Monitoring the shape of weather, soundscapes, and dynamical systems: a new statistic for dimension-driven data analysis on large datasets

Henry Kvinge
Department of Mathematics
Colorado State University
Fort Collins, CO USA
henry.kvinge@colostate.edu

Elin Farnell
Department of Mathematics
Colorado State University
Fort Collins, CO USA
elinfarnell@gmail.com

Michael Kirby
Department of Mathematics
Colorado State University
Fort Collins, CO USA
kirby@math.colostate.edu

Chris Peterson
Department of Mathematics
Colorado State University
Fort Collins, CO USA
peterson@math.colostate.edu

***Abstract*—Dimensionality-reduction methods are a fundamental tool in the analysis of large datasets. These algorithms work on the assumption that the "intrinsic dimension" of the data is generally much smaller than the ambient dimension in which it is collected. Alongside their usual purpose of mapping data into a smaller-dimensional space with minimal information loss, dimensionality-reduction techniques implicitly or explicitly provide information about the dimension of the dataset.**

In this paper, we propose a new statistic that we call the kappa-profile for analysis of large datasets. The kappa-profile arises from a dimensionality-reduction optimization problem: namely that of finding a projection that optimally preserves the secants between points in the dataset. From this optimal projection we extract kappa, the norm of the shortest projected secant from among the set of all normalized secants. This kappa can be computed for any dimension k; thus the tuple of kappa values (indexed by dimension) becomes a kappa-profile. Algorithms such as the Secant-Avoidance Projection algorithm and the Hierarchical Secant-Avoidance Projection algorithm provide a computationally feasible means of estimating the kappa-profile for large datasets, and thus a method of understanding and monitoring their behavior. As we demonstrate in this paper, the kappa-profile serves as a useful statistic in several representative settings: weather data, soundscape data, and dynamical systems data.

***Keywords*-Dimension of data, secant sets, dynamical systems, dimensionality reduction, big data.**

I. Introduction

As high-dimensional data becomes more and more plentiful, dimensionality-reduction algorithms become an increasingly important tool for any researcher seeking to extract meaningful information from their data. Indeed, it is not unusual to find that data collected in an n-dimensional space is intrinsically only k-dimensional, where $k \ll n$. A good dimensionality-reduction algorithm will find a map from $\mathbb{R}^n$ to $\mathbb{R}^{k'}$ (for some k' close to k) while preserving fundamental properties such as distances between data points. This process is essential since data in low-dimensional space is often computationally easier to store and manipulate, and hence a wider array of algorithms are available for data analytics. Furthermore, the process of reduction often coincides with the production of a more appropriate representation of the data, making many data analytics algorithms more successful as a result of a meaningful feature space.

By systematically studying how well a dataset D in $\mathbb{R}^n$ can be projected to $\mathbb{R}^m$ for $m < n$, as m varies, we can uncover geometric properties of D (such as the intrinsic dimension of an underlying manifold on which D approximately sits). In this paper we will explore this idea in the context of secant-based dimensionality-reduction algorithms, a family of dimensionality-reduction algorithms that use the secant set S of D to find projections that best preserve distances between points. We will focus on the Secant-Avoidance Projection (SAP) dimensionality-reduction algorithm [1]. By finding SAP projections for various $m < n$, we return a statistic which we call the κ-profile.

The problem of calculating the dimension of a dataset has been addressed from various perspectives. A classical approach to estimating dimension is to use a linear method such as principal component analysis (PCA). However, such methods do not capture the dimension of non-linear data well. Other methods that do not assume linearity have also been proposed, see, e.g., [2] - [8].

Though it is related to dimension, the κ-profile often carries more information such as how well the data fits into many different reduced dimensions simultaneously. This is useful for studying real-world data, which rarely conforms precisely to a manifold. Estimating the κ-profile from a dataset is relatively easy and requires very few assumptions about the data. It is thus a useful statistic, particularly in cases in which domain-specific tools are limited.

