Geometric Multi-Resolution Analysis of data sets in high-dimensions

Multiscale geometry for the analysis of high dimensional datasets

William K. Allard, Guangliang Chen, Mauro Maggioni
Duke University
BOX 90320, NC, 27708

Abstract

Data sets are often modeled as point clouds in $\mathbb{R}^D$, for D large, but having some interesting low-dimensional structure, for example that of a d-dimensional manifold $\mathcal{M}$, with $d \ll D$. When $\mathcal{M}$ is simply a linear subspace, one may exploit this assumption for encoding efficiently the data by projecting onto a dictionary of d vectors in $\mathbb{R}^D$ (found by SVD), at a cost $(n + D)d$ for n data points. When $\mathcal{M}$ is nonlinear, there are no "explicit" constructions of dictionaries that achieve a similar efficiency: typically one uses either random dictionaries, or dictionaries obtained by black-box optimization. Such constructions, which typically cast the sparsity requirement as an optimization problem, suffer from many local minima and lack of theoretical guarantees. In this paper we construct data-dependent dictionaries based on a *geometric multiresolution analysis (GMRA)* of the data, inspired by multiscale techniques in geometric measure theory, to remedy the above deficiencies.

Multiscale decomposition. We start by constructing a multiscale nested partition of $\mathcal{M}$ into dyadic cells $\{C_{j,k}\}_{k\in\Gamma_j, 0\le j\le J}$ in $\mathbb{R}^D$. There is a natural tree $\mathcal{T}$ associated to the family: For any $j \in \mathbb{Z}$ and $k \in \Gamma_j$, we let $\text{children}(j, k) = \left\{k' \in \Gamma_{j+1} : C_{j+1,k'} \subseteq C_{j,k}\right\}$.

Multiscale SVD. For every $C_{j,k}$ we define the mean (in $\mathbb{R}^D$) by $\overline{c}_{j,k} := \mathbb{E}[x|x \in C_{j,x}]$ and the covariance by $\text{cov}_{j,k} = \mathbb{E}[(x - \overline{c}_{j,k})(x - \overline{c}_{j,k})^*|x \in C_{j,k}]$. Let the rank-$d$ SVD of $\text{cov}_{j,k}$ be $\text{cov}_{j,k} = \Phi_{j,k}\Sigma_{j,k}\Phi^*_{j,k}$. The subspace spanned by the columns of $\Phi_{j,k}$, and then translated to pass through $\overline{c}_{j,k}$, $\langle\Phi_{j,k}\rangle + \overline{c}_{j,k}$, is an approximate tangent space to $\mathcal{M}$ at location $\overline{c}_{j,k}$ and scale 2^{-j}. We define the coarse approximations, at scale j, to the manifold $\mathcal{M}$ and to any point $x \in \mathcal{M}$, as follows:

$$\mathcal{M}_j := \cup_{k\in\Gamma_j} P_{j,k}(C_{j,k}), \quad x_j := P_{j,k}(x),\, x \in C_{j,k}, \tag{1}$$

where $P_{j,k}$ is the associated affine projection to $C_{j,k}$.

Multiscale geometric wavelets. We can then introduce our wavelet encoding of the difference between $\mathcal{M}_j$ and $\mathcal{M}_{j+1}$, for $j < J$. These operators are low-dimensional "detail" operators analogous to the wavelet projections in wavelet theory, and satisfy, by construction,

$$P_{\mathcal{M}_{j+1}}(x) = P_{\mathcal{M}_j}(x) + Q_{\mathcal{M}_{j+1}}(x), \quad \forall\, x \in \mathcal{M}. \tag{2}$$

Geometric Wavelet Transforms (GWT). Given a GMRA structure, we may compute a discrete Forward GWT for a point $x \in \mathcal{M}$ that maps it to a sequence of geometric wavelet coefficient vectors:

$$q_x = (q_{J,x}, q_{J-1,x}, \ldots, q_{1,x}, q_{0,x}) \tag{3}$$

where $q_{j,x} := \Psi^*_{j,x}(x_j - c_{j,x})$. Note that, for a fixed precision $\epsilon > 0$, q_x has a maximum possible length $(1 + \frac{1}{2}\log_2\frac{1}{\epsilon})d$, which is independent of D and nearly optimal in d.

Sparsity The geometric wavelet dictionary may be constructed efficiently and is associated with efficient direct and inverse transforms. Depending on the geometric regularity of the data, it provides sparse (compressible) representations for data points.