A FAST MULTISCALE FRAMEWORK FOR DATA IN HIGH-DIMENSIONS: MEASURE ESTIMATION, ANOMALY DETECTION, AND COMPRESSIVE MEASUREMENTS*

Guangliang Chen[1], *Mark Iwen*[1], *Sang Chin*[2], *Mauro Maggioni*[1,3]

[1]Mathematics and [3]Computer Science Departments, Duke University, Durham, NC 27708
[2]Applied Physics Laboratory and ECE Department, Johns Hopkins University, Laurel, MD 20723

ABSTRACT

Data sets are often modeled as samples from some probability distribution lying in a very high dimensional space. In practice, they tend to exhibit low intrinsic dimensionality, which enables both fast construction of efficient data representations and solving statistical tasks such as regression of functions on the data, or even estimation of the probability distribution from which the data is generated. In this paper we introduce a novel multiscale density estimator for high dimensional data and apply it to the problem of detecting changes in the distribution of dynamic data, or in a time series of data sets. We also show that our data representations, which are not standard sparse linear expansions, are amenable to compressed measurements. Finally, we test our algorithms on both synthetic data and a real data set consisting of a times series of hyperspectral images, and demonstrate their high accuracy in the detection of anomalies.

Index Terms— High Dimensional Data Sets, Measure Estimation, Anomaly Detection, Dictionary Learning, Hyperspectral Imaging, Compressive Sensing

1. INTRODUCTION

We study the *geometry* and *distribution* of high-dimensional data sets, and the relationships between the two. Here we model data as independent samples $X_n = \{x_1, \ldots, x_n\}$ from a probability measure μ in $\mathbb{R}^D$. We will assume that μ may be well-approximated by a measure which has support on a set of dimension $d \ll D$; this assumption is justified by many observations, empirical and, in some cases, theoretical (see [1] and references therein). This setting has been considered in much existing work on dimensionality reduction, where the (high dimensional) ambient space is compressed to a lower dimension, under a constraint of small distortion of the distances between data points, or in manifold learning research, where one seeks a parametrization of the data with a small number of parameters (ideally $O(d)$, see [2]). Another approach is that of working directly in the high-dimensional space, while using appropriate constructions to exploit the low intrinsic dimension [3].

The problems described above, and corresponding algorithms, focus on approximating the data itself. In this paper we take on the more ambitious task of approximating the probability measure μ, aiming at bypassing the curse of dimensionality in $\mathbb{R}^D$ by employing the assumption that μ is supported near a low-dimensional subset. However, the estimators for μ that we propose are based on a geometry-driven multiscale decomposition of the data and a data-driven dictionary. The basic idea, inspired by multiscale regression methods (where the object to be estimated is a function on the data), is that probability measures in the pieces of the geometric multiscale decompositions are drawn from some simple family, in order to construct an estimator of the true underlying probability distribution. We therefore fuse *geometry, multiscale analysis, dictionary learning*, and *measure estimation*. We introduce a multiscale procedure for producing an estimator $\hat{\mu}_{X_n}$ of μ, with respect to a Wasserstein metric between probability measures, with finite sample size guarantees and a sample complexity that can take advantage of the low intrinsic dimension assumption in a very controlled fashion. We also exhibit fast, parallel algorithms for computing such an estimator from sampled data. In fact, our results apply to a wide class of estimators sharing a common multiscale-geometric structure. We know of no existing results in the literature comparable to this, in terms of flexibility, finite sample guarantees and computational speed.

We then consider the case where we have a *time-varying* family of probability measures, $\{\mu_t\}_{t\geq 0}$, and we sample, for each t, a set of i.i.d. samples $X_n^{(t)} = \{x_1^{(t)}, \ldots, x_n^{(t)}\}$ from μ_t, independently for different t. We may construct an estimator $\hat{\mu}^{(t)}$ based on the sample $X_n^{(t)}$, and quantify the change of such measures with respect to a distance between probability measures, for example a Wasserstein distance. We can then use our estimators $\hat{\mu}^{(t)}$ to detect changes in the distribution of the data, for instance to monitor the development of anomalies: we may suppose that for t in a certain range T_{norm}, say $[0, \epsilon]$, there are no anomalies in the data, and use the learned measures $\{\hat{\mu}^{(t)}\}_{t\in T_{\text{norm}}}$ to test whether the subsequent observations $X_n^{(t)}$ contain anomalies, based on a likelihood test.

*THIS RESEARCH WAS PARTIALLY SUPPORTED BY DARPA MSEE FA8650-11-1-7150, ONR N00014-07-1-0625, NSF-DMS-08-47388.

