate asymptotics are those with n fixed and D going to infinity.

A key observation is that in several situations the data seems to be concentrated along low-dimensional sets in $\mathbb{R}^D$ (e.g. [10, 86, 43, 76, 4, 6, 60, 29, 44, 78, 79, 90, 87, 88]). In this case it is natural to ask what geometric properties these low-dimensional sets have, and how to exploit this phenomenon in order to better model and learn from data.

The interplay between geometry of sets, function spaces on sets, and operators on sets is if course classical in Harmonic Analysis.

This paper gives an overview of very recent work in the geometric analysis of high-dimensional point clouds and tries to briefly summarize the papers [65, 2, 24]. Material related to this paper is available at http://www.math.duke.edu/˜mauro.

2 Multiscale SVD

The quantitative study of geometric properties of sets, such as rectifiability and harmonic analysis, is classical [56, 37, 39, 40, 38, 35, 41, 36]. The applications of ideas from geometric measure theory to the analysis of point clouds are emerging, and here we would like to review a small number of very recent ones.

One of the basic tools in the analysis of data set in statistics is Principal Component Analysis (PCA), which is based on the Singular Value Decomposition (SVD). Any $n \times D$ matrix X may be decomposed as $X = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{D\times D}$ are orthonormal and $\Sigma \in \mathbb{R}^{n\times D}$ is diagonal and positive semidefinite.

The diagonal entries $\{\lambda_i\}$ of Σ, called singular values (S.V.'s), are ordered in decreasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n\wedge D} \geq 0$. This is called the SVD of X. It is useful to notice that the first d columns of V span a linear subspace that minimizes $\sum_{i=1}^n ||x_i - P_\Pi(x_i)||^2_{\mathbb{R}^D}$ over all choice of d-dimensional linear subspaces Π (here P_Π denotes the orthogonal projection onto Π). We say that the first d columns of V produce the d-dimensional least squares fit to X. If the rows $\{x_i\}_{i=1}^n$ of X represent n data points in $\mathbb{R}^D$, Principal Component Analysis consists in computing the empirical mean $m(X) = \frac{1}{n}\sum_{i=1}^n x_i$, considering the new matrix $\overline{X}$ whose rows are $x_i - m(X)$, and computing the SVD of $\overline{X}$. The columns of V are called the principal vectors. An alternative interpretation is the following: if we let

$$\begin{aligned} \mathrm{cov}(X) &= \frac{1}{n}\sum_{i=1}^n (x_i - m(X)) \otimes (x_i - m(X)) \\ &= \frac{1}{n}\sum_{i=1}^n (x_i - m(X))^T (x_i - m(X)) \end{aligned} \tag{1}$$

be the empirical covariance matrix of X, then from $\overline{X} = U\Sigma V^T$ we immediately deduce that

$$\mathrm{cov}(X) = V\Sigma^T \Sigma V^T \tag{2}$$

so that λ_i^2 are the eigenvalues of the covariance matrix. Of course in this setting one thinks of having a random variable X taking values in $\mathbb{R}^D$, and the mean and covariance are its first and second order statistics. It is clear how to generalize PCA and SVD to the case of infinitely many points.

If the data points $\{x_i\}$ lie, say, uniformly on a bounded domain in a d-dimensional linear subspace, then for n large enough (in fact, $n \gtrsim d\log d$ is enough [77] [1]), X will be of rank exactly d. The first d empirical singular values will have the same order as the true ones (i.e. the ones obtained as $n \to \infty$) and the remaining ones will be exactly 0, and the first d columns of V will span the linear subspace. Because of the least squares fit property, we expect that if we add "small" noise to the points x_i, the smallest singular values should be perturbed by a small amount and would still be much smaller than the top d singular values, indicating the existence of a d-dimensional linear subspace where the (noiseless) data lies.

We are interested in the case where the data points lie on a d-dimensional nonlinear manifold $\mathscr{M}$ embedded in a high-dimensional space $\mathbb{R}^D$ and are corrupted by high-dimensional noise. This model has gained popularity in the machine learning community in recent years [10, 86, 43, 76, 4, 6, 60, 29, 44, 78, 79, 90, 87, 88]. While only in particular cases may one expect this model to be correct for real data, it is a step beyond linear models. More general models may also be considered, such as unions of manifold, possibly intersecting each other, and possibly of different dimensions. Understanding these more general models seems to require the understanding of the simpler model with one manifold only. A particular choice of this general model is that of a union of planes, possibly of different dimensions [1, 82].

[1] These bounds are continuously being refined, in particular in rather general cases the $\log d$ term may be reduced to $\log\log d$.

In order to study these more complicated models, it may seem tempting to discard SVD, which is so well-adapted to linear models. The SVD of a matrix X representing data on a d-dimensional manifold $\mathscr{M}$ in $\mathbb{R}^D$ may not reveal d at all. As a first trivial example, consider a planar circle ($d = 1$) of radius r embedded in $\mathbb{R}^D$: $\mathrm{cov}(X)$ has exactly 2 nonzero eigenvalues equal to $\frac{r}{\sqrt{2}}$. More generally, it is easy to construct a one-dimensional manifold ($d = 1$) such that $\mathrm{cov}(X)$ has full rank (w.h.p.): it is enough to pick a curve that spirals out in more and more dimensions. A simple construction (sometimes called Y. Meyer's staircase) is the following: pick the D points $0, 1, \ldots, D-1$ on the real line, and let $\chi_{[0,2)}(x) = 1$ if $x \in [0,2)$ and 0 otherwise. Then the set

$$\{x_t := \chi_{[0,2)}(\cdot - t)\}_{t\in\mathbb{R}} \subset L^1(\mathbb{R}) \tag{3}$$

is a one-dimensional manifold, which is not contained in any subspace of dimension less than D. This may be discretized by evaluating the functions x_t on the discrete set $\{0, 1, \ldots, D-1\}$. Notice that x_{t_1} and x_{t_2} are orthogonal whenever $|t_1 - t_2| > 2$, so this curve spirals into larger and larger subspaces as t increases. Similar considerations would hold after discretization of the space and restriction of t to a bounded interval.

However, one may still make good use of PCA if one performs it locally at multiple scales: for every $r > 0$ and every $z \in \mathscr{M}$ consider

$$X_{z,r} := \mathscr{M} \cap B_z(r) \tag{4}$$

i.e. the intersection of $\mathscr{M}$ with a Euclidean ball (in the ambient space $\mathbb{R}^D$) centered at z of radius r. Perform PCA on $X_{z,r}$, and let $\{\lambda_{i,z,r}\}_{i=1}^D$ be the corresponding singular values. Also, let

$$\Delta_i(X_{z,r}) = \lambda_{i,z,r}^2 - \lambda_{i+1,z,r}^2 \qquad \text{for } 1 \le i \le D-1 \tag{5}$$

and $\Delta_D(X_{z,r}) = \lambda_{D,z,r}^2$; these are the gaps of the squared singular values of $X_{z,r}$. For a fixed z, how do these singular values behave? We expect that for small r the top d singular values $\lambda_{1,z,r}, \ldots, \lambda_{d,z,r}$ will grow linearly in r and be large compared to the remaining ones, which are associated with normal directions and grow quadratically in r. The principal components corresponding to the top d singular values will approximate the tangent space to $\mathscr{M}$ at z. This allows one to estimate d. As r grows, however, $\mathscr{M}$ will start curving inside $B_z(r)$ and the bottom singular values will start to grow, eventually (in general) becoming as large as the top d singular values, as in the examples mentioned above. Therefore the curvature of $\mathscr{M}$ inside $\mathbb{R}^D$ puts an upper bound on the set of scales that may be used to detect d via SVD.

This clear picture becomes more complicated if we add two factors crucial in applications: sampling and noise. We only have a finite number of samples X_n from $\mathscr{M}$, which will put a lower bound on the values of r: if r is too small, $X_{n,z,r} := X_n \cap B_z(r)$ will simply be empty, or may not contain enough points to be able to determine the intrinsic dimension d. If high-dimensional noise is added to the samples, so that our observations are in the form $x_i + \eta_i$, with $x_i \in \mathscr{M}$ and η_i representing noise (e.g. $\eta \sim \sigma\mathscr{N}(0, I_D)$, with $\mathscr{N}$ denoting the Gaussian distribution), then another lower

bound on r arises: if r is small compared to the "size" of the noise η, even if we have enough samples in $X_{n,z,r}$, these samples will look high-dimensional because they are scattered in $\mathbb{R}^D$ by the noise. It is only at scales r higher than the "size" of the noise that there is a chance for the top singular values to detect linear growth (in r) because the SVD will detect a noisy tangent plane. But once again, at larger scales curvature will take over. Of course, in the range of scales above that of the noise and below that dictated by curvature, $X_{n,z,r}$ must have enough samples so that the SVD may be computed reliably. We discuss this problem in some detail in Section 3, which is a short summary of the work in [65] (see also [64, 63]).

In Section 4 we discuss the problem of efficiently representing data on nonlinear d-dimensional manifolds $\mathscr{M}$ embedded in $\mathbb{R}^D$, with $d \ll D$, by constructing geometric dictionaries. If $\mathscr{M}$ were linear, then we could perform SVD, use dD numbers to encode the d-dimensional subspace $\mathscr{M}$ lies on (e.g. by the first d columns of the matrix V), and then every point on $\mathscr{M}$ would require only d coefficients, instead of D. When $\mathscr{M}$ is not linear nor contained in a low-dimensional subspace of $\mathbb{R}^D$, it is not clear how to generalize such a construction in order to efficiently store the data. We briefly discuss recent work based on so-called geometric wavelets [2, 25], which aim at efficiently encoding the multiscale family of SVD's discussed above by encoding the difference between approximate tangent planes at different scales. This encoding not only reduces the cost of encoding these planes, but yields a multiscale decomposition of every point of $\mathscr{M}$, and therefore of $\mathscr{M}$ itself, and fast but nonlinear algorithms for computing a fast geometric wavelet transform and its inverse for every point. This may be thought of as a geometric version of wavelets. Much needs to be explored in these directions. In any case, this yields multiscale matrix decompositions, that allow one to efficiently encode the data and that reveal structures in data. Section 4 is a short summary of the work in [2] (see also [25]).

