Research Statement

Henry Kvinge

1 Introduction

As data is generated at a faster and faster rate, tools for extracting information from large, high-dimensional data sets have become increasingly valuable across science, engineering, and industry. Mathematicians have a prominent role to play in helping to design such algorithms. My research interests focus on not only developing these algorithms, but also developing the mathematical framework necessary to utilize them to their full potential. To achieve these aims I am able to draw on experience working across all levels of the data science pipeline, from theory based projects that bring in tools more familiar from pure math to concrete collaboration with industry partners to produce software which actually processes real data in the field. Some of the projects that I describe below include:

- Development of a new family of "secant-based" dimensionality reduction algorithms for high-dimensional data and exploration of new statistics associated to these algorithms.
- Use of the geometry of Grassmann manifolds to address variation in signal in the endmember extraction problem in hyperspectral imaging.
- Testing and design of compressive sensing sampling and reconstruction algorithms for single-pixel imaging devices and development of a software package to run these on a hardware constrained device.

2 Secant-based dimensionality reduction

The secant set S of a data set $D \subset \mathbb{R}^n$ encodes information about the spatial relationships between points in D and consequently is a fundamental object to study when attempting to find projections of D into smaller dimension $k < n$ that preserve its structure. As was pointed out in [1], a projection $P : \mathbb{R}^n \to \mathbb{R}^k$ that preserves S is by definition bilipschitz on D and hence has a well-conditioned inverse $P^{-1} : P(D) \to \mathbb{R}^n$ for de-compression of $P(D)$. Similarly, by virtue of P being bilipschitz, P preserves the intrinsic dimension d of D. The dimension d is a fundamental statistic associated to D which can inform key decisions related to how D is handled and analyzed, so this is often a feature worth preserving.

In [2] we proposed a new algorithm called the *secant avoidance projection (SAP) algorithm* which takes as input a secant set S, a dimension $k < n$, and outputs a projection that is a local solution to the optimization problem:

$$\underset{P \in \text{Proj}(n,k)}{\arg\max} \min_{s \in S} ||Ps||_{\ell_2} \tag{1}$$

where $\text{Proj}(n, k)$ is the set of all projections from $\mathbb{R}^n$ onto a k-dimensional subspace. This flexible, iterative algorithm is designed with a GPU architecture in mind for fast convergence on even very high-dimensional data sets. It should be noted that the optimization problem (1) is distinct from the optimization problem solved by PCA for example, and hence the output of SAP is generally quite different from this more familiar approach.

Because S grows as $O(r^2)$ for r the number of data points in D, storing all secants is often impractical. One option in such situations is to calculate and store a random subset of secants. This however ignores the structure of D and we can thus miss key features in the original data set. We developed a variation on SAP called the *hierarchical secant avoidance projection (HSAP)*