Lastly, we show that, while our dictionaries and data representations do not quite fit into the standard setting of sparse representations, compressed-sensing-like results exist in our setting as well. This is because the fundamental paradigm that the data has low intrinsic complexity still holds, which leads to very fast algorithms (thanks to the multiscale organization of the data) that do not require convex optimization.

We test our algorithms both on synthetic data - a time series of manifolds developing a singularity in the form of a cusp at an unknown location, and on real world data - a sequence of hyperspectral images of a desert scene with a chemical release at some moment. In both examples our algorithms can both quickly and accurately locate the anomalies, as well as determine the correct scale at which the anomalies occur.

2. GEOMETRIC MEASURE ESTIMATION

We propose an algorithm for estimating a probability measure μ lying around a low dimensional set $\mathcal{M}$ embedded in a high dimensional ambient space $\mathbb{R}^D$, based on a finite sample X_n from μ. The estimation of μ proceeds in two steps: first, an adaptive tree-based geometric decomposition of the data set X_n is constructed; second, this tree is used to construct a multiscale family of estimators of μ, and a scale optimally balancing certain bias and variance terms is selected.

2.1. Review of Geometric Multi-Resolution Analysis

Our approach builds on the geometric multi-resolution analysis (GMRA) framework recently introduced by Allard et al. [3], which aims to learn data-dependent dictionaries on general point-cloud data. When the input data lies around a low dimensional manifold, GMRA has guarantees on the dictionary size and the sparsity of the representations for a given approximation error, on the computational complexity of the construction, and on the associated fast transforms mapping data points to sparse coefficients and vice versa.

The construction of GMRA starts by a multiscale nested decomposition of the data set X_n into a collection of subsets $\{C_{j,k}\}_{0\le j\le J, k\in\Gamma_j}$, arranged in a tree structure $\mathcal{T}$. Each node of the tree is one subset in the collection, indexed by two integers (j,k), with j representing the depth (or scale) of the node in $\mathcal{T}$ (the root of the tree has scale $j=0$ by convention) and $k\in\Gamma_j$ indexing nodes of $\mathcal{T}$ at that scale. For any fixed j, the collection $\{C_{j,k}\}_{k\in\Gamma_j}$ provides a disjoint partition of X_n. For $j>0$ each $C_{j,k}$ points to a unique parent node $C_{j-1,k'}$ containing $C_{j,k}$, and conversely any $C_{j,k}\subseteq C_{j-1,k'}$ is called a child of $C_{j-1,k'}$ (a node without child is called a leaf node). At every $C_{j,k}$ with $j>0$, one computes the following:

1. $c_{j,k}$: the (empirical) mean of the points in $C_{j,k}$;
2. $\Phi_{j,k}$: an orthogonal matrix whose columns are the top principal components of the data in $C_{j,k}$. The hyperplane spanned by the column vectors of $\Phi_{j,k}$ and passing through $c_{j,k}$ is a local linear approximation to $\mathcal{M}$;
3. $w_{j,k}$: wavelet constant associated to $C_{j,k}$, defined as $w_{j,k}=(I-\Phi_{j-1,k'}\Phi_{j-1,k'}^T)\cdot(c_{j,k}-c_{j-1,k'})$, where $C_{j-1,k'}$ is the parent of $C_{j,k}$;
4. $\Psi_{j,k}$: an orthogonal matrix whose column vectors form a basis for the projection of $\mathrm{span}(\Phi_{j,k})$ onto the orthogonal complement of $\mathrm{span}(\Phi_{j-1,k'})$ in $\mathbb{R}^D$. This implies that $\mathrm{span}(\Phi_{j,k})\subseteq\mathrm{span}([\Phi_{j-1,k'}\ \Psi_{j,k}])$, with $\Psi_{j,k}$ containing fewest columns among those meeting the same requirement.

At the root of the tree, only the mean $c_{j,k}$ and basis $\Phi_{j,k}$ are computed. The columns of the orthogonal matrices $\Phi_{j,k}, \Psi_{j,k}$ are called *geometric scaling and wavelet bases*, respectively. Collectively, the four fields at all nodes of the tree comprise the GMRA. For any $x\in C_{j,k}$, a coarse approximation of x at the associated scale j is given by $x_j=\Phi_{j,k}\Phi_{j,k}^T(x-c_{j,k})+c_{j,k}$. If $j>0$ and $C_{j,k}\subseteq C_{j-1,k'}$, the geometric wavelet basis $\Psi_{j,k}$ and constant $w_{j,k}$ provide a bridge between approximations of x at scales j and $j-1$:

$$x_j - x_{j-1} = (\Phi_{j-1,k'}\ \Psi_{j,k})\begin{pmatrix}\epsilon_{j,k}\\ q_{j,k}\end{pmatrix} + w_{j,k}, \qquad (1)$$

where $q_{j,k}$ is the so-called wavelet coefficient[1]

$$\epsilon_{j,k}=\Phi_{j-1,k'}^T(x-x_j) \quad , \quad q_{j,k}=\Psi_{j,k}^T(x_j-c_{j,k}). \qquad (2)$$