Finally, we discuss the problem of estimating the family of planes when data is modeled as lying on multiple planes, of possibly different dimensions, and possibly intersecting. This is the topic of Section 5, which is a short summary of the work [24].

3 Intrinsic dimension estimation

The problem of estimating the intrinsic dimension of a point cloud is of interest in a wide variety of situations, such as estimating the number of variables in a linear model in statistics, the number of degrees of freedom in a dynamical system, the intrinsic dimension of a data set modeled by a probability distribution highly concentrated around a low-dimensional manifold. Many applications and algorithms crucially rely on the estimation of the number of components in the data, for example spectrometry, signal processing, genomics and economics, to name only a few. Moreover, many manifold learning algorithms [10, 86, 43, 76, 4, 6, 60, 29, 44, 78, 79, 90, 87, 88] assume that the intrinsic dimension is given.

When the data lies on a plane - as for example when it is generated by a multivariate linear model - principal component analysis allows one to recover the plane. This case is well understood as the number of samples grows to infinity and also when noise is present (see e.g., out of many works, [54], [74], [81] and references therein).

The finite sample situation is less well understood. Even in this case of points on a plane we derive new results using a new approach; however the situation we are really interested in is that of data having a geometric structure more complicated than linear, such as when the data lies on a low-dimensional manifold. Several algorithms have been proposed to estimate intrinsic dimension in this setting; for lack of space we cite only [61, 50, 18, 16, 32, 14, 75, 58, 11, 71, 49, 85, 53, 47].

3.1 Multiscale Dimension Estimation

We start by describing a stochastic geometric model generating the point clouds we will study. Let $(\mathscr{M},g)$ be a compact smooth d-dimensional Riemannian manifold, isometrically embedded in $\mathbb{R}^D$. Let η be $\mathbb{R}^D$-valued with $\mathrm{e}[\eta]=0$, $\mathrm{Var}[\eta]=1$ (the "noise"), for example $\eta\sim\mathscr{N}(0,I_D)$. Let $X=\{x_i\}_{i=1}^n$ be a set of uniform (with respect to the natural volume measure on $\mathscr{M}$) independent random samples on $\mathscr{M}$. Our observations $\tilde{X}$ are noisy samples: $\tilde{X}=\{x_i+\sigma\eta_i\}_{i=1}^n$, where η_i are i.i.d. samples from η and where $\sigma>0$. These points may also be thought of as being sampled from a probability distribution $\tilde{\mathscr{M}}$ supported in $\mathbb{R}^D$ and concentrated around $\mathscr{M}$. Here and in what follows we represent a set of n points in $\mathbb{R}^D$ by an $n\times D$ matrix, whose (i,j) entry is the j-th coordinate of the i-th point. In particular X and $\tilde{X}$ are used to denote both the point cloud and the associated $n\times D$ matrices, and N is the noise matrix of the η_i's.

The problem we concern ourselves with is to **estimate** $d=\dim\mathscr{M}$**, given** $\tilde{X}$**.** We shall use multiscale SVD, as described above, and start with an example.

3.1.1 Example: d-dimensional sphere in $\mathbb{R}^D$, with noise

Let $\mathbb{S}^d=\{x\in\mathbb{R}^{d+1}:||x||_2=1\}$ be the unit sphere in $\mathbb{R}^{d+1}$, so $\dim(\mathbb{S}^d)=d$. We embed $\mathbb{S}^d$ in $\mathbb{R}^D$ via the natural embedding of $\mathbb{R}^{d+1}$ in $\mathbb{R}^D$ via the first $d+1$ coordinates. We obtain X by sampling n points uniformly at random from $\mathbb{S}^d$, and $\tilde{X}$ is obtained by adding D-dimensional white Gaussian noise of variance σ in every direction. We call this data set $\mathbb{S}^d(D,n,\sigma)$.

In Figure 2 we consider the multiscale S.V.'s of $\mathbb{S}^9(100,1000,0.1)$, as a function of r. Several observations are in order. First of all, notice that $\mathbb{R}^{d+1}$ is divided into 2^{d+1} sectors, and therefore by sampling 1000 points on $\mathbb{S}^9$ we obtain about 1 point per sector (!). Secondly, observe that the noise size, if measured by $||x_i-\tilde{x}_i||_2^2$, i.e. by how much each point is displaced, would be order $\mathrm{e}[\sigma^2\chi_D^2]\sim 1$, which is comparable with the radius of the sphere itself (!). Therefore this data set may be described as

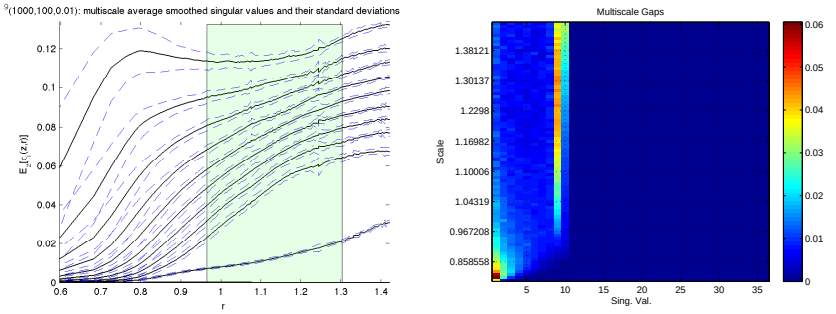

Fig. 1 $\mathbb{S}^9(100,1000,0.01)$. Left: plot of $\mathbb{E}_z[\lambda_{i,z,r}]$, and corresponding standard deviation bands (dotted), as a function of r. The top 9 S.V.'s dominate and correspond to the intrinsic dimensions; the 10-th S.V. corresponds to curvature, and slowly increases with scale (note that at large scales $\Delta_{10} > \Delta_9$); the remaining S.V.'s correspond to noise in the remaining 90 dimensions, and converge to the one-dimensional noise size σ. Right: plot of the multiscale gaps: on the x-axis we have the index i of the gap, and on the vertical axis the scale r. The entry (i,r) is the average (over z) gap $\mathbb{E}_z[\Delta_i(X_{z,r})] := \mathbb{E}_z[\lambda_i(X_{z,r}) - \lambda_{i+1}(X_{z,r})]$. At small scales the noise creates the gaps at the bottom left of the figure; at larger scales we see a large gap at $i = 9$, and at even larger scales that gap is surpassed by the gap corresponding to $i = 10$. This plane is a sort of geometric scale-"frequency" plane, where "frequency" is the index of the singular values.

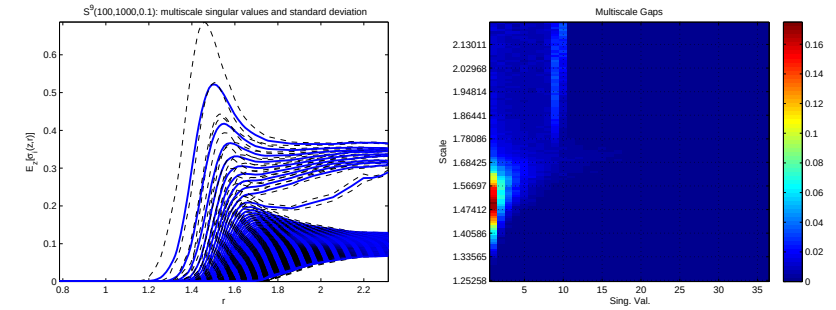

Fig. 2 Same as above, but for $\mathbb{S}^9(100,1000,0.1)$, i.e. 10 times larger noise

randomly sampling one point per sector at distance 1 from the origin in the first $D+1$ coordinates, then moved by 1 in a random direction in $\mathbb{R}^{100}$. The situation may seem hopeless.

In fact, we can detect reliably the intrinsic dimension of $\mathscr{M}$. At very small scales, $B_z(r)$ is empty or contains less than $O(d)$ points, and the rank of $\mathrm{cov}(X_{z,r})$ is even less than d. At small scales, no gap among the $\lambda_{i,z,r}$ is visible: $B_z(r)$ contains too few points, scattered in all directions by the noise, and new increasing S.V.'s keep arising for several scales. At larger scales, the top $d = 9$ S.V.'s start to separate from the others: at these scales the noisy tangent space is detected. At even larger scales, the curvature starts affecting the covariance, as indicated by the slowly growing 10th S.V., while the remaining smaller S.V.'s tend approximately to the *one-dimensional*

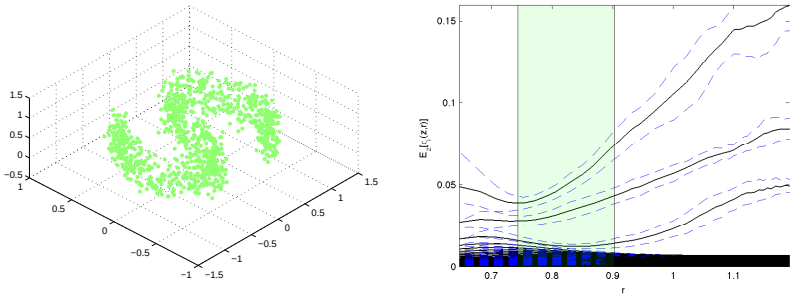

Fig. 3 Left: The S-manifold $\mathscr{S}(100,1000,0.01)$ corrupted by noise. Right: its average multiscale singular values. The green bands are the set of good scales returned by the algorithm.

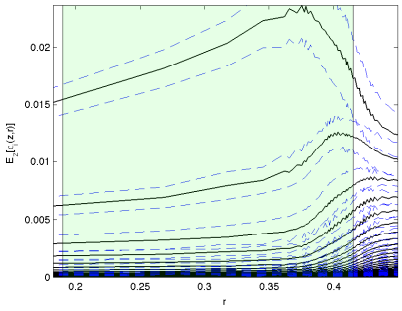

Fig. 4 The average multiscale singular values of the Meyer staircase $\mathscr{Z}^{20}(500,1000,0.05/\sqrt{1000})$ corrupted by noise. The k-th point in this Meyer staircase is a 1000-dimensional vector whose entries from $k+1$ to $k+20$ are equal to 1, and all other entries are 0. The green bands are the set of good scales returned by the algorithm.

noise variance: this is the size of the noise relevant in our procedure, rather than the much larger expected displacement measured in the full $\mathbb{R}^D$, which was of size $O(1)$.

