geometric multi-resolution analysis

Theory, Algorithms, and Applications

Nate Strawn

August 2, 2015

Georgetown University

table of contents

1. Introduction
2. Cover Trees
3. Principal Component Analysis
4. GMRA
5. Applications

introduction

introduction

Geometric Multi-Resolution Analysis [1, 4] is a method for dictionary and manifold learning that admits provable properties for a wide class of models. In order to realize this procedure as a full-fledged algorithm with an approximation theory, we need to develop an understanding of Cover Trees and Principal Component Analysis.

cover trees

metric spaces

Definition

A **metric space** is a pair $(\mathcal{X}, d)$ where $\mathcal{X}$ is a set of points and the function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (the **metric**) satisfies

1. **Positivity**: $d(x, y) \geq 0$ for all $x, y \in \mathcal{X}$ and $d(x, y) = 0$ if and only if $x = y$
2. **Symmetry**: $d(x, y) = d(x, y)$ for all $x, y \in \mathcal{X}$
3. **Triangle Inequality**: $d(x, y) \leq d(x, z) + d(y, z)$ for all $x, y, z \in \mathcal{X}$

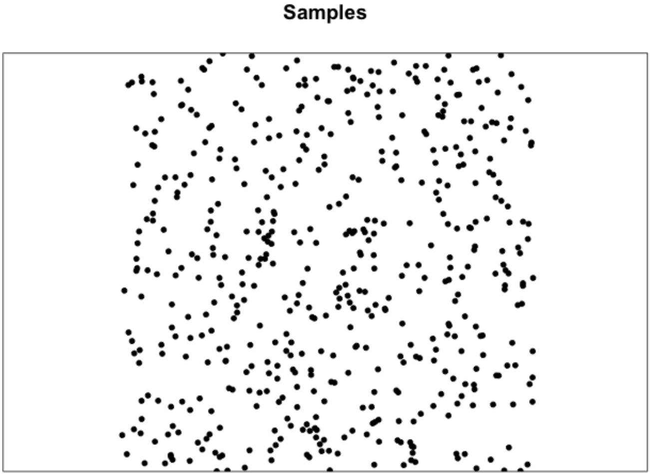
Cover Trees were introduced in [2].

Definition

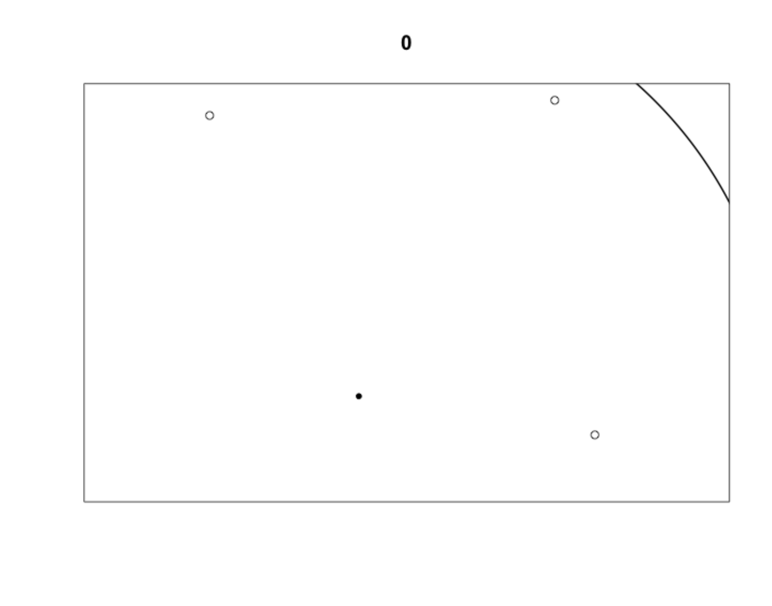
For a metric space $(\mathcal{X}, d)$, a **cover tree** over a finite set $X \subset \mathcal{X}$ is a sequence of sets ("covers") $\{C_k\}_{k \in \mathbb{Z}}$ satisfying

1. **Nesting**: $C_k \subset C_{k+1}$ and $X = \bigcup_{k \in \mathbb{Z}} C_k$
2. **Covering**: for all $x \in C_{k+1}$ there is a $y \in C_k$ such that $d(x, y) \leq 2^{-k}$
3. **Separation**: for all $x, y \in C_k$ with $x \neq y$, $d(x, y) > 2^k$