Iterating (1) for j varying on the scales of the tree, one obtains a multiscale transform of the point x, in terms of the differences x_j-x_{j-1}. Though such differences are high dimensional, they are decomposed along low dimensional subspaces (after subtracting constant terms). For more details we refer the reader to [3]. This framework provides the foundation for modeling high dimensional densities below.

2.2. Multiscale Measure Estimation

Since our target probability measure μ lives in $\mathbb{R}^D$ for D large, in general it is not feasible to model it directly due to the curse of dimensionality. However, if we assume that μ is supported near a low dimensional subset $\mathcal{M}$, we show in the following that we may estimate it efficiently by enriching the GMRA construction described in the previous section. Given X_n as above, we are interested in fast algorithms yielding an estimator $\hat{\mu}_{X_n}$ of μ, which is a random probability measure that we would like to be close, as a function of n and "regularity" assumptions on μ and with high probability, to μ. We need several ingredients, which we now detail.

Metric in $M^1(\mathbb{R}^D)$, the space of Borel probability distributions in $\mathbb{R}^D$. In order to measure distances between distributions in $M^1(\mathbb{R}^D)$, we shall use the p-Wasserstein distances

$$W_p(\nu_1,\nu_2):=\inf_{\pi\in\mathcal{C}(\nu_1,\nu_2)}\left(\int_{\mathbb{R}^D\times\mathbb{R}^D}||x-y||_{\mathbb{R}^D}^p\,\mathrm{d}\pi(x,y)\right)^{1/p},$$

[1]The coefficient $\epsilon_{j,k}$ corresponds to a correction along $\mathrm{span}(\Phi_{j-1,k'})$. In one variation of the GMRA, this coefficient can be removed [3, Sec. 6.2].

where $\mathcal{C}(\nu_1, \nu_2)$ is the set of couplings between ν_1 and ν_2, i.e., the set of measures π on $\mathbb{R}^D \times \mathbb{R}^D$ such that the marginals of π are, respectively, ν_1 and ν_2: $\pi(A \times \mathbb{R}^D) = \nu_1(A)$ and $\pi(\mathbb{R}^D \times A) = \nu_2(A)$ for all measurable subsets $A \subseteq \mathbb{R}^D$ [4]. We use these distances as they allow for comparisons between measures supported on sets of different dimensions, i.e. ν_1 may be supported near (but not exactly on) a low-dimensional set while ν_2 is supported exactly on a low-dimensional set. In our setting ν_1 is the measure μ generating the data, which may not have exactly low-dimensional support (e.g. due to model error and/or high-dimensional noise), and ν_2 is our estimator $\hat{\mu}_{X_n}$, often supported exactly on a low-dimensional set.

Local model classes $\mathcal{F} \subset M^1(\mathbb{R}^D)$. Our density estimation procedure will work by partitioning the (effective) support of μ in a treelike fashion. Each node of this tree will correspond to a subset, I, of the support of μ, and will have a corresponding local estimator for μ restricted to I. These local estimators will all belong to the local model class $\mathcal{F}$. For example, we might choose $\mathcal{F}$ to be the set of all uniform distributions on the unit cubes of the d-dimensional subspaces of $\mathbb{R}^D$, or the set of all (truncated and rescaled) Gaussian distributions with rank d covariance matrices:

$$\mathcal{F}_{d,U} := \{\mu = U([0,1])^d \cdot \Phi, \text{ for any } \Phi \in \mathbb{R}^{d \times D}\};$$
$$\mathcal{F}_{d,\mathcal{N}} := \{\mu = c_\Sigma \mathcal{N}(m, \Sigma)\mathbf{1}_{\{x:(x-m)^T\Sigma(x-m)\leq 1\}}, \text{ with } \operatorname{rank}(\Sigma) = d \text{ and } c_\Sigma \text{ such that } ||\mu|| = 1\}.$$

We think of $\mathcal{F}$ as a collection of simple measures, or "building blocks", with low-complexity (e.g., low dimensionality).

Geometric multiscale models. Note that $\mathcal{F}$ may contain only simple measures. Thus, in general, all geometric information is effectively supplied by a partition tree $\mathcal{T}$. This allows us to construct more interesting elements of $M^1(\mathbb{R}^D)$ by combining probability measures that are locally, i.e. in nodes of $\mathcal{T}$, in our family of models while having significantly more complicated global geometry.

