D: Multiphysics and multiscale coupling
"Sensing multiscale structures in high-dimensional data"

B. W. Ong, Michigan Technological University, ongbw@mtu.edu
M. A. Iwen, Michigan State University
April 24, 2015

Motivation

Plasma behavior is well understood to span many temporal and spatial scales. Consequently, many well-resolved numerical simulations generate massive amounts of data, resulting in data management, analysis, and visualization challenges. This will clearly be exacerbated as we move towards exascale computations. A key observation that "big" in big data typically refers to a naive measure of size, for example, the number of (possibly adaptively selected) grid points in a simulation, or the number of time slices. In many instances, the "complexity" of the data, or perhaps more accurately the "complexity up to precision ϵ", is much smaller [3, 26, 10, 24, 12, 14, 11, 5, 6, 7]. In principle, this phenomenon should be exploited to reduce the computational cost of algorithms and aid in the data management of the simulation results. One approach is to approximate the data using low-dimensional geometric models (e.g., by low-dimensional manifolds [19, 20, 4, 2, 13, 26, 10, 1], or by a union of hyperplanes [17, 8, 25, 23, 9]).

Challenges/Opportunities:

1. The paradigm of compute, store, then analyze is no longer going to be feasible for large scale computations. In-situ/real time compression will be needed, requiring a significant cultural shift in how researchers handle simulation data. Researchers are going to demand guarantees about how well the low dimensional manifolds approximate the data, as well as the detection of anomalies before discarding any data. A software package that attempts to perform real time compression should construct sufficiently resolved low dimensional manifolds that satisfy the user specified tolerance, as well as provide error guarantees. This is especially challenging because the software also needs to be computationally efficient, and the algorithms will often only get "one pass" at the data.

2. There are surprisingly few existing libraries which provide data structures and functions for constructing low dimensional manifolds, certainly none of which are "mainstream" or production ready. A partial list includes:

   - scikit-learn [18]: Machine learning in Python – beta software still in development.
   - GMRA [16, 1]: MATLAB code base for finding multiscale low dimensional manifolds – not optimized, hard to incorporate into a larger project
   - SOM toolbox [22]: MATLAB code base using an approach called self-organizing maps – not actively developed.

   In preparation for large scale computing, researchers need access to production-ready toolboxes, libraries and software that are able to capture lower dimensional manifolds that can approximate the multiscale simulation data. Ideally, the software product will provide APIs for researchers to store, access and manipulate the manifolds that support the data.

3. Furthermore, the above software libraries are but a first step of what will eventually be required: a distributed (likely hierarchical) approach to constructing low dimensional manifolds, with error guarantees. A single data curation node that is on the tail end of a fire hose of data creates bottlenecks, and is impractical for exascale computations. Rather, a distributed (data-parallel) approach is needed, where multiple data curation tasks process locally available data to create local sketches, which are then merged to give a sketch of the entire data set. The mathematical foundations for analyzing the precision of merged manifolds are not well developed as yet. There are at least two distinct scenarios where merging manifolds are important:

   (a) "Pleasantly parallel" large scale computations. For example, imagine a set of smaller tightly-coupled simulations running independently to study the effect of electron collision; each simulation might utilize different cross-section data for electron collisions. The vast amount of data needs to be processed dynamically to provide useful information.

   (b) Often, there is a natural spatial decomposition upon discretization. Merging manifolds should be relatively straightforward in this scenario, although some care must be taken to provide reasonable error estimates..

4. Visualization tools will be needed to aid researchers analyze simulation results. There are many visualization platforms, for example the yt-project framework [21], a Python based analysis and visualization toolkit or paraview [15]. Such software will need "hooks" to process the (possibly non-linear) low dimensional manifolds. Ideally, a researcher will have access to a slider bar which tunes the scale of the multiscale structures which are being presented.

5. Time dependent multiscale features are probably best represented by time dependent low dimensional manifolds. Is there a way to evolve existing low-dimensional manifolds in a computationally efficient manner? How do error estimates and guarantees change in this scenario? How can this be leveraged to aid analysis, visualization and data management?

6. Can the reduced and associated orthogonal basis representing the (potentially time-dependent) low-dimensional manifold be used to inform the development of new simulation algorithms that are memory, communication, and computationally efficient? What sort of bias does one generate when using the reduced basis?

7. A multiphysics software has many different components and solvers to resolve the many spatial and temporal time scales. For example, one might use a Eulerian mesh with a fluid solver in a region where MHD approximations are valid, and a Lagrangian mesh with a direct summation solver in a region where kinetic effects are dominant. The software constructing the low dimensional manifold approximations will need to have the flexibility to process different data structures (possibly in overlapping regions) with differing precision, leading to potentially tricky decisions on how to weight data for the construction of the low dimensional manifolds.

References

[1] 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.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] M. Belkin and P. Niyogi. Using manifold structure for partially labelled classification. *Advances in NIPS*, 15, 2003.

[4] M. Brand. Charting a manifold. In *Advances in neural information processing systems*, pages 961–968, 2002.

[5] 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.

[6] 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.

[7] W. Cao and R. Haralick. Nonlinear manifold clustering by dimensionality. *ICPR*, 1:920–924, 2006.

[8] G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9(5):517–558, 2009.

[9] G. Chen and M. Maggioni. Multiscale geometric and spectral analysis of plane arrangements. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2825–2832. IEEE, 2011.

[10] 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. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[11] J. Costa and A. Hero. Learning intrinsic dimension and intrinsic entropy of high dimensional datasets. In *Proc. of EUSIPCO*, Vienna, 2004.

[12] D. L. Donoho and C. Grimes. When does isomap recover natural parameterization of families of articulated images? Technical Report 2002-27, Department of Statistics, Stanford University, August 2002.

[13] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[14] D. L. Donoho, O. Levi, J.-L. Starck, and V. J. Martinez. Multiscale geometric analysis for 3-d catalogues. Technical report, Stanford Univ., 2002.

[15] A. Henderson, J. Ahrens, and C. Law. *The ParaView Guide.* Kitware Clifton Park, NY, 2004.

[16] M. A. Iwen and M. Maggioni. Approximation of points on low-dimensional manifolds via random linear projections. *Inference & Information*, 2(1):1–31, 2013.

[17] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3):413–458, 2008.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, December 2000.

[20] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[21] M. J. Turk, B. D. Smith, J. S. Oishi, S. Skory, S. W. Skillman, T. Abel, and M. L. Norman. yt: A multi-code analysis toolkit for astrophysical simulation data. *The Astrophysical Journal Supplement Series*, 192(1):9, 2011.

[22] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orei, T. Honkela, and H. Lhdesmki. Self-organization and missing values in {SOM} and {GTM}. *Neurocomputing*, 147(0):60 – 70, 2015.

[23] R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, 2011.

[24] 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.

[25] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1927–1934. IEEE, 2010.

[26] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2002.