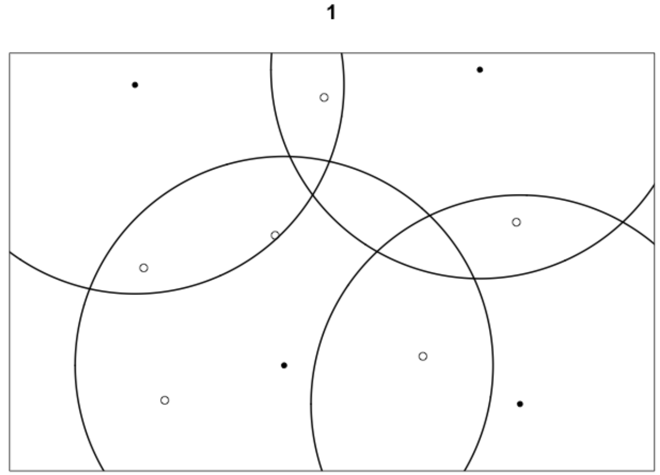
example



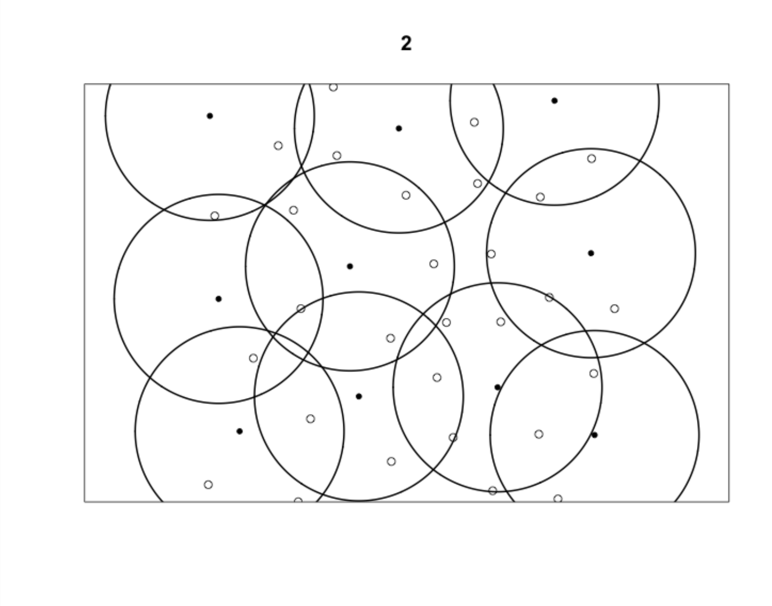
example



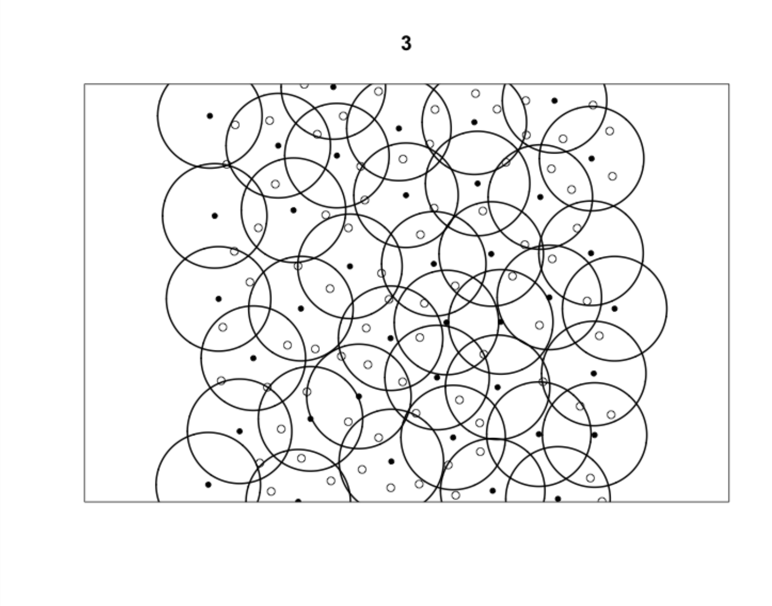
example



example



example



insertion

$\mathrm{Insert}(\mathcal{C}, p, Q, k)$ where $\mathcal{C} = \{C_k\}_{k \in \mathbb{Z}}$ is a cover tree in the metric space $(\mathcal{X}, d)$, $p \in \mathcal{C}$ is the point to be inserted, $Q \subset C_k$ is the current cover, and k is the insertion level

1. $Q' \leftarrow \{q \in C_{k+1} : d(q, Q) \leq 2^{-k}\}$
2. IF $d(p, Q') > 2^{-k}$ return FAIL
3. ELSE
   - 3.1 $Q'' \leftarrow \{q \in Q' : d(p, q) \leq 2^{-k}\}$
   - 3.2 IF $\mathrm{Insert}(\mathcal{C}, p, Q'', k+1) = \mathrm{FAIL}$ AND $d(p, Q) \leq 2^{-k}$
     - $C_j \leftarrow C_j \cup \{p\}$ for all $j > k$
     - return SUCCESS
   - 3.3 ELSE return FAIL

insertion proof

For a cover tree $\mathcal{C}$ on $X \subset \mathcal{X}$, let K be the largest number such that $C_{-\infty} = C_K$, suppose $p \notin X$, and let κ denote the largest number such that $d(p, C_{-\infty}) \leq 2^{-\kappa}$. Then, we initialize insertion at scale $\min(\kappa, K)$.

insertion proof

Since $X \subset \mathcal{X}$ is finite and $p \notin X$, we have that $d(p, X) > 0$, and the recursion is destined to return FAIL at some finite index. This triggers the Step 3.2 to test that the input cover Q satisfies $d(p, Q) \leq 2^{-k}$. Since we know that $d(p, C_{\min(\kappa, K)}) \leq 2^{-\min(\kappa, K)}$, we conclude there is some index $k \leq \min(\kappa, K)$ where Step 3.2 succeeds.

insertion proof

For this K, we know the following:

1. p is added to all cover sets C_j for $j > k$
2. $d(p, Q) \leq 2^{-k}$ for the local variable $Q \subset C_K$
3. $\text{Insert}(\mathcal{C}, p, Q'', k+1)$ returned FAIL, and hence $d(p, Q'') > 2^{-k-1}$

Nesting follows from this first observation, and covering follows from the second.

insertion proof

We now demonstrate that $d(p, C_j) > 2^{-j}$ holds for all $j > k$, and hence the separation condition is satisfied when p is added to each C_j with $j > K$. Let $a \in C_j$. If $q \in Q$, then the third observation above implies that $d(p, q) > 2^{-k-1} \geq 2^{-j}$. If $q \notin Q$, then there is a scale i and an ancestor of q, $r \in C_{i+1}$, such that r was eliminated in Step 3.1 of $\mathrm{Insert}(\mathcal{C}, p, \widetilde{Q}, i)$. Thus, $d(p, r) > 2^{-i}$ and the covering property imply

$$d(p, q) \geq d(p, r) - d(q, r) > 2^{-i} - \sum_{t=i+1}^{j-1} 2^{-t} = 2^{-i} - 2^{-i} + 2^{-j} = 2^{-j}.$$

From these two cases, we conclude that the separation condition holds.

nearest neighbor

$\mathrm{NN}(p, \mathcal{C})$ where $\mathcal{C} = \{C_k\}_{k \in \mathbb{Z}}$ is a cover tree on the finite set X in the metric space $(\mathcal{X}, d)$ and $p \in \mathcal{X}$