For a partition Λ consisting of elements of $\mathcal{T}$ we define

$$P_{\Lambda,I}(\mu) = \begin{cases} \mu|_I & \text{if } \mu(I) = 0, \\ \operatorname{argmin}_{\nu \in \mathcal{F}} W_p\left(\nu, \frac{\mu|_I}{\mu(I)}\right) & \text{else} \end{cases}$$

for each $I \in \Lambda$, where $\mu|_I(A) = \mu(A \cap I)$ for all (measurable) $A \subseteq \mathbb{R}^D$. We then define $P_\Lambda(\mu) = \sum_{I \in \Lambda} \mu(I) P_{\Lambda,I}(\mu)$. We consider here the case $\Lambda = \Lambda_j$ for some j, where $\Lambda_j = \{C_{j,k}\}_k$ is the GMRA partition at scale j.

Complexity constant for $\mathcal{F}$. We also utilize a complexity parameter ζ which bounds the number of samples required for accurate approximation of a measure projected onto $\mathcal{F}$, with high probability. More exactly, we require that a *concentration inequality for empirical approximations* holds for $\mathcal{F}$, i.e., that there exists a *sample value* $\zeta > 0$, and constants $C_1, C_2, C_3 > 0$ such that the following holds for all $\mu \in M^1(\mathbb{R}^D)$ and samples X_n from μ^n. Let $\mu_{X_n} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the (random) empirical measure based on X_n, and $n_{X_n,I} = |X_n \cap I|$ for any $I \subset \mathbb{R}^D$. We assume that

$$\mathbb{P}\left(W_2\left(\frac{P_\Lambda(\mu_{X_n}|_I)}{\mu_{X_n}(I)}, \frac{P_\Lambda(\mu|_I)}{\mu(I)}\right) > \sqrt{\frac{\zeta}{n_{X_n,I}}}\, t\right) \leq C_1 e^{-C_2 t^2}$$

for all $t > 1$, $I \in \Lambda$, and partitions $\Lambda = \Lambda_j$ with $j \leq C_3 \ln n$.

Approximation spaces. We define our uniform approximation spaces in $M^1(\mathbb{R}^D)$ as follows. We let $E_j(\mu) = W_p(P_{\Lambda_j}(\mu), \mu)$. Our approximation space $\mathcal{A}^s$ consists of those $\mu \in M^1(\mathbb{R}^D)$ for which there is a constant C such that $E_j(\mu) \leq C\,(\#\Lambda_j)^{-s}$ for all $j \geq 0$. The infimum C for which this equation holds defines a seminorm $|\mu|_{\mathcal{A}^s}$ on $\mathcal{A}^s$.

The measure estimator $\hat{\mu}_{X_n}$ will essentially be $P_{\Lambda_j}(\mu_{X_n})$ for an appropriately chosen j. As j increases our model space becomes larger (albeit not necessarily in a monotonic fashion) and therefore it contains elements closer to μ. However, both the model complexity and variance of $\hat{\mu}_{X_n}$ grow with this increased flexibility. A bias-variance tradeoff balancing these two components allows us to pick an optimal scale j at which estimation should occur. Under a few additional technical assumptions one can prove the following result [4].

Theorem 2.1. *Let $\mathcal{T}$ be a fixed partition tree of $\mathcal{M} \subset \mathbb{R}^D$, μ a probability measure with $\mu(\mathcal{M}) = 1$, and $\mathcal{F}$ a model class, as above. Assume that $\mu \in \mathcal{A}^s$, and let ζ be the complexity constant for $\mathcal{F}$. Then for any $\beta > 0$, there exists a constant $C_{\mathcal{F},\beta}$ such that if j is the smallest index with*

$$\#\Lambda_j \geq C_{\mathcal{F},\beta}\left(\frac{n}{\zeta \cdot \operatorname{diam}(\mathcal{M})^4 \cdot \ln n}\right)^{\frac{1}{4s+1}},$$

then the estimator $\hat{\mu}_{X_n} := P_{\Lambda_j}(\mu_{X_n})$ satisfies

$$W_2(\hat{\mu}_{X_n}, \mu) \leq C(1 + |\mu|_{\mathcal{A}^s})\left(\frac{\zeta \cdot \operatorname{diam}(\mathcal{M})^4 \cdot \ln n}{n}\right)^{\frac{s}{4s+1}}$$

with probability at least $1 - n^{-\beta}$.