Motivated by applications to large data sets in high-dimensional spaces that are assumed to be intrinsically low-dimensional, we are interested in the regime where D is large, $d << D$, and will ask how large n needs to be in order to estimate d correctly with high probability (w.h.p.). In a classical statistical framework one may rather be interested in the regime where D,d are fixed and n tends to infinity, but in that case one would conduct the analysis as $r \to 0$ and this would lead essentially to the problem of consistency of PCA, and noise would be a relatively minor complication. In many applications D is large and n cannot be taken much larger than D itself: we will therefore be interested in the regime when D and n are large but $\frac{n}{D} = O(1)$.

3.2 Results

In the setting above, we are interested in non-asymptotic results that hold for finite n, d, and D, since they will imply finite sample inequalities w.h.p.

We fix a center z for our computations. Let T_z and N_z be the tangent and normal planes to $\mathscr{M}$ at z. Their dimension is of course d and $D-d$. Let $X_{z,r}{}^{||}$ and $N^{||}$ be the projections of $X_{z,r}$ and N onto T_z, and let $X_{z,r}{}^{\perp}$ and $N^{\perp}$ be the projections onto N_z, so that $X_{z,r} = X_{z,r}{}^{||} + X_{z,r}{}^{\perp}$ and $N = \sigma_{||} N^{||} + \sigma_{\perp} N^{\perp}$. We assume that for each fixed z, there exist parameters $R_{\max}, R_{\min}, \lambda_{\max}, \lambda_{\min}, \kappa$, and $v_{\min}$ (which in general will depend on z) such that for every $r \in [R_{\min}, R_{\max}]$:

$$\begin{aligned} \lambda(\mathrm{cov}(X_{z,r}{}^{||})) &\subseteq d^{-1} r^2 [\lambda_{\min}^2, \lambda_{\max}^2] \\ ||\mathrm{cov}(X_{z,r}{}^{\perp})||_{\mathrm{F}} &\leq \kappa^2 r^4 \\ \mathrm{vol}(X_{z,r}) &\geq v_{\min} \mu_{\mathbb{R}^d}(\mathbb{B}^d) r^d, \end{aligned} \tag{6}$$

where $\lambda(\mathrm{cov}(X_{z,r}{}^{||}))$ denotes the set of eigenvalues of $\mathrm{cov}(X_{z,r}{}^{||})$ and $\mathbb{B}^d$ is the Euclidean unit ball in $\mathbb{R}^d$.

We would like to detect the unknown intrinsic dimension d by estimating a range of "good" scales where the d-th gap is the largest gap. We define

$$\tilde{\Lambda}_{z,r} := \{ r > 0 : \Delta_d(\tilde{X}_{z,r}) = \max_{i=1,\ldots,D} \Delta_i(\tilde{X}_{z,r}) \} \tag{7}$$

Observe that $\tilde{\Lambda}_{z,r}$ is a random set, and we are interested in finding an interval which is contained in $\Lambda_{z,r}$ with high probability given n noisy samples as above.

Proposition 1 ($n \to \infty$). *Assume $\lambda_{\max} = \lambda_{\min} = \lambda$, $r > \sigma_{||}\sqrt{d} + \sigma_{\perp}\sqrt{D}$, $r < \frac{\lambda_{\max}}{\kappa\sqrt{d}}$. Then for n large enough, a sufficient condition for $r \in [R_{min}, R_{max}]$ being in $\tilde{\Lambda}_{z,r}$ is that:*

$$\underbrace{\frac{\lambda^2 r^2}{d}}_{\text{tangent term}} + \underbrace{\sigma_{||}{}^2}_{\text{tangent noise}} \geq \underbrace{2\kappa^2 r^4 + \frac{\lambda \kappa r^3}{\sqrt{d}}}_{\text{curvature term}} + \underbrace{\sigma_{\perp}{}^2}_{\text{normal noise}} + O(n^{-\frac{1}{2}}) \tag{8}$$

As $\sigma_{||}, \sigma_{\perp} \to 0$, this is implied by $r \leq \frac{\lambda}{2\kappa\sqrt{d}}$.

Proposition 2 ($n, D \to \infty, \frac{n}{D} \to \gamma$). *Assume $\lambda_{\max} = \lambda_{\min} = \lambda$, $r > \sigma_{||}\sqrt{d} + \sigma_{\perp}\sqrt{D}$, $r < \frac{\lambda_{\max}}{\kappa\sqrt{d}}$. Then for n, d large enough, and $\gamma = \frac{n}{d}$, a sufficient condition for $r \in [R_{min}, R_{max}]$ being in $\tilde{\Lambda}_{z,r}$ is that:*

$$\underbrace{(\sigma_{||}\sqrt{d} + \sigma_{\perp}\sqrt{D})}_{\text{noise}} \lesssim r \lesssim \underbrace{\frac{\lambda}{\kappa\sqrt{d}}}_{\text{curvature}} \tag{9}$$

Proposition 3 ($D \to \infty$,$\sigma_\perp = \frac{\sigma}{\sqrt{D}}$). *A sufficient condition for $r \in [R_{min}, R_{max}]$ being in $\tilde{\Lambda}_{z,r}$ is that:*

$$\underbrace{(\sigma_{||}\sqrt{d}+\sigma)}_{noise} \vee \underbrace{\left(\frac{d\log(d)}{n}\frac{\mathrm{vol}(\mathscr{M})}{\lambda\ v_{min}\mu_{\mathbb{R}^d}(\mathbb{B}^d)}\right)^{\frac{1}{d}}}_{sampling} \lesssim r \lesssim \underbrace{\frac{\lambda}{\kappa\sqrt{d}}}_{curvature} \tag{10}$$

In fact, in all of the above results, the conditions are sufficient not only in the limit, but with high probability for finite values of n, D. These results are more technical and the interested reader is referred to [65]. They essentially imply, in this context, that as soon as $n_r := |X_{z,r}| \gtrsim_{\kappa,\sigma_{||},\sigma_\perp,v_{\min}} d\log d$ for $r < R_{\max}$, then $\tilde{\Lambda}_{z,r}$ is non-empty with high-probability. An efficient algorithm for finding r's in $\tilde{\Lambda}_{z,r}$ is also developed in [65], and tested against the leading competitors (see the following section). Finally, the setting in [65] is much more general than the one presented above; in particular no manifold assumption is made. Instead, the existence of a set of scales is assumed, at which the data set looks d-dimensional plus smaller detail structure and noise.

In the special case of $\mathbb{S}^d$, if we have no noise and $n \to \infty$ one can show that:

$$\begin{aligned}
\lambda^2_{i,z,r} &= \frac{1}{d+2}r^2 + O(r^4) && \text{for } 1 \le i \le d \\
\lambda^2_{d+1,z,r} &= \frac{d}{(d+2)^2(d+4)}r^4 + O(r^6) \\
\lambda^2_{i,z,r} &= 0 && \text{for } i > d+1
\end{aligned}$$

So here, $\lambda_{\max} = \lambda_{\min} = \frac{d}{d+2} \sim 1$ and $\kappa \sim \frac{1}{d}$. Although on first glance it appears that Prop 1 gives us that Δ_d is the largest gap when $r \lesssim \sqrt{d}$, this is in fact not the case since $R_{\max}$ (the upper bound on the region where the curvature S.V. $\lambda_{d+1,z,r}$ grows quadratically with respect to the tangent S.V.'s) is actually small: $R_{\max} = O(1)$.

In all of these results a "curse of intrinsic dimension" is visible. By "curse of dimensionality" one usually means the large number of samples needed for estimating functions of many parameters. For example, if one tries to approximate a continuous function on the d-dimensional unit cube $\mathbb{Q}^d$ up to precision ε, one needs in general one sample in every little d-dimensional cube of side ε contained in $\mathbb{Q}^d$: the number of such cubes is ε^{-d}, which is large as soon as ε is not large and d is not small (for example: if $\varepsilon = 10^{-3}$ and $d = d$, one would need 10^{3d} samples to approximate the function up to only 3 digits). From the geometric perspective, the curse of dimensionality may manifest itself in terms of concentration of measure phenomena. In our particular situation, for example, covariance matrices of intrinsically high-dimensional objects tend to be small, and therefore easily corrupted by noise. For example, the covariance matrix of the $d-1$-dimensional unit sphere $\mathbb{S}^{d-1}$ is $\frac{1}{d}I_d$ (and not I_d as one may have expected). In particular, if Gaussian noise $\sigma\mathscr{N}(0, I_d)$ is added to points sampled on $\mathbb{S}^{d-1}$, then the covariance "signal to noise

ratio" is $d^{-\frac{1}{2}}/\sigma$, which goes to 0 as $d \to +\infty$. In the last two Propositions one sees this curse in the upper bound on the right hand side, which contains the factor $d^{-\frac{1}{2}}$. However, notice that if $\kappa \sim d^{-\frac{1}{2}}$, such bounds become independent of d. In other words, the curse of intrinsic dimensionality, in this context, is not only rather mild, but disappears by decreasing the curvature κ as the intrinsic dimension increases. This is a analogue of sort to assuming smoothness dependent on d to break the curse in approximating functions in high-dimensions, an interesting approach taken in functional analysis, approximation theory and statistics.