1. LET K be the largest value such that $C_{-\infty} = C_K$
2. LET L be the smallest value such that $C_L = X$
3. IF $K < \infty$, FOR $k = K$ to $L - 1$
    3.1 Set $Q \leftarrow \{q \in C_{k+1} : d(q, Q_k) \leq 2^{-k}\}$
    3.2 Form the cover set $Q_{k+1} = \{q \in Q : d(p, q) \leq d(p, Q) + 2^{-k}\}$

    and return $\arg\min_{q \in Q_L} d(p, q)$
4. ELSE return the single member of X

nearest neighbor proof

Let $q^* \in X$ denote the nearest neighbor to p in X. For any $q \in C_{k+1}$, the distance between q and any descendant q' is bounded by

$$d(q, q') \leq \sum_{j=k+1}^{\infty} 2^{-j} = 2^{-k}.$$

Consequently, Step 2.2 can never remove an ancestor of the nearest neighbor because

$$d(p, q) \leq d(p, q^*) + d(q, q^*) \leq d(p, Q) + 2^{-k}$$

for all ancestors $q \in Q$ of q^*. Thus, Q_L must contain q^*. $\square$

expansion constant

Given a finite subset $X \subset \mathcal{X}$, define

$$B_X(p, r) = \{q \in X : d(p, q) \leq r\}$$

The **expansion constant** of X is the smallest value $c \geq 2$ such that

$$|B_X(p, 2r)| \leq c|B_X(p, r)|$$

for all $p \in \mathcal{X}$ and $r > 0$, where $|A|$ is the number of elements in a finite set A.

expansion constant

Example: The expansion constant of the integer lattice in $\mathbb{R}^D$ is 2^D with the metric

$$\|x\|_\infty = \max_{i \in [D]} |x_i|.$$

algorithmic properties

- Space Complexity: $O(n)$
- Time Complexity of Insertion: $O(c^6 n \log n)$
- Time Complexity of Construction: $O(c^6 n \log n)$
- Time Complexity of Nearest Neighbor: $O(c^{12} \log n)$

partitions

For a cover tree $\mathcal{C}$ and any scale k, the cover set $C_k = \{x_j\}_{j=1}^J$ induces a **partition** $\{Q_j\}_{j=1}^J$ of $\mathcal{X}$ ($Q_i \cap Q_j = \emptyset$ if $i \neq j$ and $\bigcup_{j=1}^J Q_j = \mathcal{X}$):

$$Q_j = \{x \in \mathcal{X} : j = \min\{i : d(x, x_i) = d(x, C_k)\}\}.$$

These Q_j are called Voronoi regions.

principal component analysis

projections

An **orthogonal projection** from $\mathbb{R}^D$ to $\mathbb{R}^d$ is a linear map P such that the adjoint P^* satisfies $P \circ P^* = Id_{\mathbb{R}^d}$. When viewed as a matrix,

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_d \end{pmatrix}$$

where the p_i are orthonormal row vectors. If $D = d$, then P is an **orthogonal matrix**.

singular value decomposition

For a d by N matrix X with rank r, a **singular value decomposition** of X has the form

$$X = U\Sigma V^*$$

where U^T is an r by d orthogonal projection, V^T is a r by N orthogonal projection, and Σ is a diagonal matrix with non-negative, non-increasing entries along the diagonal. The nonzero entries of X are the **singular values** of X

principal components

The rank d **matrix of Principal Components** of a D by N dataset matrix X is the matrix P consisting of the first d columns of U where $X = U\Sigma V^T$. The d by N matrix P^TX is the projection onto these components.

affine approximations

Let μ be the mean vector of the columns of the D by N matrix X, and let $\widetilde{X}$ denote the matrix obtained by subtracting μ from each column of $\widetilde{X}$. Let P be the rank d matrix of Principal Components of $\widetilde{X}$. Then the rank d **affine projection** of X is the map

$$\mathcal{P}(x) = PP^T(x - \mu) + \mu.$$

choosing the rank

How do you choose the rank for the Principal Component Projection?

- Fix a level $\alpha > 0$ and let d be the first d such that

$$\frac{\sum_{i=1}^{d} \sigma_i}{\sum_{i=1}^{D} \sigma_i} \geq 1 - \alpha$$

- Find d which maximizes the sphericity:

$$\frac{\left(\sum_{i=d+1}^{D} \sigma_i\right)^2}{\sum_{i=d+1}^{D} \sigma_i^2}$$

- In high dimensions, consider the Sparse PCA procedure of Vu, Cho, Rohe, and Jing Lei

gmra

gmra

Given a dataset $X \subset \mathbb{R}^D$, a sequence of multiscale partitions $\mathcal{Q}_k(X) = \{Q_j^{(k)}\}_{j=1}^{m_k}$, and affine approximations $\mathcal{P}_{j,k}$ to $X \cap Q_j^{(k)}$, the **GMRA** projections are the sequence of functions

$$\mathcal{P}_k(x) = \sum_{j=1}^{m_k} \mathbf{1}_{Q_j^{(k)}}(x)\mathcal{P}_{j,k}(x).$$

good partitions

For "good" partitions of $\mathbb{R}^D$ which are compatible with an underlying probability distribution Π on $\mathbb{R}^D$, we can prove good approximations properties for the GMRA construction.

Concentration

$$\Pi(Q_k) \geq c_1 2^{-jd}$$

Centrality

$$\|X - \mu_k\| \leq c_2 2^{-j}, \; \Pi_k \text{ a.s.}$$

good partitions

Complexity

$$\lambda_d^{(k)} \geq c_3 \frac{-2j}{d}$$

$$\sum_{l>d} \lambda_l^{(k)} \leq c_4(\sigma^2 + 2^{-4j}) \leq \frac{1}{2}\lambda_d^{(k)}$$

mean square error

Theorem (Maggioni, Minsker, Strawn, 2014)

Suppose the partition $\{Q_k\}_{k=1}^K$ *is "good" for* Π *at a scale* j*, which is above the "noise level"* σ*. If the i.i.d. samples* $X_1, \ldots, X_N$ *are drawn according to* Π *and* $\widehat{\mathcal{P}}$ *is the associated empirical GMRA, then with high probability we have*

$$\mathbb{E}\|X - \widehat{\mathcal{P}}(X)\|^2 \leq C_1 \left(\sigma^2 + 2^{-4j}\right) + C_2 2^{-2j} \frac{\log(d)}{N2^{-jd}}.$$

mse proof: bias variance decomposition

$$\mathbb{E}\|X - \widehat{\mathcal{P}}(X)\|^2 \leq 2\mathbb{E}\|X - \mathcal{P}(X)\|^2 + 2\mathbb{E}\|\mathcal{P}(X) - \widehat{\mathcal{P}}(X)\|^2$$

mse proof

See [5].