In other words, given n samples from μ and a partition tree $\mathcal{T}$, if the measure μ is in the regularity class $\mathcal{A}^s$ with respect to $\mathcal{T}$, then the optimal scale j is such that the estimator $\hat{\mu}_{X_n}$ constructed by "locally projecting" the empirical measure μ_{X_n} onto $\mathcal{F}$ has, with high probability, nearly the best possible (at that scale) approximation to μ. In practice we do not have the partition tree $\mathcal{T}$; we may pick the GMRA tree constructed from a separate set of μ-samples. We also do not know s. Thus, we estimate the best level j for our estimator via cross-validation (see Section 4).

3. ALGORITHMS

Given a finite sample X_n from μ, our first step is to construct the GMRA. At any node $C_{j,k}$ of the GMRA tree, we consider the low-dimensional subspace $\operatorname{span}([\Phi_{j-1,k'}\ \Psi_{j,k}])$: the multiscale transform of a point x at this scale is encoded,

via (1), by the coefficients $\epsilon_{j,k}$ and $q_{j,k}$ defined in (2). This step is a significant reduction in the dimensionality, appropriate for the local portion of the data. We estimate the density of the coefficients $(\epsilon_{j,k}, q_{j,k})$ of the local data, using a density estimator in $\mathcal{F}$. In our examples we let $\mathcal{F}$ include mixtures of (truncated) Gaussians, and use the kernel density estimator (KDE) toolbox [5] (mainly due to its simplicity and fast implementation). Let $\hat{\mu}_{j,k}$ be the estimated density for $(\epsilon_{j,k}, q_{j,k})$ at $C_{j,k}$. Since $X_n = \cup_{k\in\Gamma_j} C_{j,k}$, we obtain a model for the underlying measure μ per scale, namely $\hat{\mu}_j := \frac{1}{n}\sum_{k\in\Gamma_j} |C_{j,k}|\,[\Phi_{j-1,k'}\,\Psi_{j,k}]\,\hat{\mu}_{j,k}$. The collection $\{\hat{\mu}_j\}_{j\geq 0}$ provide a family of density estimates for the measure μ, at multiple scales. Algorithm 1 summarizes these steps.

Algorithm 1 Multiscale-transform based Density Estimation

Input: Data set X_n

Output: Multiscale densities $\{\hat{\mu}_{j,k}\}_{j\geq 0, k\in\Gamma_j}$

1: Apply GMRA to the training data to obtain a multiscale dictionary $\{c_{j,k}, \Phi_{j,k}, w_{j,k}, \Psi_{j,k}\}_{j\geq 0, k\in\Gamma_j}$

2: For each $j > 0, k \in \Gamma_j$, apply the transform in (1) to the data in $C_{j,k}$ and obtain low dimensional coefficients

3: Apply a density estimator (e.g. KDE) to the above coefficients corresponding to each $C_{j,k}$, and obtain density estimates $\hat{\mu}_{j,k}$

Computational considerations. These algorithms are extremely fast: If d is the intrinsic dimension of data, n the sample size, and D the ambient dimension, the complexity is $O(c^d nD)$, for some small universal constant $c < 4$, for constructing a multiscale partition of the data (via cover trees [6]), plus $O(nD(\log(n) + d^2))$ for constructing a low-dimensional geometric approximation, plus the cost of computing a density estimator in d dimensions. Most of the steps in the constructions of GMRA and the measure estimator are trivially parallelizable thanks to the tree structure, and greedy updates for new incoming data are trivial as well provided the tree $\mathcal{T}$ is only grown or pruned.

4. APPLICATION, EXAMPLES AND EXTENSION

Application in Anomaly Detection. Given a sequence of data sets $\{X_n^{(t)}\}_{t\geq 0}$, each of which is a random realization of the underlying measure possibly accompanied by anomalies at unknown time and location, we want to determine when and where such anomalies occur.

With the multiscale GMRA dictionary and density estimates learned on a training data set X_n, for example $X_n = X_n^{(0)}$, this may be performed as follows. We obtain a multiscale partition of the data $X_n^{(t)}$ by assigning points to the nearest leaf node centers $c_{j,k}$ in the tree corresponding to the training data. At coarser scales, the partitions are uniquely determined by the leaf nodes. Denote this new multiscale decomposition of $X_n^{(t)}$ by $\{C_{j,k}^{(t)}\}$. We apply the transform in (1) to the data in each node $C_{j,k}^{(t)}$ to obtain joint coefficients, and then compute the likelihoods of these coefficients relative to our model $\hat{\mu}_{j,k}$ at $C_{j,k}$. When anomalies occur, they will be captured by the extreme (i.e. small) likelihoods at proper location and scale.