The κ-profile can be used not only for analysis of a static dataset, but also for settings such as time-series analysis and anomaly detection. In particular, the sensitivity of the κ-profile to changes in the geometry of a dataset makes it a prime candidate as an indicator of fundamental changes in the behavior of an underlying system as a function of either

This paper is based on research partially supported by the National Science Foundation under Grants No. DMS-1513633, and DMS-1322508 as well as DARPA awards N66001-17-2-4020 and D17AP00004.

time or a set of parameters. In fact, by utilizing a time-delay embedding, the κ-profile can in some cases also give indication of the dimension of the underlying dynamics from which the data is drawn.

This paper is organized as follows. In Section II we review dimension of a dataset and secant-based dimensionality-reduction algorithms, we define the κ-profile, and we provide a short illustrative example. We also include a brief review of geodesic distance on a Grassmann manifold and ways to format a collection of time-parametrized datasets prior to calculating the κ-profile. In Sections III and IV, we calculate the κ-profile on weather data and ambient noise data, respectively. In Section V, we discuss a synthetic example where the dimension of the data, the solution set of a well-known partial differential equation, is already approximately known.

II. Background

A. Dimension Estimation

We follow the background on dimension estimation from [9]. The motivation for consideration of dimension is that locally a dataset often has hidden constraints that allow one to consider the data as a noisy sampling of some underlying manifold. When this underlying manifold is d dimensional, we say that the data is d-dimensional. Locally, a d-dimensional manifold can be parameterized by d free variables. Dimensionality is a coarse measure of manifold complexity. The estimation of dimension from data has been addressed by numerous authors including, e.g., [2] - [11].

An important result for characterizing dimension-preserving transformations revolves around the definition of a *bi-Lipschitz* function. A function $f : X \to Z$, for $X \subseteq \mathbb{R}^n$, $Z \subseteq \mathbb{R}^m$, is said to be *bi-Lipschitz* on X if there exist $a, b > 0$ such that for all $x, y \in X$

$$a\|x - y\|_{\ell_2} \leq \|f(x) - f(y)\|_{\ell_2} \leq b\|x - y\|_{\ell_2}. \qquad (1)$$

Thus, pairs of points neither collapse nor are blown apart by application of f. When f is a projection, the class of functions considered in this paper, we have $b = 1$. Asking for $a > 0$ satisfying (1) is equivalent to asking that f preserve secants. Such a goal is justified by the fact that if f is bi-Lipschitz then we know $\dim(X) = \dim(Z)$, where the dimension can be taken as the topological dimension, or the Hausdorff dimension; see [12] for details. Thus projection-based algorithms that maximally avoid decreasing the length of secants are, in some sense, optimally dimension-preserving. An additional argument for this approach, based on invoking Whitney's easy embedding theorem, is made in [13].

B. The SAP algorithm and κ-profiles

Following the context in [1], data residing in a high-dimensional ambient space can both be a computational burden and difficult to analyze. Data-reduction algorithms offer a way to reduce these difficulties by mapping the dataset into a lower-dimensional space with the goal of retaining as much information as possible. A classical example of a data-reduction algorithm for this context is PCA. In [11], [13], [14], Broomhead and Kirby developed a new framework for data-reduction which focuses on preserving the normalized secant set

$$S := \left\{ \frac{x - y}{||x - y||_2} \mid x, y \in D \right\}$$

corresponding to a dataset $D \subset \mathbb{R}^n$. When successful, projections produced by such methods not only retain differential structure but also have a well-conditioned inverse. This property, not necessarily found for projections obtained from other popular methods such as PCA, means that the projection provides not only a method to compress the data, but also to decompress the data without loss of information.

In practice, producing projections from a dataset $D \subset \mathbb{R}^n$ into $\mathbb{R}^m$ for $m < n$ amounts to finding projections, P, for which the smallest value of $||P(s)||_{\ell_2}$ over all $s \in S$ is maximized. In [1] the authors propose the SAP algorithm for producing a projection that best preserves the secant set of a dataset.