Theorem (Minsker, 2013)

Let $Z_1, \dots, Z_N \in \mathbb{R}^{D \times D}$ be an independent sequence of symmetric random matrices such that $\mathbb{E} Z_i = 0$ and $\|Z_i\| \le U$ almost surely for all $1 \le i \le N$. If

$$\sigma^2 = \left\| \sum_{i=1}^{N} \mathbb{E} Z_i^2 \right\| \text{ and } \rho = 4 \frac{\mathrm{trace}\left(\sum_{i=1}^{N} \mathbb{E} Z_i^2\right)}{\sigma^2},$$

then for any $t \ge 1$ we have

$$\left\| \sum_{i=1}^{N} Z_i \right\| \le 2 \max\left(\sigma \sqrt{t + \log \rho}, U(t + \log \rho)\right)$$

with probability exceeding $1 - e^{-t}$.

mse proof

- Bound $\|\mu_k - \widehat{\mu}_k\|$ and $\|\Sigma_k - \widehat{\Sigma}_k\|$ with high probability using this last theorem.

mse proof

- Bound $\|\mu_k - \widehat{\mu}_k\|$ and $\|\Sigma_k - \widehat{\Sigma}_k\|$ with high probability using this last theorem.
- Infer a bound on $\|P_k - \widehat{P}_k\|$ using Davis-Kahan:

mse proof

- Bound $\|\mu_k - \widehat{\mu}_k\|$ and $\|\Sigma_k - \widehat{\Sigma}_k\|$ with high probability using this last theorem.

- Infer a bound on $\|P_k - \widehat{P}_k\|$ using Davis-Kahan:

Theorem (Davis and Kahan, 1970)
Let $\delta = \frac{1}{2}(\lambda_d^{(k)} - \lambda_{d+1}^{(k)})$. *If* $\|\Sigma_k - \widehat{\Sigma}_k\| \leq \delta/2$, *then* $\|P_k - \widehat{P}_k\| \leq \frac{1}{\delta}\|\Sigma_k - \widehat{\Sigma}_k\|$.

mse proof

- Bound $\|\mu_k - \widehat{\mu}_k\|$ and $\|\Sigma_k - \widehat{\Sigma}_k\|$ with high probability using this last theorem.

- Infer a bound on $\|P_k - \widehat{P}_k\|$ using Davis-Kahan:

Theorem (Davis and Kahan, 1970)
Let $\delta = \frac{1}{2}(\lambda_d^{(k)} - \lambda_{d+1}^{(k)})$. *If* $\|\Sigma_k - \widehat{\Sigma}_k\| \leq \delta/2$, *then* $\|P_k - \widehat{P}_k\| \leq \frac{1}{\delta}\|\Sigma_k - \widehat{\Sigma}_k\|$.

- Infer individual bounds on $\|\mathcal{P}_k(x) - \widehat{\mathcal{P}}_k(x)\|^2$, and accumulate. $\square$

good partitions for good models

It remains to demonstrate that there is a partition strategy for some class of models such that good partitions are achievable. Our partition strategy will be from the Cover Tree algorithm. We'll need a few definitions before we can define a class of models that yield "good" partitions.

manifolds

Definition

A C^k **manifold** of dimension d is a topological space $\mathcal{M}$ along with an **atlas** of **coordinate charts** of the form $u : U \to \mathbb{R}^d$ where

1. Each domain U is an open subset of $\mathcal{M}$
2. Each coordinate map u is a homeomorphism
3. For any two coordinate maps $u : U \to \mathcal{M}$ and $v : V \to \mathcal{M}$, the coordinate transition map $u \circ v^{-1} : v(U \cap V) \to u(U \cap V)$ is k-times differentiable
4. The union over the domains of the coordinate charts is all of $\mathcal{M}$

Definition

A Euclidean **embedding** of a C^k manifold $\mathcal{M}$ is a map $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^D$ such that $\iota \circ u^{-1}$ is a C^k map for any chart u on $\mathcal{M}$, and the Jacobians $D_{u(x)}(\iota \circ u^{-1})$ have full rank for each chart u and each point x in the domain of u. If a subset $\mathcal{X} \subset \mathbb{R}^D$ is the image of such an ι, we say that $\mathcal{X}$ is an **embedded manifold** and we write $\mathcal{X} \hookrightarrow \mathbb{R}^D$.

normal bundles

If $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^D$ is an embedding and u and v are charts around $x \in \mathcal{M}$ then it can be shown that the Jacobians

$$D_u(x)(\iota \circ u^{-1}) \text{ and } D_v(x)(\iota \circ v^{-1})$$

share the same column space, which we call the **tangent space** of $\mathcal{M}$ at x (denoted $T_x\mathcal{M}$). The orthogonal complement of $T_x\mathcal{M}$ is call the normal space at x, and is denoted $N_x\mathcal{M}$.

normal bundles

If $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^D$ is an embedding and u and v are charts around $x \in \mathcal{M}$ then it can be shown that the Jacobians

$$D_u(x)(\iota \circ u^{-1}) \text{ and } D_v(x)(\iota \circ v^{-1})$$

share the same column space, which we call the **tangent space** of $\mathcal{M}$ at x (denoted $T_x\mathcal{M}$). The orthogonal complement of $T_x\mathcal{M}$ is call the normal space at x, and is denoted $N_x\mathcal{M}$.