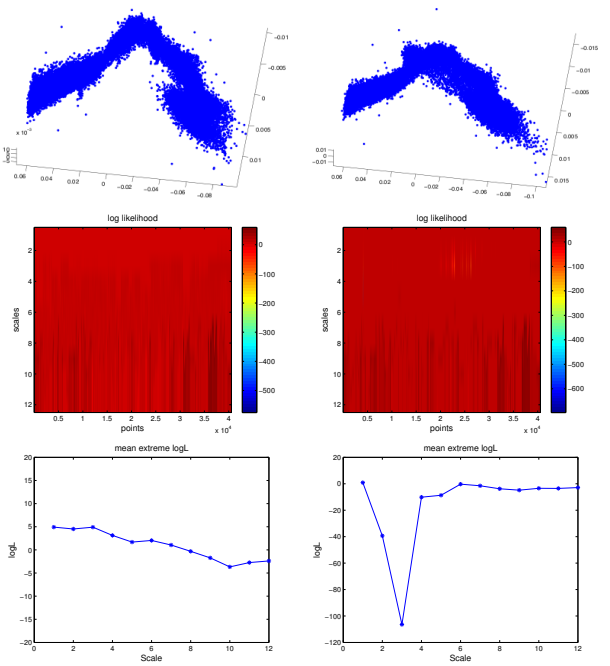

Fig. 2. Top row: two hyperspectral cubes shown in top three principal dimensions: chemical-free (left) and with chemical release (right). They are the two frames in the first column of Fig. 3. Second row: multiscale log likelihoods. Bottom row: mean values of 1% extreme log likelihoods.

We propose the following strategy for automatic selection of the scale at which anomalies can be detected. We know that anomalies tend to yield extreme likelihoods at correct scale(s), but otherwise generate comparable likelihoods with normal parts of the data. If we are given a lower bound α for the percentage of points that can be anomalies, we may compute, at every scale, a mean value of the α smallest likelihoods. The scale at which such mean values attain a local minimum is identified as the optimal scale.

Anomaly Detection on Synthetic Data. We generate a uniform probability distribution supported on a two-dimensional swissroll manifold, embedded in $\mathbb{R}^{50}$, and sample 2000 points i.i.d. from this distribution as training data. We now draw another sample of 2000 points and grow a cusp singularity at a randomly chosen location of the swissroll. We produce four data sets, the first having no anomaly and the next three having increasingly larger singularities at that fixed location. We report the results by our algorithms in Fig. 1.

Anomaly Detection in Hyperspectral Imaging. We apply our algorithms to hyperspectral imaging data for chemical detection. In a hyperspectral image (HSI), each "pixel" is a