The SAP algorithm proceeds through an iterative procedure. An outline of the algorithm is as follows. Let $D \subset \mathbb{R}^n$ be a dataset and define S to be the set of all normalized secants of D. Initialize a projection from $\mathbb{R}^n$ to $\mathbb{R}^m$. At iteration i, use the current projection to compute the projection of the secant set S and determine the secant vector that is least well-preserved, i.e. has the shortest projection. Define the $(i + 1)$-th projection by rotating the i-th projection subspace toward the orthogonal complement of the projection of the shortest projected secant. The $(i + 1)$-th projection is the projection corresponding to this rotated m-dimensional subspace.

For a version of SAP that addresses big data settings, see the Hierarchical Secant-Avoidance Projection (HSAP) algorithm [9]. Both SAP and HSAP are polynomial-time algorithms [1], [9], and SAP has been shown to be straightforward to implement on a GPU [1]. In general, the computational complexity for the κ-profile will be highly dependent on the choice of secant-based dimensionality-reduction technique.

The key mathematical object of this paper is a κ-profile.

Definition II.1. Let D be a finite set of points in $\mathbb{R}^n$, and let S be the set of normalized secants for D. For fixed $m < n$, define $\mathcal{P}_m$ to be the collection of matrices in $\mathbb{R}^{n \times m}$ with orthonormal columns (equivalently the set of all orthogonal projections from $\mathbb{R}^n$ into an m-dimensional subspace). If

$$P_m^* = \underset{P \in \mathcal{P}_m}{\arg\max} \left(\min_{s \in S} \|P^T s\|_{\ell_2} \right), \text{ then}$$

$$\kappa_m = \min_{s \in S} \|(P_m^*)^T s\|_{\ell_2}.$$

One may then construct a tuple of κ_m values for a range of m. Such a tuple $(\kappa_{m_1}, \kappa_{m_2}, \ldots, \kappa_{m_\ell})$ is a κ-*profile*. Note that in this definition, $m_1 \geq 1$ and $m_\ell \leq n$.

Intuitively, if $\mathcal{M}$ is a manifold from which data is drawn, the κ-profile provides a measure of how successfully a projection P^* embeds $\mathcal{M}$ into $\mathbb{R}^m$ for varying m. Not only does this serve as a means of extracting information about the dimension of the underlying manifold, the κ-profile itself contains useful

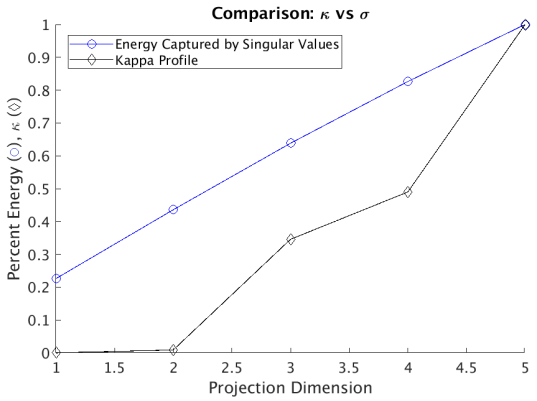

Fig. 1. A comparison of the κ-profile and the cumulative energy captured by the singular values for a set of points from the trigonometric moment curve in $\mathbb{R}^5$.

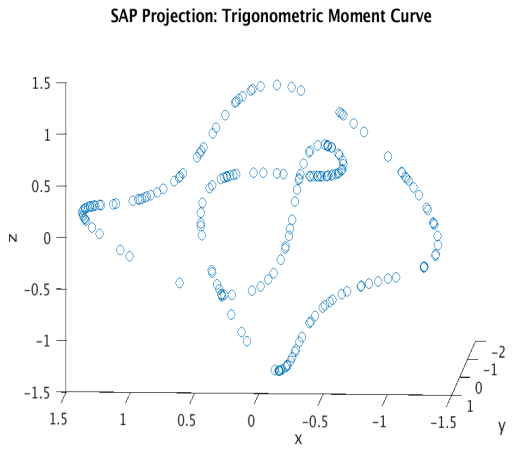

Fig. 2. Projection of points from trigonometric moment curve in $\mathbb{R}^5$ to $\mathbb{R}^3$ using SAP.

information about the nature of the dataset. We demonstrate, for example, in Section III that the κ-profile reflects important features in data, such as the changing characteristics of weather data during the presence or absence of extreme weather events.