The set of points $(x, v) \in \iota(\mathcal{M}) \times \mathbb{R}^D$ such that $v \in N_x\mathcal{M}$ and $\|v\| < r$ for $r > 0$ constitutes the radius-r **normal bundle** of $\iota(\mathcal{M})$, which we denote $N(\iota(\mathcal{M}), r)$. It can be shown that $N(\iota(\mathcal{M}), r)$ is a C^{k-1} manifold of dimension D embedded in $\mathbb{R}^D \times \mathbb{R}^D$.

tubular neighborhoods

Definition

A **tubular neighborhood** of a manifold $\mathcal{M} \hookrightarrow \mathbb{R}^D$ is the set

$$\mathcal{T}_r(\mathcal{M}) = \left\{ x \in \mathbb{R}^D : \min_{y \in \mathcal{M}} \|x - y\| < r \right\}.$$

It can be shown that $\mathcal{T}_r(\mathcal{M})$ is the image of $N(\mathcal{M}, r)$ under the map $(x, v) \mapsto x + v$.

reach of a manifold

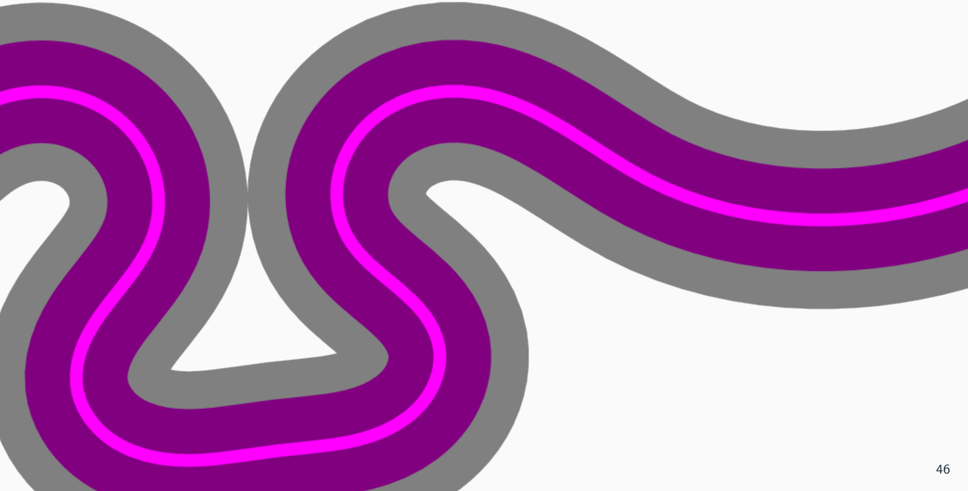
Definition

The **reach** of a manifold $\mathcal{M} \hookrightarrow \mathbb{R}$ is the supremum over the values τ such that $N(\mathcal{M}, r) \hookrightarrow \mathbb{R}^D$.

noisy manifold

Definition

A **noisy manifold** is a probability distribution Π is supported on a radius σ tubular neighborhood of a closed C^2 manifold $\mathcal{M}$ with reach $\tau > 0$.



good partition theorem

Theorem (Maggioni, Minsker, Strawn, 2014)

Suppose

- $\mathcal{M} \hookrightarrow \mathbb{R}^D$ *is a closed* C^2, d*-dimensional manifold with reach* $\tau > 0$
- Π *is mutually absolutely continuous with respect to the uniform distribution on the radius* $\sigma < \tau$ *tubular neighborhood of* $\mathcal{M}$
- Y_1, Y_2,..., Y_N *is an i.i.d. sample from* Π *with* $N \geq C\sigma^{-d}(t - \log \sigma)$

Then for appropriate scales j, *we have that the cover tree partition at scale* j, $\{Q_k\}_{k=1}^K$ *is "good" with probability exceeding* $1 - e^{-t}$.

prerequisite theory

Apart from Centrality condition, all of our conditions are integral conditions. Thus, we need to be able to estimate probabilities on tubular neighborhoods, which ultimately means that we need to estimate volumes on the tubular neighborhoods.

prerequisite theory

Apart from Centrality condition, all of our conditions are integral conditions. Thus, we need to be able to estimate probabilities on tubular neighborhoods, which ultimately means that we need to estimate volumes on the tubular neighborhoods.

First, it is important to note that all closed manifolds embedded in $\mathbb{R}^D$ inherit a volume or Hausdorff measure: Let $\mathrm{Proj}_{\mathcal{M}}$ denote projection onto the manifold, then

$$\mathrm{Vol}_{\mathcal{M}}(U) = \lim_{r \to 0} \frac{\mathrm{Vol}(\mathrm{Proj}_{\mathcal{M}}^{-1}(U) \cap \mathcal{M}_r)}{r^{D-d}\mathrm{Vol}(B_{D-d}(0,1))}.$$

prerequisite theory

Apart from Centrality condition, all of our conditions are integral conditions. Thus, we need to be able to estimate probabilities on tubular neighborhoods, which ultimately means that we need to estimate volumes on the tubular neighborhoods.

First, it is important to note that all closed manifolds embedded in $\mathbb{R}^D$ inherit a volume or Hausdorff measure: Let $\mathrm{Proj}_{\mathcal{M}}$ denote projection onto the manifold, then

$$\mathrm{Vol}_{\mathcal{M}}(U) = \lim_{r \to 0} \frac{\mathrm{Vol}(\mathrm{Proj}_{\mathcal{M}}^{-1}(U) \cap \mathcal{M}_r)}{r^{D-d}\mathrm{Vol}(B_{D-d}(0,1))}.$$