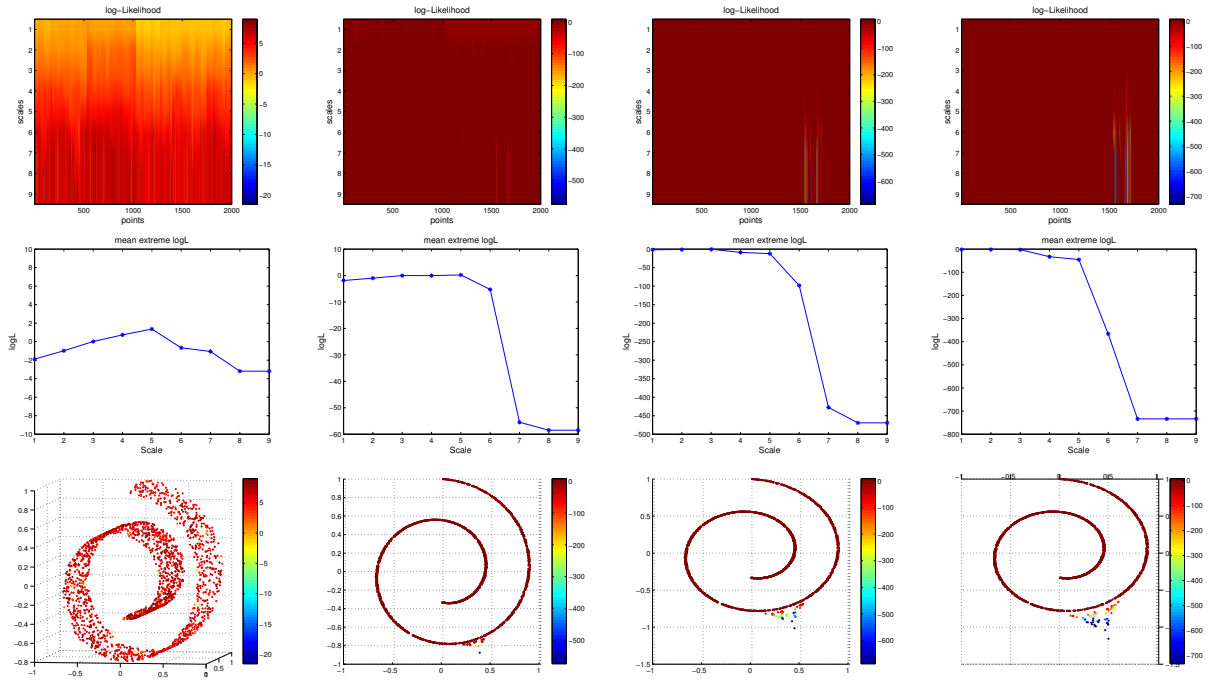

Fig. 1. Top row: multiscale likelihoods (in log scale) computed relative to the multiscale measures learned on training data. For each data set, the likelihoods are arranged into a matrix whose rows correspond to coarse (top) to fine (bottom) scales and whose columns represent the data points. Middle row: mean values of 1% extreme likelihoods at all scales. In each plot, the location of the local minimum determines an optimal scale at which the anomalies are revealed. Observe also that as the singularity grows, anomalies can be identified at coarser scales. Last row: the four data sets colored by their likelihoods at the optimal scales.

vector whose entries correspond to hundreds of narrow spectral bands, and thus the HSI is represented by a 3-D cube. The HSI data addressed in this paper contains chemical plumes that need to be detected from a desert background. Each 3-D HSI cube contains of 256×320 pixel arrays in 129 bands. The first frame in the HSI sequence is chemical-free and will be used as training data. Our goal is to identify when and where chemical plume is present in the atmosphere.

We assume that the background pixels have a manifold structure. Indeed, principal component analysis applied to these 129 dimensional vectors shows that they concentrate along a low dimensional manifold (see Fig. 2). Before the chemical is released, the hyperspectral cubes are just different observations of the background, which might vary slowly over time (e.g. due to weather changes). Anomaly will be present in the manifold when the chemical is released in the air. We display in Fig. 2 the multiscale likelihoods of two representative hyperspectral cubes, one without anomaly and the other having anomaly. Clearly, there is a block of extreme likelihoods corresponding to the anomaly frame, and the optimal scale is three. It is also interesting to note that the likelihoods of the anomalies increase rapidly after that scale, indicating considerable overfitting of the data at finer scales. In contrast, for the normal frame, the mean values are approximately constant, though slowly decreasing.

Finally, we display in Fig. 3 eight hyperspectral cubes from the dataset as two-dimensional images whose pixel intensities are the likelihoods at automatically selected scales.

Connection to Compressive Sensing. Very recently it has been demonstrated that the GMRA dictionaries can facilitate signal approximation via compressive measurements [7]. In this section we show that the HSI data can be accurately recovered using significantly fewer than 129, the ambient dimension, linear measurements.

In these experiments we constructed the GMRA using a 3-D HSI cube. We then evaluated the compressed sensing method from [7] on $10,000$ pixels independently and uniformly sampled from a second HSI cube. We obtain our compressive measurements using a different fixed random (with respect to Haar measure) orthogonal projections for each of the 12 different scale partitions of the HSI data produced by GMRA. The range of the projection, $M_{j,m}$, has dimension md_j at scale j, where m is an oversampling factor with a value from $\{1, 2, 4, 6\}$, and d_j is the maximum range dimension over all affine projectors associated with a j^{th}-scale GMRA partition. This intrinsic dimension d_j of the HSI cube is 1 for scales 1 through 7. The intrinsic dimension increases adaptively for scales 8 through 12 thereafter, as described in [3]. The actual dimension values are reported in Figure 4,