3.3 Algorithm

$[\hat{d}, \hat{R}_{\min}, \hat{R}_{\max}]$ = `EstDimMSVD` $(\tilde{X_n}, K)$

// **Input:**
// $\tilde{X_n}$: an $n \times D$ set of noisy samples
// K : upper bound on the intrinsic dimension k
// **Output:**
// $\hat{d}$: estimated intrinsic dimension
// $(\hat{R}_{\min}, \hat{R}_{\max})$: estimated interval of good scales

Nets = MultiscaleNets($\tilde{X_n}, K$)
$\lambda_{K+1,z,r}$ = FindLargestNoiseSingularValue($\tilde{X_n}$,Nets)
$\hat{R}_{\min}$ = Smallest scale for which $\lambda_{K+1,z,r}$ is decreasing and $|B_z(\hat{R}_{\min})| \gtrsim K \log K$
$\hat{R}_{\max}$ = Largest scale for which $\lambda_{1,z,r}$ is nonincreasing
$\hat{k}$ = Largest i such that:

- for $r \in (\hat{R}_{\min}, \hat{R}_{\max})$, $\lambda_{i,z,r}$ is linear and $\lambda_{i+1,z,r}$ is quadratic in r, and
- $\Delta_i^{(z,r)}$ is largest gap for r in a large fraction of $(\hat{R}_{\min}, \hat{R}_{\max})$

$(\hat{R}_{\min}, \hat{R}_{\max})$ = Largest interval in which $\Delta_{\hat{d}}^{(z,r)}$ is the largest gap

Fig. 5 Pseudo-code for the Intrinsic Dimension Estimator based on multiscale SVD.

The results above suggest the following algorithm: for each $z \in \mathscr{M}$, $r > 0$, $i = 1, \ldots, D$, we compute $\lambda_{i,z,r}$. When r is large, if $\mathscr{M}$ is contained in a linear subspace of dimension K ($K \geq d$) we will observe K large eigenvalues and $D - K$ smaller noise eigenvalues, in the regime for the values of K, D, σ, n suggested by our results. Clearly, $d \leq K$. Moreover, $\{\lambda_{i,z,r}\}_{i=K+1,\ldots,D}$ will be highly concentrated and we use them to estimate σ, which is useful per se. By viewing $\{\lambda_{i,z,r}\}_{i=K+1,\ldots,D}$, we identify an interval in r where the noise is almost flat, i.e. we remove the small scales where the distortion due to noise dominates.

We look at the first $\{\lambda_{i,z,r}\}_{i=1,\ldots,K}$, and the goal is to decide how many of them are due to the extrinsic curvature of $\mathscr{M}$. But the curvature S.V.'s grow quadratically

w.r.t. the "tangential" (non-curvature) S.V.'s: a best least-square linear and quadratic fit to $\lambda_{i,z,r}$, as a function of r, is enough to tell the curvature S.V.'s from the tangential S.V.'s.

MATLAB code and a User Interface for navigating the multiscale S.V.'s are available at www.math.duke.edu/~mauro.

3.4 Examples

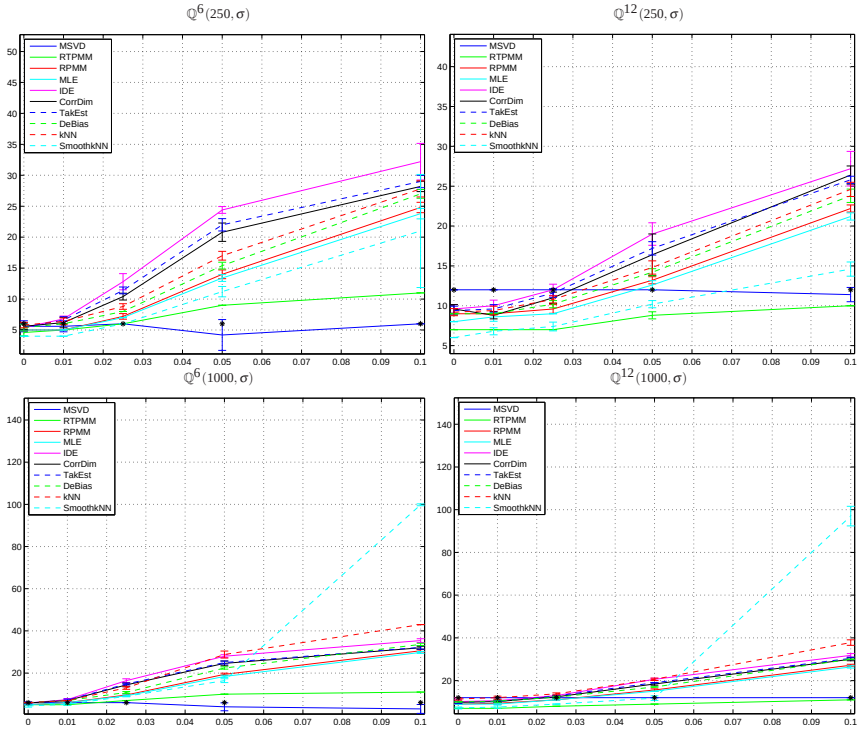

Fig. 6 Benchmark data sets: cube. The horizonal axis is σ, the one-dimensional standard deviation of the noise, the vertical axis is the estimated dimension. Black dots mark the correct intrinsic dimension.

We test our algorithm on several data sets obtained by sampling manifolds, and compare it with existing algorithms. The test is conducted as follows. We fix the ambient space dimension to $D = 100$. We let $\mathbb{Q}^d$, $\mathbb{S}^d$, $\mathscr{S}$, $\mathscr{Z}^d$ be, respectively, the unit d-dimensional cube, the d-dimensional sphere of unit radius, a manifold product of an S-shaped curve of roughly unit diameter and a unit interval, and the Meyer's staircase $\{\chi_{0,d}(\cdot - l)\}_{l=0,\ldots,D}$. Each of these manifolds is embedded isometrically in

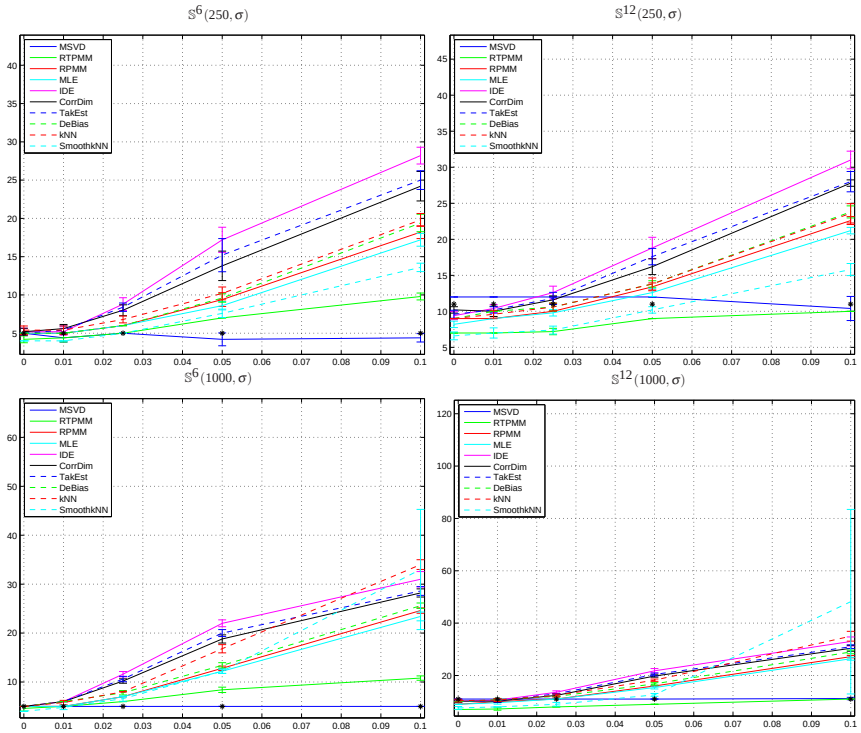

Fig. 7 Benchmark data sets: sphere. The horizonal axis is σ, the one-dimensional standard deviation of the noise, the vertical axis is the estimated dimension. Black dots mark the correct intrinsic dimension.

$\mathbb{R}^K$, where $K = d$ for $\mathbb{Q}^d$, $K = d+1$ for $\mathbb{S}^d$, $K = 3$ for $\mathscr{S}$ and $K = D$ for $\mathscr{Z}^d$, and $\mathbb{R}^K$ is embedded naturally in $\mathbb{R}^D$. Finally, a random rotation is applied (this is irrelevant since all the algorithms considered are invariant under isometries). We draw n samples uniformly (with respect to the volume measure) at random from each manifold, and add noise $\eta \sim \frac{\sigma}{\sqrt{D}}\mathscr{N}(0, I_D)$. We incorporate these parameters in the notation by using $\mathbb{Q}^d(n,\sigma)$ to denote the set of n samples obtained as described above, where the manifold is the d dimensional unit cube and the noise has variance σ. We also consider a variation of these sets, where we dilate $\mathbb{R}^K$ (after embedding the manifold but before any other operation) by a diagonal dilation with factors drawn at random from the multiset $\{1,1,1,1,0.9,0.9,0.9,0.8,0.8\}$.