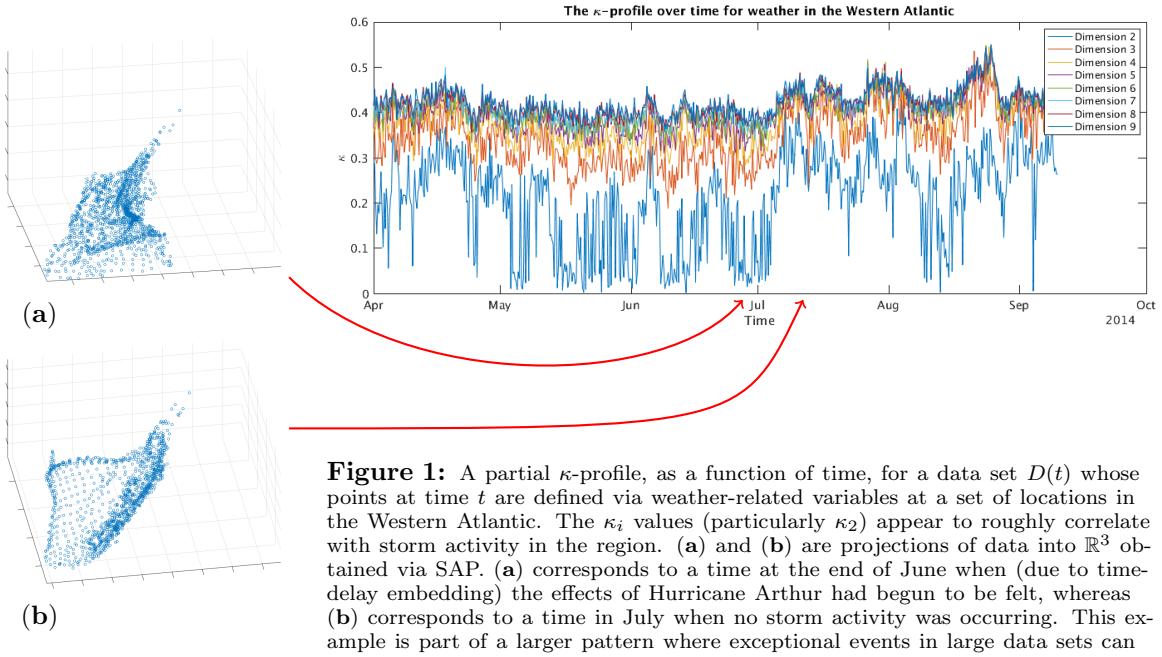

Figure 1: A partial κ-profile, as a function of time, for a data set $D(t)$ whose points at time t are defined via weather-related variables at a set of locations in the Western Atlantic. The κ_i values (particularly κ_2) appear to roughly correlate with storm activity in the region. (**a**) and (**b**) are projections of data into $\mathbb{R}^3$ obtained via SAP. (**a**) corresponds to a time at the end of June when (due to time-delay embedding) the effects of Hurricane Arthur had begun to be felt, whereas (**b**) corresponds to a time in July when no storm activity was occurring. This example is part of a larger pattern where exceptional events in large data sets can be detected via the κ-profile and reflect changes in the intrinsic dimension of the data.

.

algorithm to address this issue [3]. HSAP utilizes the natural structure of D to inform subsampling and approximation of secants to vastly reduce the number of secants while still retaining key features of the data.

Finally, part of the output of both SAP and HSAP is the length of the projected secant least well preserved by the final output projection. If P^* is the output projection from $\mathbb{R}^n$ to a k-dimensional subspace, we write $\kappa_k := \min_{s \in S} ||P^* s||_{\ell_2}$. By computing κ_i over a range of $1 \leqslant i \leqslant n$, we arrive at a statistic which we call the *κ-profile*. When the data sits on an m-dimensional manifold in $\mathbb{R}^n$ for $m < n$, the κ-profile can be related to m via the constructive proof of Whitney's embedding theorem. In [4], we investigated how the κ-profile can capture fundamental structure in D. This is particularly relevant when $D(t)$ is a data set that is changing with respect to a time parameter t. It seems that fundamental changes in $D(t)$, irregardless of the nature of the data, are often reflected in dramatic changes in $\kappa(t)$. See Figure 1 for a partial κ-profile of an Atlantic weather dataset as a function of time. The drops in the κ_2 value from July to September seem to reflect Hurricane events.

Some projects that I am either working on or plan to work on include:

- *Deeper development of the theory of secant-based dimensionality reduction algorithms:* We are currently investigating properties of the optimization problem (1), including the relationship between solutions of (1) for different values of k.
- *Deeper investigation of the κ-profile:* The κ-profile has clear connections to dimension of a data set via the constructive proof of Whitney's embedding theorem for example. We would like to understand

3 Searching for pure signals in hyperspectral imagery via the geometry of the Grassmannian

Unlike RGB images, which only sample from three different spectral bands of light (red, green, and blue), hyperspectral imagery can sometimes sample more than 200. While hyperspectral images thus

have strong discriminatory power, it can be difficult to extract information from them because they are often very high dimensional. One particularly important problem is understanding how to isolate the "most pure" signals from a hyperspectral image. Geometrically, this corresponds to picking out those spatial locations in the image whose spectral curve sits on the convex hull formed by all spectral curves in the image. In [5] we show how if one considers local patches rather than individual pixels, this endmember extraction problem is most naturally realized on a Grassmann manifold. We propose an algorithm that solves the problem in this setting.

It has become increasingly clear that finding the right geometric setting to capture the natural structure of a data set can result dramatic improvements in our ability to extract the relevant information from the data. Our paper [5] is another piece of evidence that many traditional signal detection and data analytics techniques for hyperspectral imaging would be better adapted to the setting of Grassmannian or flag manifolds (see also [6]). In future work we plan to examine further examine how geometry might be used to better utilize hyperspectral data.