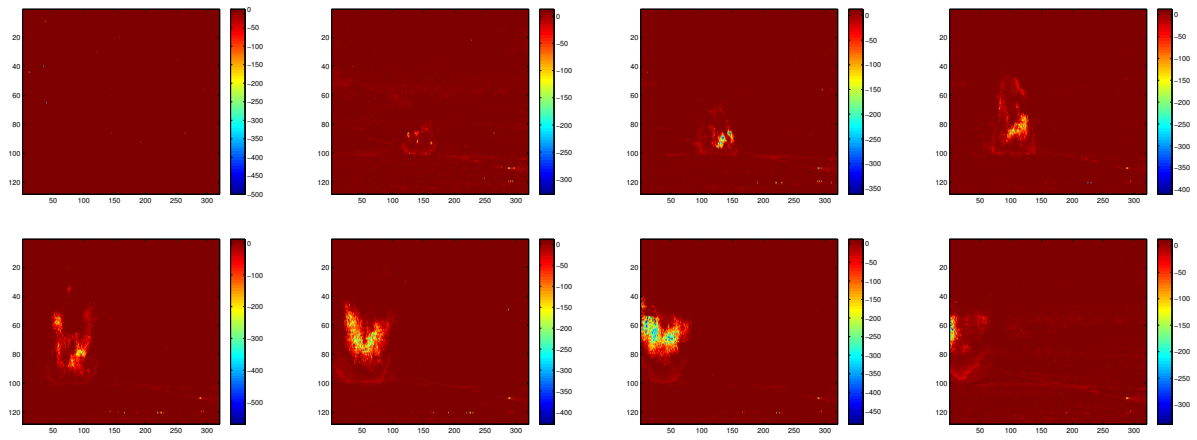

Fig. 3. Log-likelihoods of selected hyperspectral cubes in the HSI sequence relative to the training frame.

where we plot the following approximation error,

$$\mathrm{MRE}(j,m) := \max_{1\le i\le 10{,}000} \|x_i - \mathcal{A}(M_{j,m}x_i)\|_2/\|x_i\|_2,$$

in which the x_i are the data points used for evaluation, j denotes the GMRA scale, and $\mathcal{A} : \mathbb{R}^{md_j} \to \mathbb{R}^{129}$ is the compressed sensing recovery algorithm proposed in [7]. Fig. 4 also plots the maximum relative error between each point and its GMRA approximation with no compression.

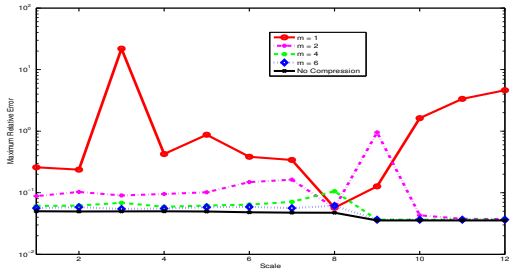

Fig. 4. Hyperspectral data reconstruction errors incurred during reconstruction with compressive measurements. The intrinsic dimension, d_j, is 1 for scales $j = 1-7$. Thereafter intrinsic dimension is: $d_8 = 8, d_9 = 32, d_{10} = 50, d_{11} = 57$, and, $d_{12} = 58$.

We can see that a relatively modest number of random linear measurements suffice to approximately recover all the tested pixels nearly as accurately as GMRA with no compression. For example, six linear measurements at scales 1 to 7 (see the curve for $m = 6$) appear to perform nearly as well as recovery in the ambient 129-dimensional space. The computational complexity of the recovery algorithm at scale j is dominated by the time required to find an (approximate) nearest neighbor for each evaluation point from the set of all scale-j GMRA centers, $c_{j,k}$. In practice, the method has been demonstrated to be several orders of magnitude faster than other sparse reconstruction algorithms (see [7] for details).

5. SUMMARY

We introduced a novel framework for estimating measures in high dimensions that are supported near intrinsically low dimensional sets. The construction of our estimators is based on a geometric multiscale decomposition of the given data and performing best local fits, while controlling the overall model complexity. We proved strong finite sample performance bounds, essentially dependent only on the intrinsic complexity of the data and not on the ambient dimension, for a large variety of models and target probability measures. The algorithms implementing this construction are fast and parallelizable, and produced accurate results when applied to synthetic and real data for anomaly detection. Finally, we showed that this framework is compatible with a generalized compressive sensing procedure.

6. REFERENCES

[1] A.V. Little, M. Maggioni, and L. Rosasco, "Multiscale geometric methods for data sets I: Multiscale covariances, noise and curvature," *submitted*, 2012.

[2] P.W. Jones, M. Maggioni, and R. Schul, "Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels," *PNAS*, 105(6):1803–1808, Feb. 2008.

[3] W.K. Allard, G. Chen, and M. Maggioni, "Multiscale geometric methods for data sets II: Geometric multiresolution analysis," *ACHA*, 32(3): 435-462, May 2012. Available online Sep. 2011.

[4] G. Chen, M.Iwen, and M.Maggioni, "Fast geometric multiscale approximation of measures in high dimensions," *in preparation*.

[5] A. Ihler and M. Mandel, "Kernel density estimation toolbox," Available at http://www.ics.uci.edu/ ihler/code/kde.html, 2003.

[6] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *Proc. ICML*, 2006.

[7] M. Iwen and M. Maggioni, "Approximation of points on low-dimensional manifolds via random linear projections," *submitted*, 2012. Preprint available on arXiv:1204.3337v1.