We consider here $d = 6,12,24,48$ for $\mathbb{Q}^d$ and $\mathbb{S}^d$, $d = 10,20,50$ for $\mathscr{Z}^d$. The samples size is set as $n = 250,500,1000,2000$. We let the noise parameter $\sigma = 0,0.1,0.25,0.5,1,1.5,2$. For each combination of these parameters we generate 5 realizations of the data set and report the most frequent (integral) dimension returned by the set of algorithms specified below, as well as the standard deviation of the estimated dimension. We test the following algorithms: "Debiasing" [16], "Smooth-

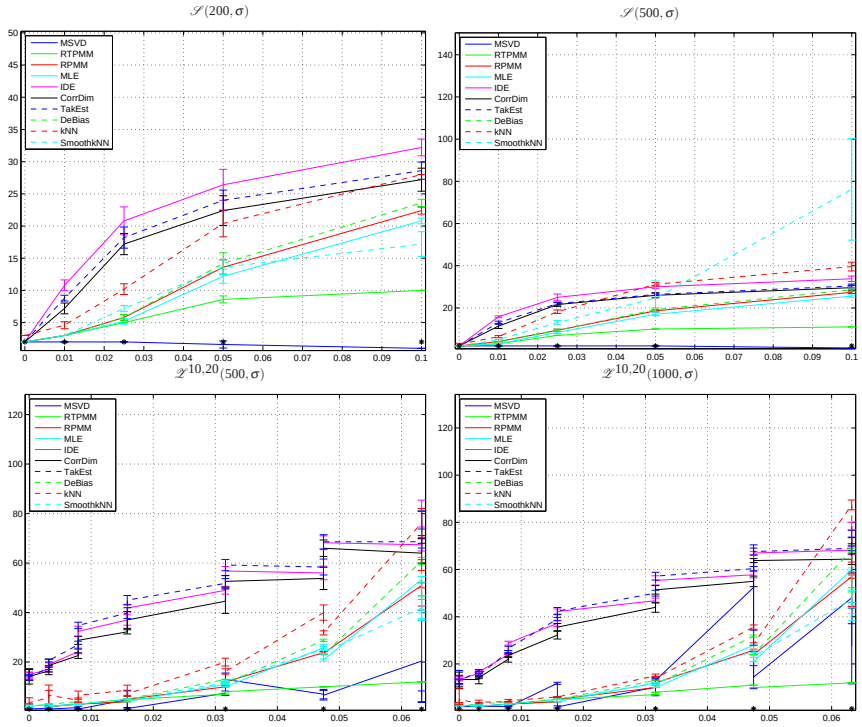

Fig. 8 Benchmark data sets: S-shaped manifold $\mathscr{S}$ and Meyer's staircase $\mathscr{Z}$. The results for $\mathscr{Z}^{20}$ are consistently better than those for $\mathscr{Z}^{10}$, once the number of points and the level of noise is fixed. This is consistent with the fact that $\mathscr{Z}^{20}$ has a smaller effective curvature than $\mathscr{Z}^{10}$.

ing" [18], RPMM in [51], "MLE" [62], "DeBias" [17], "kNN" [33], "SmoothKNN" [19], as well as the classical Correlation Dimension and Taken estimator [85, 46]. The "MFA" label in table 1 refers to one of the state-of-art Bayesian approaches to dimension estimation [26]. For MFA, the authors of [26] ran the code, given only the information that no data set would have intrinsic dimension larger than 100 (this is essentially the only parameter in our algorithm, and it is used only for speeding up computations); in all the other cases we ran the code ourselves, after finding a reasonable range of the parameters that worked on toy examples. There is a large disparity in the number of parameters in these algorithms, ranging from 1 to 7. We will make the data sets publicly available at www.math.duke.edu/˜mauro so that other researchers may try their algorithms (present and future) on a common set of examples.

Finally, in 9 we consider a data set with different dimensionality in different regions, and run the algorithm pointwise. We show both the pointwise estimated dimensionality, and the maximal value of r the algorithm returns as a good scale.

Table 1 This table contains the dimension estimates for a quite benign regime with 1000 samples and no noise. Even in this straightforward setting the estimation of dimension is challenging for most methods.

	RTPMM	RPMM	MLE	IDE	CorrDim	TakEst	DeBias	kNN	SmoothkNN	MFA	MSVD
$\mathbb{Q}^6$	5	5	5	**6**	5	**6**	**6**	**6**	4	2	**6**
$\mathbb{Q}^{12}$	7	9	9	10	10	10	10	**12**	7	4	**12**
$\mathbb{Q}^{24}$	9	16	16	17	17	17	17	20	11	1	**24**
$\mathbb{Q}^{48}$	11	26	25	29	28	28	27	32	19	2	**48**
$\mathbb{Q}^6$	5	5	5	**6**	**6**	**6**	**6**	**6**	5	1	**6**
$\mathbb{Q}^{12}$	7	9	9	10	10	10	10	**12**	7	2	**12**
$\mathbb{Q}^{24}$	9	16	15	17	17	17	17	20	11	3	**24**
$\mathbb{Q}^{48}$	11	26	25	28	27	28	27	31	17	2	**48**
$\mathbb{S}^5$	4	**5**	**5**	**5**	**5**	**5**	**5**	**5**	4	2	**5**
$\mathbb{S}^{11}$	7	9	9	10	10	10	10	10	8	1	**11**
$\mathbb{S}^{23}$	10	16	16	18	18	18	18	18	14	2	24
$\mathbb{S}^{47}$	11	27	26	31	30	31	29	29	21	3	48
$\mathbb{S}^5$	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	4	2	**5**
$\mathbb{S}^{11}$	7	9	9	10	10	10	10	10	8	1	**11**
$\mathbb{S}^{23}$	9	16	16	18	18	18	18	18	13	1	**23**
$\mathbb{S}^{47}$	11	27	26	31	30	30	29	29	21	3	48
$\mathscr{S}$	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	1	**2**
$\mathscr{S}$	**2**	**2**	**2**	**2**	**2**	**2**	**2**	3	**2**	**2**	**2**
$\mathscr{Z}^1$	NaN	NaN	3	340	0	29	3	87	7	4	2
$\mathscr{Z}^1$	NaN	NaN	2	93	0	14	2	67	3	2	**1**
$\mathscr{Z}^1$	NaN	NaN	3	14	12	14	3	3	2	2	**1**
$\mathscr{Z}^1$	NaN	NaN	2	13	13	13	2	5	2	2	**1**

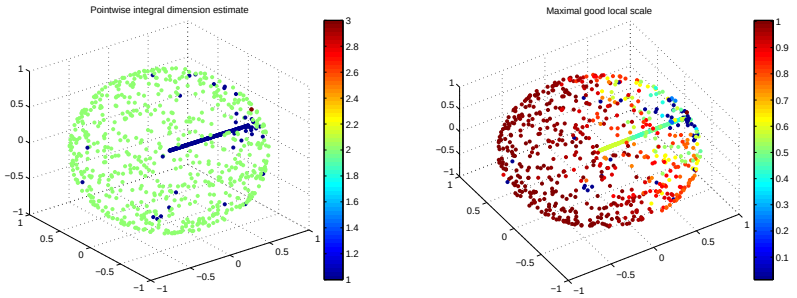

Fig. 9 This data set is a union of two manifolds of different dimensions: a 2-dimensional sphere and a 1 dimensional segment. Left: the estimated pointwise dimensionality. Right: the largest "good" scale returned by the algorithm, for every point. Note how the largest good scale is large for points far from the intersection of the sphere and the segment and decreases as we approach the intersection.

4 Geometric Wavelets

Recent work, both in machine learning, signal processing, image analysis, and harmonic analysis, has focused on either the construction of carefully hand-crafted dictionaries for large classes of data sets (wavelet for 1-D signals, curvelets for certain classes of images, chirplets, etc.), or on dictionaries that are tuned to specific data sets [1, 82, 91] (and references therein). The latter approach typically is formulated by requesting to find a dictionary Φ with I elements, such that every element in the data set may be represented, up to a certain precision ε, by at most m elements of the dictionary. This sparsity requirement of the representation is very natural in statistics, signal processing, and interpretation of the representation. Of course, the smaller I and m are, for a given ε, the better. Current constructions of such dictionaries, such as K-SVD [1], k-flats [82], and Bayesian methods [91], cast these requirements as an optimization problem and rely on black-box optimization to find solutions. Typically no guarantees are provided about the size of I and m (as functions of ε), the computational costs, and the uniqueness of the solution (in practice, it is observed that these algorithms depend heavily on the initial guess). On the other hand, when these algorithms do give solutions that a practitioner considers "good", the use of these data-dependent dictionaries can yield very impressive results in such diverse problems as data modeling, classification, image compression and impainting, and more (e.g. [69] and references therein). Another drawback of existing constructions of data-dependent dictionaries is that the output dictionary is in general highly overcomplete and completely unstructured (even if, at least in the case of images, one may empirically observe certain structures, symmetries and regularities in the dictionary elements). As a consequence, in general there is no fast algorithm for computing the coefficients of the representation of a data point in the dictionary (nor, but less importantly, to sum a long linear combination of dictionary elements), which requires appropriate sparsity-seeking algorithms.

In [24], the authors construct data-dependent dictionaries using a multiscale geometric analysis of the data, based on the geometric analysis in the work [55]. These dictionaries are structured in a multiscale fashion and can be computed efficiently; the expansion of a data point on the dictionary elements is guaranteed to have a certain degree of sparsity m and can be computed by a fast algorithm; the growth of the number of dictionary elements I (as a function of ε) is controlled theoretically and easy to estimate in practice. The elements of these dictionaries are called *geometric wavelets* [24], since in some respects they generalize wavelets from vectors that analyze functions to affine vectors that analyze point clouds. The multiscale analysis associated with geometric wavelets shares some similarities with that of standard wavelets (e.g. fast transforms, a version of two-scale relations, etc.), but is in fact quite different in many crucial respects. It is highly nonlinear, as it adapts to arbitrary nonlinear manifolds, albeit every scale-to-scale step is linear (which is key to efficient computation and fast algorithms); translations or dilations do not play any role here, while they are often considered crucial in classical wavelet constructions. Geometric wavelets may allow the design of new algorithms for manipulating point clouds similar to those used for wavelets to manipulate functions. Dictionaries of

basis functions have a large number of applications in mathematics and engineering [70, 34, 42, 89, 20, 28, 7, 27].

4.1 Construction of Geometric Wavelets

Let $(\mathscr{M},g)$ be a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. We are interested in the regime $d \ll D$. Assume we have n samples drawn i.i.d. from $\mathscr{M}$, according to the natural volume measure $d\text{vol}$ on $\mathscr{M}$ [2]. We construct a multiscale decomposition of the manifold $\mathscr{M}$ as follows.

