Multiscale SVD and Geometric Multi-Resolution Analysis for noisy point clouds in high dimensions

Mauro Maggioni Duke University
Durham, NC, USA
mauro.maggioni@duke.edu

ABSTRACT

Data sets are often modeled as samples from a probability distribution in $\mathbb{R}^D$, for D large. It is often assumed that the data has some interesting low-dimensional structure, for example that of a d-dimensional manifold $\mathcal{M}$, with d much smaller than 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 $\mathbb{R}^D$ (for example found by SVD), at a cost $(n + D)d$ for n data points. When $\mathcal{M}$ is nonlinear, there are no "explicit" and algorithmically efficient constructions of dictionaries that achieve a similar efficiency: typically one uses either random dictionaries, or dictionaries obtained by black-box global optimization. The recent construction in [1] yields data-dependent multi-scale dictionaries that aim at efficiently encoding and manipulating the data. Their construction is fast, and so are the algorithms that map data points to dictionary coefficients and vice versa, in contrast with L^1-type sparsity-seeking algorithms, but alike adaptive nonlinear approximation in classical multiscale analysis. In addition, data points are guaranteed to have a compressible representation in terms of the dictionary, depending on the assumptions on the geometry of the underlying probability distribution.

We start by considering the problem of estimating the **intrinsic dimension** of data sets modeled as samples from a probability distribution supported on d-dimensional set $\mathcal{M}$ (in particular, a manifold) embedded in $\mathbb{R}^D$, in the regime $d \ll D$, and corrupted by high-dimensional noise. This setting has been recognized as important in various applications, ranging from the analysis of sounds, images (RGB or hyperspectral), to gene arrays, EEG signals, and other types of manifold-valued data, and has been at the center of much investigation in the applied mathematics and machine learning communities during the past several years. This has lead to a flurry of research on several problems, old and new, such as estimating the intrinsic dimensionality of point clouds [3], parametrizing sampled manifolds [2], constructing dictionaries tuned to the data or for functions on the data, and their applications to machine learning and function approximation. To this aim, we consider the singular values of the covariance matrix of the data restricted to Euclidean balls of radius r, as a function of r, and show that not only they contain useful information about local geometric properties of the data, including local intrinsic dimension and a certain stable notion of curvature, but they have very low sample complexity and are extremely robust to high-dimensional noise. We discuss applications to machine learning, hyper spectral imaging, and molecular dynamics [4].

Then we discuss the construction of **Geometric Multi-Resolution Analyses** for analyzing data sets as above [1]. We focus on obtaining multi-scale representations in order to organize the data in a natural fashion, and obtain efficient data structures for data storage, transmission, manipulation, at different levels of precision that may be requested or needed for particular tasks. This work ties with a significant amount of recent work in different directions: (a) Harmonic analysis and efficient representations of signals; (b) Data-adaptive signal representations in high dimensional spaces and dictionary learning; (c) Hierarchical structures for organization of data sets; (d) Geometric analysis of low-dimensional sets in high-dimensional spaces. This leads to multi scale transforms mapping data into compressible sets of coefficients, with associated fast transforms and sparse representations. An extension of compressive sensing to this framework will also be discussed.

BIO

I graduated from the Universitá degli Studi in Milano in 1999, and was a Ph.D. student of G. Weiss at the Washington University in St. Louis, from where I graduated in 2002. I am currently a professor of mathematics, and computer science, at Duke University working in harmonic analysis, probability and graph theory applied to the study of high-dimensional data sets. My web page contains recent publications, code, and various applications thereof: `http://www.math.duke.edu/~mauro`.

1. REFERENCES

[1] W. Allard, G. Chen, and M. Maggioni. Multiscale geometric methods for data sets II: Geometric wavelets. *Appl. Comp. Harm. Anal., accepted*, May 2011.

[2] P. Jones, M. Maggioni, and R. Schul. Universal local manifold parametrizations via heat kernels and eigenfunctions of the Laplacian. *Ann. Acad. Scient. Fen.*, 35:1–44, January 2010. http://arxiv.org/abs/0709.1975.

[3] A. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets I: Estimation of intrinsic dimension. *submitted*, 2010.

[4] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, (134):124116, 2011.

Analysis and CG June 18, 2012, Chapel Hill, North Carolina, USA.