For this closed manifold $\mathcal{M}$, we define the **uniform distribution** on $\mathcal{M}$ by

$$\mathcal{U}_{\mathcal{M}}(U) = \frac{\mathrm{Vol}_{\mathcal{M}}(U)}{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})}.$$

volumes of tubular neighborhoods

For 1-dimensional manifolds, volumes of tubular neighborhoods admit a simple formula. If $U \subset \mathcal{M} \hookrightarrow \mathbb{R}^D$ is open and $\widetilde{U} = \mathrm{Proj}_{\mathcal{M}}^{-1}(U)$, then

$$\mathrm{Vol}(\widetilde{U}) = \sigma^{D-1}\mathrm{Vol}(B_{D-1}(0,1)) \cdot \mathrm{Vol}_{\mathcal{M}}(U).$$

volumes of tubular neighborhoods

For 1-dimensional manifolds, volumes of tubular neighborhoods admit a simple formula. If $U \subset \mathcal{M} \hookrightarrow \mathbb{R}^D$ is open and $\widetilde{U} = \mathrm{Proj}_{\mathcal{M}}^{-1}(U)$, then

$$\mathrm{Vol}(\widetilde{U}) = \sigma^{D-1}\mathrm{Vol}(B_{D-1}(0,1)) \cdot \mathrm{Vol}_{\mathcal{M}}(U).$$

In higher dimensions, there is a Weyl tube formula which expresses volumes in terms of Lipschitz-Killing curvatures:

$$\mathrm{Vol}(\mathcal{M}_\sigma) = \sum_{n=0}^{d} \mu_{d-n}(\mathcal{M})\mathrm{Vol}(B_n(0,1))\sigma^n$$

volumes of tubular neighborhoods

For 1-dimensional manifolds, volumes of tubular neighborhoods admit a simple formula. If $U \subset \mathcal{M} \hookrightarrow \mathbb{R}^D$ is open and $\widetilde{U} = \text{Proj}_{\mathcal{M}}^{-1}(U)$, then

$$\text{Vol}(\widetilde{U}) = \sigma^{D-1}\text{Vol}(B_{D-1}(0,1)) \cdot \text{Vol}_{\mathcal{M}}(U).$$

In higher dimensions, there is a Weyl tube formula which expresses volumes in terms of Lipschitz-Killing curvatures:

$$\text{Vol}(\mathcal{M}_\sigma) = \sum_{n=0}^{d} \mu_{d-n}(\mathcal{M})\text{Vol}(B_n(0,1))\sigma^n$$

Now, this only holds for the full manifold and it involves some unwieldy computation. Instead, we want estimates saying that the 1-dimensional formula approximately holds.

volume bounds

Theorem (Maggioni, Minsker, Strawn, 2014)

Let $\mathcal{M} \hookrightarrow \mathbb{R}^D$ be a d-dimensional submanifold with reach $\tau > 0$, suppose $\sigma < \tau$ and that the measurable subset $U \subset \mathcal{M}$ satisfies $Vol_{\mathcal{M}}(U) > 0$ for the volume measure on $\mathcal{M}$, and finally let $Proj_{\mathcal{M}} : \mathcal{M}_\sigma \to \mathcal{M}$ denote the map that assigns a point in the tubular neighborhood of $\mathcal{M}$ to the closest point on $\mathcal{M}$. Then

$$\left(1 - \frac{\sigma}{\tau}\right)^d \leq \frac{Vol_{\mathbb{R}^D}(Proj_{\mathcal{M}}^{-1}(U))}{Vol_{\mathcal{M}}(U) Vol_{\mathbb{R}^{D-d}}(B_{D-d}(0,\sigma))} \leq \left(1 + \frac{\sigma}{\tau}\right)^d$$

where $Vol_{\mathbb{R}^{D-d}}(B_{D-d}(0,\sigma))$ is the volume of the ball of radius σ in $\mathbb{R}^{D-d}$.

Corollary

Let $\mathcal{U}_{\mathcal{M}}$, $\mathcal{U}_{\mathcal{M}_\sigma}$, and $\widetilde{\mathcal{U}}_{\mathcal{M}_\sigma}$ denote the uniform measure on $\mathcal{M}$, the uniform measure on $\mathcal{M}_\sigma$, and the push forward of $\mathcal{U}_{\mathcal{M}_\sigma}$ under $\mathrm{Proj}_{\mathcal{M}}$, respectively. Then the Radon-Nikodym derivative satisfies

$$\left(\frac{\tau - \sigma}{\tau + \sigma}\right)^d \leq \frac{d\,\widetilde{\mathcal{U}}_{\mathcal{M}_\sigma}}{d\,\mathcal{U}_{\mathcal{M}}} \leq \left(\frac{\tau + \sigma}{\tau - \sigma}\right)^d .$$

idea of the proof

- Locally, we may approximate the manifold by a multivariable function of the tangent space, $v \mapsto (v, f(v))$. A local embedding of the normal bundle is then given by

$$\begin{pmatrix} v \\ \beta \end{pmatrix} \longmapsto \begin{pmatrix} v \\ f(v) \end{pmatrix} + \begin{pmatrix} Df(v)^* \\ -I_{(D-d)\times(D-d)} \end{pmatrix} \beta$$

idea of the proof

- Locally, we may approximate the manifold by a multivariable function of the tangent space, $v \mapsto (v, f(v))$. A local embedding of the normal bundle is then given by

$$\begin{pmatrix} v \\ \beta \end{pmatrix} \longmapsto \begin{pmatrix} v \\ f(v) \end{pmatrix} + \begin{pmatrix} Df(v)^* \\ -I_{(D-d)\times(D-d)} \end{pmatrix} \beta$$

- The Jacobian of this map is

$$\begin{pmatrix} I_{d\times d} + \sum_{i=d+1}^{D} \beta_i D^2 f_i(v) & Df(v)^* \\ Df(v)^* & -I_{(D-d)\times(D-d)} \end{pmatrix}$$

idea of the proof

- The necessary invertibility of this Jacobian can be used to derive the conditions (for $\varepsilon < \tau/8$)

$$\sup_{v \in B_d(0,\varepsilon)} \|Df(v)\| \leq \frac{2\varepsilon}{\tau - 2\varepsilon}$$

and

$$\sup_{v \in B_d(0,\varepsilon)} \sup_{u \in \mathcal{S}^{D-d-1}} \left\| \sum_{i=d+1}^{D} u_i D^2 f_i(v) \right\| \leq \frac{\tau^2}{(\tau - 2\varepsilon)^3}.$$

idea of the proof

- The necessary invertibility of this Jacobian can be used to derive the conditions (for $\varepsilon < \tau/8$)

$$\sup_{v \in B_d(0,\varepsilon)} \|Df(v)\| \leq \frac{2\varepsilon}{\tau - 2\varepsilon}$$

and

$$\sup_{v \in B_d(0,\varepsilon)} \sup_{u \in \mathcal{S}^{D-d-1}} \left\| \sum_{i=d+1}^{D} u_i D^2 f_i(v) \right\| \leq \frac{\tau^2}{(\tau - 2\varepsilon)^3}.$$