We start by decomposing $\mathscr{M}$ into multiscale nested partitions $\mathscr{P}_j$. For $j \le J$, let $\mathscr{P}_j = \{C_{j,k}\}_{k\in\Gamma_j}$ be a disjoint cover of $\mathscr{M}$, each $C_{j,k}$ contains a ball of radius $\sim 2^{-j}$, has diameter $\sim 2^{-j}$ and piecewise smooth boundary. Moreover, we assume that every $C_{j,k} = \cup_{k'\in\text{children}(j,k)} C_{j+1,k'}$; this also defines children(j,k). There is a natural tree structure $\mathscr{T}$ associated with this family of partitions. For $x \in \mathscr{M}$, we let $C_{j,x}$ be the unique element of $\mathscr{P}_j$ that contains x. We also note in advance that we will adopt similar notation ($P_{j,x}, \Phi_{j,x}, \Psi_{j,x}$, etc.) for objects associated with $C_{j,x}$. In practice, we use METIS [57] on a nearest-neighbor weighted graph in order to compute the multiscale partitions $\mathscr{P}_j$.

For every $C_{j,k}$ we may compute the top d eigenvalues and eigenvectors of the covariance matrix $\text{cov}_{j,k}$ of the data in $C_{j,k}$, as we did before for the matrix of the data in a ball of radius r centered around a point z. Let $\Phi_{j,k}$ be the $D\times d$ orthogonal matrix of the top d eigenvectors of $C_{j,k}$ and $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$ the corresponding eigenvalues. We think of the span of the columns of $\Phi_{j,k}$ as an approximate tangent space to $\mathscr{M}$ at scale j and location marked by $C_{j,k}$. Let $P_{j,k}$ be the associated affine projection

$$P_{j,k}(x) = \Phi_{j,k}\Phi_{j,k}^*(x-\overline{c}_{j,k}) + \overline{c}_{j,k}\,. \tag{11}$$

where $\overline{c}_{j,k}$ is avg$(C_{j,k})$. We let, for $\forall x \in \mathscr{M}$ and $j \le J$,

$$x_{j;J} \equiv P_{\mathscr{M}_{j;J}}(x) := P_{j,x}\circ P_{j+1,x}\circ\cdots\circ P_{J,x}(x) \tag{12}$$

and define the approximation $\mathscr{M}_{j;J}$ to $\mathscr{M}$ at scale j as the set

$$\mathscr{M}_{j;J} := \cup_{k\in\Gamma_j} P_{\mathscr{M}_{j;J}}(C_{j,k}). \tag{13}$$

Although $\mathscr{M}_{j;J}$ and $x_{j;J}$ depend on the finest scale J, we will from now on drop the J subscript for simplicity of notation. $\mathscr{M}_j$ is a coarse approximation of $\mathscr{M}$ at scale j, analogous to what the projection of a function onto a scaling function subspace is in wavelet theory. Under suitable assumptions, $\mathscr{M}_j \to \mathscr{M}$ in the Hausdorff distance, as $J \to +\infty$.

[2] More general hypotheses on the sampling procedure or on the measure μ are possible but we do not consider them here.

The set $P_{j,k}(C_{j,k})$ of pieces of affine planes centered at $\overline{c}_{j,k}$ and spanned by $\Phi_{j,k}$ is an approximation of the manifold $\mathscr{M}$ at scale j. Just as a scaling function approximation to a function is a coarse version of the function, so this set of pieces of planes is an approximation of $\mathscr{M}$.

We can now construct wavelet planes that span the space needed to complete $\Phi_{j,k}$ into the span of $\{\Phi_{j+1,k'}\}_{k' \in \text{children}(j,k)}$. We leave the description of the construction to [2]. In its simplest form, this construction yields, for every $(j+1,k')$, a $D \times d'_{j+1,k'}$ orthogonal matrix $\Psi_{j+1,k'}$ spanning the subspace $(I - \Phi_{j,k}\Phi_{j,k}^*)\langle \Phi_{j+1,k'} \rangle$. Let $Q_{j+1,k'}$ be the corresponding affine projection: we have the fundamental two-scale relation

$$P_{\mathscr{M}_{j+1}}(x) = P_{\mathscr{M}_j}(x) + Q_{j+1,x}(P_{\mathscr{M}_{j+1}}(x)) \tag{14}$$

for every $x \in C_{j+1,k'}$. By iterating, we obtain a wavelet sum representation of any point $x \in \mathscr{M}$.

The geometric scaling and wavelet coefficients $\{p_{j,x}\}, \{q_{j,x}\}$ of a point $x \in \mathscr{M}$ are defined by the equations

$$x_j = \Phi_{j,x} p_{j,x} + \overline{c}_{j,x}; \tag{15}$$

$$Q_{j+1,x}(x_{j+1}) = \Psi_{j+1,x} q_{j+1,x} + w_{j+1,x}, \tag{16}$$

where $x_j = P_{j,x}(x)$. The computation of the coefficients (and translations), from fine to coarse, is simple and fast. For any $x \in \mathscr{M}_J$ and $j_0 < J$, the set of coefficients

$$\hat{x} = \left(q_{J,x}, q_{J-1,x}, \ldots, q_{j_0+1,x}, p_{j_0,x}\right) \in \mathbb{R}^{d+\sum_{j=j_0+1}^{J} d_{j,x}} \tag{17}$$

is called the discrete geometric wavelet transform of x. Since $d_{j,x} \le d$, $d + \sum_{j>j_0} d_{j,x} \le (J - j_0 + 1)d$.

Observe that we may immediately extend this transform to points not on $\mathscr{M}$, but within the so called reach($\mathscr{M}$), which is the set of points in the ambient space which have a unique closest point in $\mathscr{M}$. This set of points may be thought of as a maximal tube, of variable radius, around $\mathscr{M}$, which does not "self-intersect".

Finally, one may show that the algorithm for constructing the $\Phi_{j,k}$'s and $\Psi_{j,k}$'s only costs $\mathscr{O}(Dn(log(n) + d^2 + k))$ [2].

4.2 Examples

We conduct numerical experiments in this section to demonstrate the performance of the algorithm.

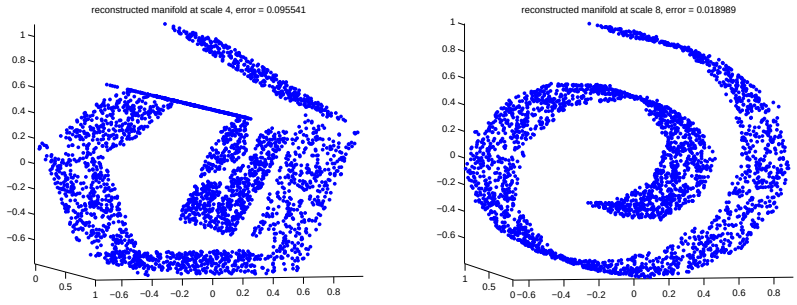

Fig. 10 Geometric wavelet transform of an *S*-shaped manifold, from which 3000 points are randomly sampled. Left: the reconstructed manifold $\mathscr{M}_4$ at scale 4. Right: the reconstructed manifold $\mathscr{M}_8$ at scale 8.

4.2.1 Low-dimensional smooth manifolds

We first consider a simple data set of a 2-dimensional *S*-shaped manifold in $\mathbb{R}^3$ and apply the algorithm to obtain the geometric wavelet transform of the sampled data (3000 points) in Figure 10-11. The resulting wavelet coefficients matrix is very sparse (with about 63% of the coefficients below 1 percent of the maximal magnitude). The reconstructed manifolds also approximate the original manifold well.

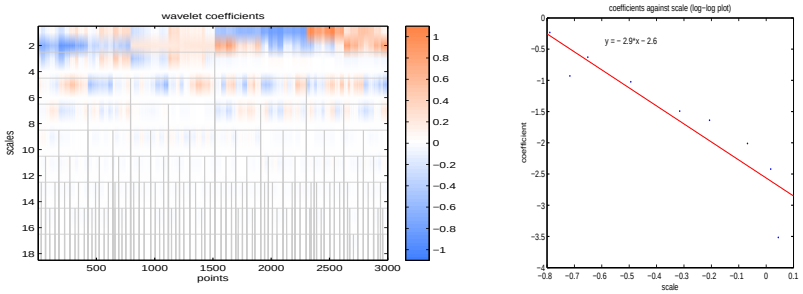

Fig. 11 Left: Geometric wavelet representation of the data. The *x*-axis indexes the points, and the *y* axis indexes the wavelet coefficients as in (17), with the coarsest scale at the top and the finest scale at the bottom. The wavelet subspaces have dimension at most 2, and in fact numerically their dimension is, up to two digits of accuracy, 1. This "matrix" is sparse, with about 37% entries above 10^{-2}. Observe that this "matrix" representation is not an actual matrix, since the set of rows is not in one-to-one correspondence with the dictionary elements, since each cell in the tree has its own local dictionary. Right: Average error in approximating a point on the manifold, as a function of scale (smaller scales on the right).

4.2.2 A data set

We next consider a data set of images from the MNIST data set [3]. We consider the handwritten digit 7. Each image has size 28×28. We randomly sample 5000 such images from the database and then project the samples into the first 120 dimensions by SVD. We apply the algorithm to construct the geometric wavelets and show the reconstructions of the data and the wavelet coefficients at all scales in Figure 12. We observe that the magnitudes of the coefficients stops 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 dimension of the wavelet in future work to lead to an efficient representation also in this case.

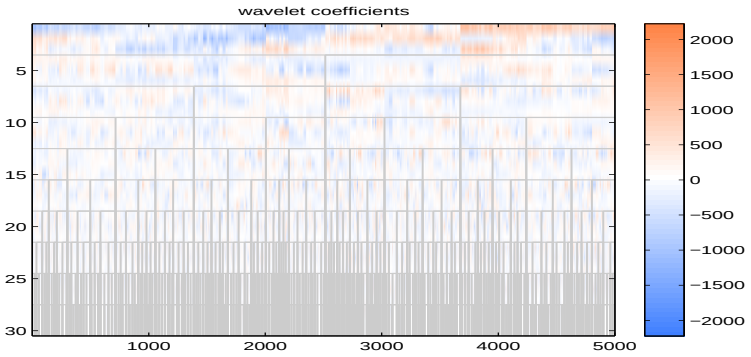

Fig. 12 Geometric wavelet representation of the data for the images of handwritten 7's. This matrix is less sparse than what we would expect for a manifold. This is an artifact of the construction of geometric wavelets we presented here, in which the dimension of the planes $\langle \Phi_{j,k} \rangle$ is chosen independent of j,k. This constraint is not necessary and is removed in [2], which allows one to tune this dimension, as well as the dimension of the wavelet spaces, to the local (in space and scale) properties of the data.