4 Compressive sensing algorithm design and implementation for a single-pixel camera[1]

I am have been a primary contributor to algorithm and data pipeline development for two projects with industry partner Physical Science Inc., which aim to produce innovative, low-cost imaging devices using a single-pixel architecture. The goal of the first project is to develop a single-pixel compact infrared flash 3D imaging sensor capable of capturing a stream of highly accurate depth images, while the goal of the second is to develop a single-pixel hyperspectral imaging device with real-world chemical plume detection capability.

My contributions to these projects center around algorithm design and testing, software development, and integration of a unified data processing system that takes as input compressively sensed depth (respectively, hyperspectral) image data and returns as output fully-reconstructed and post-processed depth (respectively hyperspectral) images.

In order to reconstruct depth and hyperspectral images accurately with limited sampling, we have developed an innovative, data-driven approach to sampling which differs significantly from the standard theoretical framework used to understand sampling matrix optimization [7]. A description of this approach is currently under review [8]. In fact, one consequence in the hyperspectral setting we have observed that CS sampling and reconstruction can actually amplify chemical signals [9].

Finally, the main consideration when producing a software package was choosing a framework in which our algorithms would run quickly. Choosing the split Bregman method was one way of addressing this challenge. The other way that we ensured fast data processing was by implementing nearly all algorithms in Nvidia's GPU programming language CUDA. Since most steps in our algorithms are either linear algebra operations or parallelizable, this step was essential.

Some new projects that I am either currently working on or plan to work on in the future include:

- *Schubert varieties and attribute discovery for machine learning*: One foundational problem in machine learning and data analysis is attribute discovery within a data set. We have evidence to suggest that when data can be represented on a Grassmann manifold, then Schubert varieties are a natural framework in which to differentiate fundamental attributes associated. We currently in the process of not only making this association rigorous, but also adapting various optimization algorithms from Euclidean space to the Grassmannian framework with Schubert constraints.
- *Decomposition of the regular representation and dimensionality reduction*: The multidimensional scaling (MDS) algorithm is a foundational algorithm which can be used both for dimensionality reduction and more generally for realization of a finite configuration of points from an abstract metric space in Euclidean space. It is thus a basic algorithm used in many areas of science and engineering. In the special case when the points of interest live on a group, then the algorithm can be given a representation theoretic interpretation. This observation both sheds new light on the algorithm itself and suggests alternative implementations.

[1] Work described in this section is ITAR-protected and hence some details have been omitted.

References

[1] D. Broomhead and M. Kirby, "A new approach for dimensionality reduction: Theory and algorithms," *SIAM J. of Applied Mathematics*, vol. 60, no. 6, pp. 2114–2142, 2000.

[2] H. Kvinge, E. Farnell, M. Kirby, and C. Peterson, "A GPU-oriented algorithm design for secant-based dimensionality reduction," in *2018 17th International Symposium on Parallel and Distributed Computing (ISPDC)*, June 2018, pp. 69–76.

[3] ——, "Too many secants: a hierarchical approach to secant-based dimensionality reduction on large data sets," in *to appear in Proceedings of the 2018 IEEE High Performance Extreme Computing Conference*. IEEE, 2018.

[4] ——, "Monitoring the shape of weather, soundscapes, and dynamical systems: a new statistic for dimension-driven data analysis on large data sets," in *accepted for publication in the Proceedings of the IEEE International Conference on Big Data 2018*, August 2018.

[5] E. Farnell, H. Kvinge, M. Kirby, and C. Peterson, "Endmember extraction on the Grassmannian," in *2018 IEEE Data Science Workshop (DSW)*, June 2018, pp. 71–75.

[6] T. Marrinan, J. R. Beveridge, B. Draper, M. Kirby, and C. Peterson, "Flag-based detection of weak gas signatures in long-wave infrared hyperspectral image sequences," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXII*, vol. 9840, May 2016, p. 98401N.

[7] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, p. 969, 2007. [Online]. Available: http://stacks.iop.org/0266-5611/23/i=3/a=008

[8] E. Farnell, H. Kvinge, M. Kirby, and C. Peterson, "A data-driven approach to sampling matrix selection for compressive sensing," in *abstract submitted to SPIE: Defense + Commercial Sensing*, 2018.

[9] H. Kvinge, E. Farnell, M. Kirby, and C. Peterson, "More chemical detection through less sampling: amplifying chemical signals in hyperspectral data cubes through compressive sensing," in *abstract submitted to SPIE: Defense + Commercial Sensing*, 2018.