- These bounds imply upper and lower comparison bounds for the determinant of the Jacobian of the embedding, from which we gather the volume bounds. □

volume lemma

Lemma

Suppose $Y = [y_1|\cdots|y_d]$ *is symmetric* d *by* d *matrix such that* $\|Y\| \leq q < 1$*. Then*

$$\begin{aligned} Vol\begin{pmatrix} I+Y \\ X \end{pmatrix} &\leq (1+q)^d \, Vol\begin{pmatrix} I \\ X \end{pmatrix} \\ Vol\begin{pmatrix} I+Y & X^T \\ X & -I \end{pmatrix} &\geq (1-q)^d \, Vol\begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}. \end{aligned}$$

volume lemma proof

For the first inequality, let

$$A = \begin{pmatrix} I \\ X \end{pmatrix} \text{ and } B = \begin{pmatrix} Y \\ 0 \end{pmatrix},$$

and for every $T \subset [d]$, we let V_T denote the volume of $\{a_i\}_{i \in T^c} \cup \{b_i\}_{i \in T}$, where a_i and b_i denote the ith columns of A and B respectively.

volume lemma proof

By submultilinearity of the volume we have

$$\mathrm{Vol}(A+B) \le \sum_{T \in 2^{[d]}} V_T,$$

where $2^{[d]} = \{S : S \subset \{1, \ldots, d\}\}$. We now show that $V_T \le q^{|T|}\mathrm{Vol}(A)$ for every $T \in 2^{[d]}$. The bound $\|Y\| \le q$ implies $\|y_i\| \le q$ for all $i = 1, \ldots, d$, and so the fact that the volume is a submultiplicative function implies that

$$V_T \le q^{|T|}\mathrm{Vol}(A_{T^c}).$$

volume lemma proof

On the other hand, letting $a_1^\perp$ be the orthogonal projection of a_1 onto $\text{span}^\perp\{a_i\}_{i=2}^d$, we note that $\|a_1^\perp\| \geq 1$, and thus

$$\text{Vol}(A_{\{1\}^c}) \leq \|a_1^\perp\|\text{Vol}(A_{\{1\}^c}) = \text{Vol}(A).$$

By induction and invariance of the volume under permutations, we see that $\text{Vol}(A_{T^c}) \leq \text{Vol}(A)$ for all $T \in 2^{[d]}$. Thus,

$$\text{Vol}(A+B) \leq \sum_{T\in 2^{[d]}} q^{|T|}\text{Vol}(A) = (1+q)^d\text{Vol}(A).$$

volume lemma proof

For the second inequality, since Y is symmetric, we can represent it as $Y = F - G$ where F and G are symmetric positive semidefinite, $FG = GF = 0$, and $\|F\|, \|G\| \leq \|Y\|$. Indeed, if $Y = Q\Lambda Q^T$ is the eigenvalue decomposition of Y with $\Lambda = \mathrm{diag}(\lambda)$, set $\lambda_+ := (\max(0, \lambda_1), \ldots, \max(0, \lambda_d))^T$, $\lambda_- := \lambda_+ - \lambda$, and define $F := Q\mathrm{diag}(\lambda_+)Q^T$, $G = Q\mathrm{diag}(\lambda_-)Q^T$.

volume lemma proof

Recall the *matrix determinant lemma*: let $T \in \mathbb{R}^{k \times k}$ be invertible, and let $U, V \in \mathbb{R}^{k \times l}$. Then

$$\text{Vol}(T + UV^T) = \text{Vol}(I + V^T T^{-1} U)\text{Vol}(T).$$

Applying it in our case with $U = \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix}$, $V = \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}$, and $T = \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}$, we have that

$$\text{Vol}\begin{pmatrix} I + Y & X^T \\ X & -I \end{pmatrix} = \text{Vol}\left(I + \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^T \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix} \right) \text{Vol}\begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}.$$

volume lemma proof

By orthogonality of the columns in

$$\begin{pmatrix} I \\ X \end{pmatrix}$$

with the columns in

$$\begin{pmatrix} X^T \\ -I \end{pmatrix},$$

we have that

$$\left\| \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \right\| \geq \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|,$$

and hence

$$\left\| \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^T \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix} \right\| \leq \sqrt{q} \cdot 1 \cdot \sqrt{q} = q.$$

volume lemma proof

Therefore, we conclude that

$$\mathrm{Vol}\left(I + \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^{T} \begin{pmatrix} I & X^{T} \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix}\right) \geq (1-q)^{d},$$

and combining this with the expression from the matrix determinant lemma completes the proof. □

covariance bounds

Theorem (Maggioni, Minsker, Strawn, 2014)

Suppose Π *is a distribution supported on* $\mathcal{M}_\sigma$*, and let* $r < \tau/2$*. Further assume that* Z *is the random variable drawn from* Π *conditioned on the event* $Z \in Q$ *where* $\mathcal{M}_\sigma \cap Q \subset B(y, r)$ *for some* $y \in \mathcal{M}$*. If* Σ *is the covariance matrix of* Z*, then*

$$\sum_{i=d+1}^{D} \lambda_i(\Sigma) \leq 2\sigma^2 + \frac{8r^4}{\tau^2},$$

where $\lambda_i(\Sigma)$ *are the eigenvalues of* Σ *arranged in the decreasing order.*

covariance bounds

Theorem (Maggioni, Minsker, Strawn, 2014)