We then fix two data points (i.e. two images) and show in Figure 13 and 14 their reconstructed approximations at all scales and the corresponding wavelet bases (all of which are also images). We see that at every scale we have a handwritten digit, 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.

[3] available, together with detailed description and state-of-art results, at http://yann.lecun.com/exdb/mnist/.

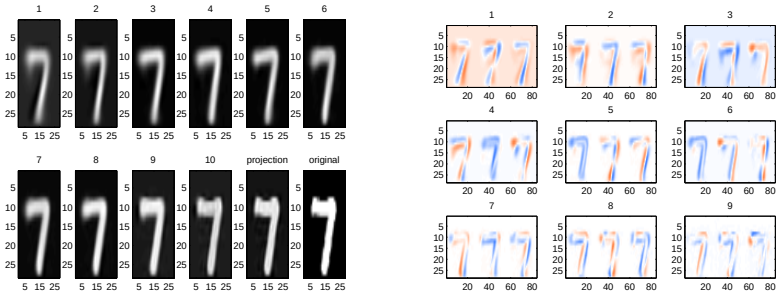

Fig. 13 An image representation of the reconstructed data point and the corresponding subset of the wavelet dictionary. Left: in images 1-10 we plot coarse-to-fine geometric wavelet approximations of the original data point represented in the last two images (projection and original) on the bottom. Right: elements of the wavelet dictionary (ordered from coarse to fine in 1-10) used in the expansion above.

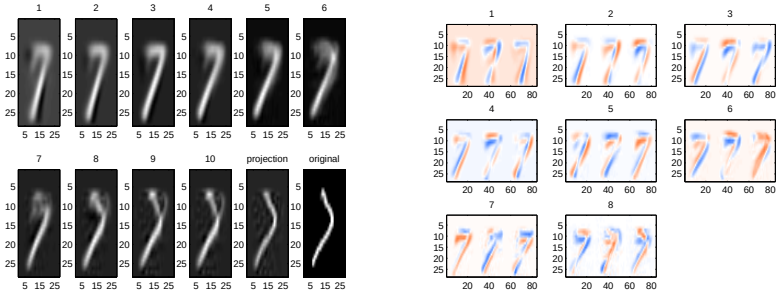

Fig. 14 An image representation of the reconstructed data point and the corresponding subset of the wavelet dictionary. Left: in images 1-10 we plot coarse-to-fine geometric wavelet approximations of the original data point represented in the last two images (projection and original) on the bottom. Right: elements of the wavelet dictionary (ordered from coarse to fine in 1-10) used in the expansion above.

5 Multiple planes

Recent work in machine learning, imaging and harmonic analysis has focused on modeling data, or spaces of signals, as a union of linear subspaces (in some cases, affine). More precisely, assume that the data is generated by sampling (uniformly, say) from

$$\cup_{i=1}^{K} \underbrace{a_i(\mathbb{Q}^{d_i})}_{=:\pi_i} \tag{18}$$

where the a_i are affine maps $\mathbb{R}^{d_i} \to \mathbb{R}^D$ that are invertible on their range. In other words, the data is sampled from affine images of unit cubes - pieces of planes - of different dimensions. These pieces may intersect each other. A typical problem is the

following: given n samples, K and d_i, find a_i. Current approaches are based on black-box optimization [1, 82], "higher-order" spectral clustering methods [48, 22, 23] and references therein, or Monte-Carlo sampling within a Bayesian framework [91]. These cited papers also contain many references to related work. In the first two approaches, K and d_i (or, in some cases, only an upper bound on d_i) are needed as input. In the latter work, experimentation shows that hyper-parameters need to be set very close to the correct values unless the number of points n is very large.

Based on the work on the estimation of intrinsic dimension and determination of "good scales" described in Section 2 (see [65]), we can attack this problem not only without having to know K and d_i, but also with guarantees that given a very small number of samples we are going to determine the correct a_i's. None of the algorithms described above has this type of guarantee with so little information.

We shall assume that each of the K pieces of planes contains at least c/K points, with high probability, where c is some numerical constant. We shall also assume that a fraction of each piece of plane is far from the intersection with other pieces of plane. The algorithm we propose is the following: pick a sample point x_0 at random. We may assume without loss of generality that x_0 belongs to π_0. We run the intrinsic dimension estimator of [65] (see Section 2). If the estimator succeeds, it will return a range of "good scales", as well as an estimate of d_0, which is correct w.h.p.. If the estimator fails, the point was too close to an intersection between π_0 and some other π_i. We simply consider another sample point. By assumption w.h.p. after $O(1)$ samples we will find a sample, which we call again x_0, for which the dimension estimator succeeds. We have therefore found d_0 (say), and from the ball centered at x_0 of radius equal to the largest "good scale", we estimate π_0. At this point we assign to π_0 all the points that are no farther than δ from π_0 (we may choose $\delta = 0$ if there is no noise). We now repeat the above on the remaining points, till all the points have been assigned to some plane. After all the points have been assigned, a polishing step is performed: since we now have all the π_i's, we recompute the assignment of each point to the nearest π_i. Notice that the algorithm will succeed with high probability, as soon as each π_i has a fraction of points far enough from the intersection with other π_j's for which the dimension estimator is going to succeed. Recall that this estimator only requires $\mathscr{O}(d_i \log d_i)$ points in order to assess the dimension. The only remaining issue is the selection of δ, which we perform by estimating the noise variance empirically, as already done in the intrinsic dimension estimator. In fact, an even more robust procedure may be used, based on denoising and clustering the matrix of affinities between points and estimated candidates for the π_i's: such a procedure determines K and the assignment of the points to the π_i's at the same time: the details may be found in [24]. Finally, assuming pre-computation of any structure needed to compute nearest neighbors, and assuming that nearest neighbors may be computed in $O(\log n)$, the computational cost of the algorithm to find the pieces of planes and assign points to them is $\mathscr{O}(n \log n K \max_i d_i^2)$. If the assignment of points to planes is not required, and only the pieces of planes are requested, then randomization allows one to reduce the computational cost to $\mathscr{O}(K^2 \max_i d_i^2)$.

We consider a simple example in Figure 15, which in fact uses a more robust, less greedy version of the algorithm just described [24].

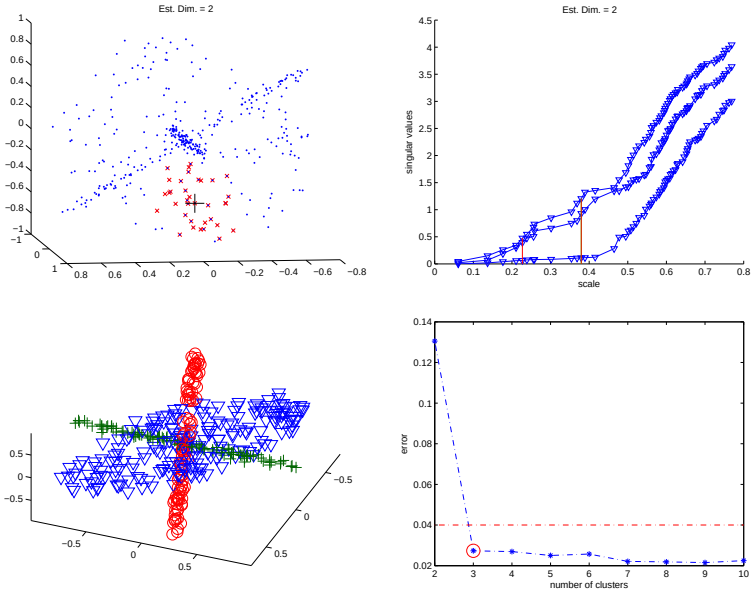

Fig. 15 Global spectral analysis of local subspaces for modeling and clustering hybrid linear data [24]. In this example the given data set consists of points sampled around 2 lines and 1 plane in $\mathbb{R}^3$. We first apply the intrinsic dimension estimator to the data and find a collection of good local scales with dimension estimates at many locations (see one such location (black plus symbol) in the top left figure, with region indicated by the red points) by analyzing the local singular values (see top right, the region between red vertical lines is the range of good scales). We then perform spectral analysis to integrate the local planes so that we recover the model as well as the underlying clusters (see bottom left; the misclassification rate is 2.5%). All we need so far, including for finding the intrinsic dimensions of the underlying planes, is the coordinates of the data plus knowledge of the number of clusters. When the number of clusters is not known, we detect it by looking at the errors associated with different choices (see bottom right; the red line indicates the model error which we can estimate by multiscale SVD).