Suppose that $Q \subseteq \mathbb{R}^D$ *is such that*

$$B(y, r_1) \subset Q \text{ and } \mathcal{M}_\sigma \cap Q \subset B(y, r_2)$$

for some $y \in \mathcal{M}$ *and* $\sigma < r_1 < r_2 < \tau/8 - \sigma$. *Let* Z *be drawn from* $U_{\mathcal{M}_\sigma}$ *conditioned on the event* $Z \in Q$, *and suppose* Σ *is the covariance matrix of* Z. *Then*

$$\lambda_d(\Sigma) \geq \frac{1}{4\left(1 + \frac{\sigma}{\tau}\right)^d} \left(\frac{r_1 - \sigma}{r_2 + \sigma}\right)^d \left(\frac{1 - \left(\frac{r_1 - \sigma}{2\tau}\right)^2}{1 + \left(\frac{2(r_2+\sigma)}{\tau - 2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1 - \sigma)^2}{d}.$$

sampling nets

This is a slight modification from the corresponding result in [6]

Theorem (Niyogi, Smale, Weinberger, 2008)

Suppose $0 < \varepsilon < \frac{\tau}{2}$, *and also that* n *and* t *satisfy*

$$n \geq \varepsilon^{-d} \frac{1}{\phi_1} \left(\frac{\tau + \sigma}{\tau - \sigma} \right)^d \beta_1 \left(\log(\varepsilon^{-d} \beta_2) + t \right),$$

where $\beta_1 = \frac{Vol_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_1) Vol(B_d(0,1/4))}$, $\beta_2 = \frac{Vol_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_2) Vol(B_d(0,1/8))}$, $\delta_1 = \sin^{-1}(\varepsilon/8\tau)$, *and* $\delta_2 = \sin^{-1}(\varepsilon/16\tau)$. *Let* $\mathcal{E}_{\varepsilon/2,n}$ *be the event that*

$$\mathcal{Y} = \{Y_j = Proj_{\mathcal{M}}(X_j)\}_{j=1}^n$$

is $\varepsilon/2$*-dense in* $\mathcal{M}$ *(that is,* $\mathcal{M} \subset \bigcup_{i=1}^n B(Y_i, \varepsilon/2)$*). Then,* $\Pi^n(\mathcal{E}_{\varepsilon,n}) \geq 1 - e^{-t}$, *where* Π^n *is the n-fold product measure of* Π *with* $\phi_1 \leq \frac{d\Pi}{d\mathcal{U}_{\mathcal{M}_\sigma}}$.

conclusion

- Building a cover tree on this net and invoking the separation property implies that each Voronoi region contains a small ball, and our probability bounds then supply the Concentration property

conclusion

- Building a cover tree on this net and invoking the separation property implies that each Voronoi region contains a small ball, and our probability bounds then supply the Concentration property
- The Voronoi regions from the cover trees are also bounded by balls, which implies the Centrality property

conclusion

- Building a cover tree on this net and invoking the separation property implies that each Voronoi region contains a small ball, and our probability bounds then supply the Concentration property
- The Voronoi regions from the cover trees are also bounded by balls, which implies the Centrality property
- The above eigenvalue bounds yield the Complexity property

conclusion

- Building a cover tree on this net and invoking the separation property implies that each Voronoi region contains a small ball, and our probability bounds then supply the Concentration property
- The Voronoi regions from the cover trees are also bounded by balls, which implies the Centrality property
- The above eigenvalue bounds yield the Complexity property
- All the constants are explicit in terms of the reach, noise level, and volume of the noisy manifold

applications

application

Sloan Digital Sky Survey:

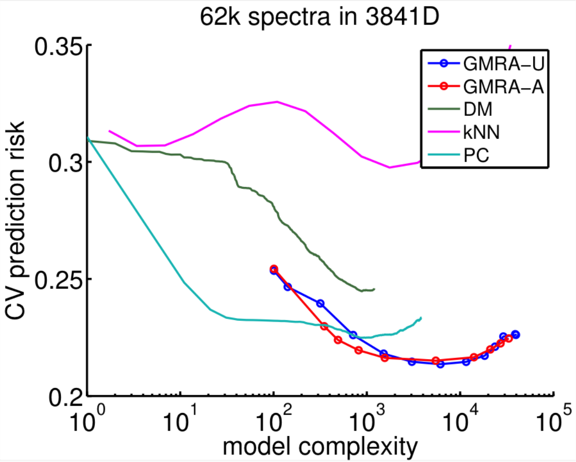
- Estimate redshift as a function of galactic spectra
- 62k data points in 3841 dimensional space
- Measure prediction risk using 10-fold cross validation
- Comparison of techniques:
    - GMRA using Uniform partitions
    - GMRA using Adaptive partitions
    - Diffusion Maps [3]
    - k-Nearest Neighbor regression
    - Principal Components

execution times

Timing (s)	Train		Test	
method	mean	std	mean	std
GMRA-U	673	22.8	72.3	21.9
GMRA-A	667	25.9	65.4	11.9
DM	6.02e4	4.93e3	1.10e3	254
kNN	8.07e3	537	16.4	2.47
PC	2.26e4	1.08e3	68.1	165

Table 1: Timing (in seconds) for the various methods.

complexity versus risk



Questions?

references I

W. K. Allard, G. Chen, and M. Maggioni.
Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis.
Applied and Computational Harmonic Analysis, 32(3):435–462, 2012.
(submitted:5/2011).

A. Beygelzimer, S. Kakade, and J. Langford.
Cover trees for nearest neighbor.
In *Proceedings of the 23rd international conference on Machine learning*, pages 97–104. ACM, 2006.

references II

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.

M. Maggioni, S. Minsker, and N. Strawn.
Multiscale dictionary learning: Non-asymptotic bounds and robustness.
Accepted to JMLR, 2015.

S. Minsker.
On some extensions of Bernstein's inequality for self-adjoint operators.
arXiv preprint arXiv:1112.5448, 2013.

references III

P. Niyogi, S. Smale, and S. Weinberger.
Finding the homology of submanifolds with high confidence from random samples.
Discrete & Computational Geometry, 39(1-3):419–441, 2008.