References

1. M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. In *PROCEEDINGS OF SPARS 05'*, pages 9–12, 2005.
2. W.K. Allard, G. Chen, and M. Maggioni. Multiscale geometric methods for data sets II: Geometric wavelets. *in preparation*, 2010.
3. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 585–591. MIT Press, Cambridge, 2001.
4. M. Belkin and P. Niyogi. Using manifold structure for partially labelled classification. *Advances in NIPS*, 15, 2003.
5. M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(Invited Special Issue on Clustering):209–239, 2004. TR-2001-30, Univ. Chicago, CS Dept., 2001.
6. M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*,

(7):2399–2434, Nov. 2006.
7. J.J. Benedetto and M.W. Frazier eds. *Wavelets, Mathematics and Applications*. CRC Press, 1993.
8. P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *J. Mach. Learn. Res.*, 6:1297–1321, 2005.
9. P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part ii: piecewise polynomial functions. *Constr. Approx.*, 26(2):127–152, 2007.
10. I. Borg and P. Groenen. *Modern Multidimensional Scaling : Theory and Applications*. Springer, 1996.
11. S. Borovkova, R. Burton, and H. Dehling. Consistency of the Takens estimator for the correlation dimension. *Ann. Appl. Probab.*, 9(2):376–390, 1999.
12. J. Bourgain. On Lipschitz embedding of finite metric spaces into Hilbert space. *Isr. Journ. Math.*, pages 46–52, 1985.
13. F. Camastra and A. Vinciarelli. Intrinsic dimension estimation of data: An approach based on grassberger-procaccia's algorithm. *Neural Processing Letters*, 14(1):27–34, 2001.
14. F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE P.A.M.I.*, 24(10):1404–10, 2002.
15. Wenbo Cao and Robert Haralick. Nonlinear manifold clustering by dimensionality. *icpr*, 1:920–924, 2006.
16. K. Carter, A. O. Hero, and R. Raich. De-biasing for intrinsic dimension estimation. *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pages 601–605, Aug. 2007.
17. Kevin M. Carter, Alfred O. Hero, and Raviv Raich. De-biasing for intrinsic dimension estimation. *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pages 601–605, Aug. 2007.
18. K.M. Carter and A.O. Hero. Variance reduction with neighborhood smoothing for local intrinsic dimension estimation. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3917–3920, 31 2008-April 4 2008.
19. K.M. Carter and A.O. Hero. Variance reduction with neighborhood smoothing for local intrinsic dimension estimation. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3917–3920, 31 2008-April 4 2008.
20. Tony F. Chan and Jianhong Shen. *Image processing and analysis*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005. Variational, PDE, wavelet, and stochastic methods.
21. M. Chaplain, M. Ganesh, and I.Graham. Spatio-temporal pattern formation on spherical surfaces: numerical simulation and application to solid tumor growth. *J. Math. Biology*, 42:387–423, 2001.
22. G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Found. Comput. Math.*, 9:517–558, 2009. DOI 10.1007/s10208-009-9043-7.
23. G. Chen and G. Lerman. Spectral curvature clustering (scc). *Int. J. Comput. Vis.*, 81:317–330, 2009. DOI 10.1007/s11263-008-0178-9.
24. G. Chen and M. Maggioni. Multiscale geometric methods for data sets III: multiple planes. *in preparation*, 2010.
25. G. Chen and M.Maggioni. Multiscale geometric wavelets for the analysis of point clouds. *to appear in Proc. CISS 2010*, 2010.
26. M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Processing*, 2010. submitted.
27. Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
28. C.K. Chui. *An introduction to wavelets*. Academic Press, San Diego, 1992.
29. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.
30. R.R. Coifman and S. Lafon. Diffusion maps. *Appl. Comp. Harm. Anal.*, 21(1):5–30, 2006.

31. J. Costa and A.O. Hero. Learning intrinsic dimension and intrinsic entropy of high dimensional datasets. In *Proc. of EUSIPCO*, Vienna, 2004.
32. J.A. Costa and A.O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *Signal Processing, IEEE Transactions on*, 52(8):2210–2221, Aug. 2004.
33. J.A. Costa and A.O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *Signal Processing, IEEE Transactions on*, 52(8):2210–2221, Aug. 2004.
34. I Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.
35. G. David and S. Semmes. *Uniform Rectifiability and Quasiminimizing Sets of Arbitrary Codimension*. AMS.
36. G. David and S. Semmes. Singular integrals and rectifiable sets in $\mathbf{R}^n$: Au-delá des graphes lipschitziens. *Astérisque*, (193):152, 1991.
37. Guy David. `http://www.math.u-psud.fr/~gdavid/Notes-Parkcity.dvi`.
38. Guy David. Morceaux de graphes lipschitziens et intégrales singulières sur une surface. *Rev. Mat. Iberoamericana*, 4(1):73–114, 1988.
39. Guy David. *Wavelets and singular integrals on curves and surfaces*, volume 1465 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1991.
40. Guy David. *Wavelets and Singular Integrals on Curves and Surfaces*. Springer-Verlag, 1991.
41. Guy David and Stephen Semmes. *Analysis of and on uniformly rectifiable sets*, volume 38 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1993.
42. D. L Donoho and Ana G Flesia. Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics? *Network: Comput. Neural Syst.*, 12:371–393, 2001.
43. D. L. Donoho and C. Grimes. When does isomap recover natural parameterization of families of articulated images? Technical Report Tech. Rep. 2002-27, Department of Statistics, Stanford University, August 2002.
44. D. L Donoho and Carrie Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proc. Nat. Acad. Sciences*, pages 5591–5596, March 2003. also tech. report, Statistics Dept., Stanford University.
45. D. L. Donoho, O. Levi, J.-L. Starck, and V. J. Martinez. Multiscale geometric analysis for 3-d catalogues. Technical report, Stanford Univ., 2002.
46. A. M. Farahmand and C. Szepesv?and J.-Y. Audibert. Manifold-adaptive dimension estimation. *Proc. I.C.M.L.*, 2007.
47. A. M. Farahmand, Cs. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, page 265272, 2007.
48. S. Atev G. Chen and G. Lerman. Kernel spectral curvature clustering (kscc). In *The 4th ICCV International Workshop on Dynamical Vision*, Kyoto, Japan, 2009.
49. Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Phys. D*, 9(1-2):189–208, 1983.
50. G. Haro, G. Randall, and G. Sapiro. Translated poisson mixture model for stratification learning. *Int. J. Comput. Vision*, 80(3):358–374, 2008.
51. Gloria Haro, Gregory Randall, and Guillermo Sapiro. Translated poisson mixture model for stratification learning. *Int. J. Comput. Vision*, 80(3):358–374, 2008.
52. X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Trans. pattern analysis and machine intelligence*, 27(3):328–340, 2005.
53. M. Hein and Y. Audibert. Intrinsic dimensionality estimation of submanifolds in euclidean space. In S. Wrobel De Raedt, L., editor, *ICML Bonn*, pages 289 – 296, 2005.
54. Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, 29(2):295–327, April 2001.
55. Peter W. Jones. Rectifiable sets and the traveling salesman problem. *Invent. Math.*, 102(1):1–15, 1990.
56. P.W. Jones. Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae*, 102:1–15, 1990.

57. G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1999.
58. V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Ann. Stat.*, 28(2):591–629, 2000.
59. R. Krauthgamer, J. Lee, M. Mendel, and A. Naor. Measured descent: A new embedding method for finite metrics, 2004.
60. S Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.
61. E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. *In Advances in NIPS 17,Vancouver, Canada*, 2005.
62. Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, Cambridge, MA, 2005.
63. A.V. Little, Y.-M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *Proc. A.A.A.I.*, 2009.
64. A.V. Little, J. Lee, Y.-M. Jung, and M. Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale *SVD*. In *Proc. S.S.P.*, 2009.
65. A.V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets I: Estimation of intrinsic dimension. *in preparation*, 2010.
66. P-C. Lo. Three dimensional filtering approach to brain potential mapping. *IEEE Tran. on biomedical engineering*, 46(5):574–583, 1999.
67. S. Mahadevan, K. Ferguson, S. Osentoski, and M. Maggioni. Simultaneous learning of representation and control in continuous domains. In *AAAI*. AAAI Press, 2006.
68. S. Mahadevan and M. Maggioni. Value function approximation with diffusion wavelets and laplacian eigenfunctions. In *University of Massachusetts, Department of Computer Science Technical Report TR-2005-38; Proc. NIPS 2005*, 2005.
69. Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, page 87, 2009.
70. S.G. Mallat. *A wavelet tour in signal processing*. Academic Press, 1998.
71. Benoit B. Mandelbrot and Richard L. Hudson. *The (mis)behavior of markets*. Basic Books, New York, 2004. A fractal view of risk, ruin, and reward.
72. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm, 2001.
73. P. Niyogi, I. Matveeva, and M. Belkin. Regression and regularization on large graphs. Technical report, University of Chicago, Nov. 2003.
74. Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.
75. M. Raginsky and S. Lazebnik. Estimation of intrinsic dimensionality using high-rate vector quantization. *Proc. NIPS*, pages 1105–1112, 2005.
76. ST Roweis and LK Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
77. M. Rudelson. Random vectors in the isotropic position. *J. of Functional Analysis*, 164(1):60–72, 1999.
78. L.K. Saul, K.Q. Weinberger, F.H. Ham, F. Sha, and D.D. Lee. *Spectral methods for dimensionality reduction*, chapter Semisupervised Learning. MIT Press, 2006.
79. F. Sha and L.K. Saul. Analysis and extension of spectral methods for nonlinear dimensionality reduction. *Proc. ICML*, pages 785–792, 2005.
80. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE PAMI*, 22:888–905, 2000.
81. Jack Silverstein. On the empirical distribution of eigenvalues of large dimensional information-plus-noise type matrices. *Journal of Multivariate Analysis*, 98:678–694, 2007.
82. A. Szlam and G. Sapiro. Discriminative *k*-metrics. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1009–1016, 2009.
83. A.D. Szlam, M. Maggioni, and R.R. Coifman. Regularization on graphs with function-adapted diffusion processes. *JMLR*, (9):1711–1739, Aug 2008.

84. A.D. Szlam, M. Maggioni, R.R. Coifman, and J.C. Bremer Jr. Diffusion-driven multiscale analysis on manifolds and graphs: top-down and bottom-up constructions. volume 5914-1, page 59141D. SPIE, 2005.
85. Floris Takens. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations (Groningen, 1984)*, volume 1125 of *Lecture Notes in Math.*, pages 99–106. Springer, Berlin, 1985.
86. J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
87. M. B. Wakin, D. L. Donoho, H. Choi, and R. G. Baraniuk. The multiscale structure of non-differentiable image manifolds. In *SPIE Wavelets XI*, San Diego, July 2005.
88. K.Q. Weinberger, F. Sha, and L.K. Saul. Leaning a kernel matrix for nonlinear dimensionality reduction. *Proc. ICML*, pages 839–846, 2004.
89. Mladen Victor Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A K Peters Ltd., Wellesley, MA, 1994. With a separately available computer disk (IBM-PC or Macintosh).
90. Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignement. Technical Report CSE-02-019, Department of computer science and engineering, Pennsylvania State University, 2002.
91. M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *Neural and Information Processing Systems (NIPS)*, 2009.
92. Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.