Throughout this paper, we use the SAP algorithm to estimate the κ-profile in various settings; for simplicity we refer to the estimated profile as the κ-profile.

C. An illustrative synthetic example

We calculate the κ-profile for a small example. We randomly sample 192 points from the trigonometric moment curve $\phi : \mathbb{R} \to \mathbb{R}^5$ defined by $\phi(t) = (\cos(t), \sin(t), \cos(2t), \sin(2t), \cos(3t))$. While $\phi(\mathbb{R})$ is fundamentally 1-dimensional, it is not contained in any nontrivial subspace of $\mathbb{R}^5$. In Figure 1 we compare the κ-profile obtained from application of SAP against the cumulative energy captured by the singular values of the data $\left(\text{for each } k \text{ from 1 to 5} : \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{5} \sigma_i^2}\right)$. By heuristic, if $\kappa > 0.2$, we consider the embedding to be good. This choice is based on experiments where we calculated the κ-profile for data sets drawn from manifolds of known dimension (see [1, Section VI] for example). Notice that the κ-profile alone tells us that data is either 1, 2, or 3 dimensional. On the other hand, the energy distribution from the singular values is indistinguishable from what one might see from a draw from a 5-dimensional Gaussian distribution. The projection of our sample into $\mathbb{R}^3$ (obtained via SAP), is pictured in Figure 2.

D. Geodesic Distance on the Grassmann Manifold

In this paper, we use the notion of geodesic distance on a Grassmann manifold to compare the projection returned by the SAP algorithm to the common dimensionality-reduction technique of PCA. We briefly review the Grassmannian and distance metrics on this manifold.

As described in [9], the *Grassmannian* $Gr(k, n)$ is a manifold whose points parametrize the k-dimensional subspaces of a fixed n-dimensional vector space. Distance metrics on this manifold are often given via principal angles between corresponding vector spaces; such metrics are orthogonally invariant. Principal angles are defined below and are readily computable through a singular value decomposition.

Consider the subspaces U and V of a vector space $\mathbb{R}^n$ and let $q = \min\{\dim U, \dim V\}$. The principal angles between U and V are the angles $\theta_1, \theta_2, \ldots, \theta_q \in [0, \frac{\pi}{2}]$ between pairs of principal vectors $\{u_k, v_k\}$ with $u_1, \ldots, u_q$ a distinguished orthonormal set of vectors in U and $v_1, \ldots, v_q$ a distinguished set of orthonormal vectors in V. These vectors are obtained recursively, for each $1 \leq k \leq q$, by

$$\cos\theta_k = \max_{u \in U, v \in V} u^T v = u_k^T v_k$$

subject to $||u||_2 = ||v||_2 = 1$, $u^T u_i = 0$ and $v^T v_i = 0$ for $i = 1, 2, \ldots, k-1$.

The geodesic distance between $X, Y \in Gr(q, n)$ is defined in terms of the principal angles, $\theta_1, ..., \theta_q$, between the vector spaces represented by X and Y as

$$d_{geodesic}(X, Y) = \sqrt{\theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}.$$

While there are many other interesting orthogonally invariant metrics, we utilize the geodesic distance for this paper.

E. The κ-profile and time series

The κ-profile can provide information about the dynamics of data over time. Suppose $\{D_t\}$ is a collection of datasets in $\mathbb{R}^n$ parametrized by time parameter $t \in \mathbb{Z}_{\geq 0}$, with bijective maps $f_t : D_t \to D_{t+1}$. Then we can identify points in D_t and D_{t+1}, so that $f_t(x) \in D_{t+1}$ is the same point as $x \in D_t$, but one time step later.

Given such a collection $\{D_t\}$ there are various ways of calculating the κ-profile depending on how we structure the input data. One approach is to create a κ-profile for each D_t independently. That is, the secant set we apply our dimensionality-reduction algorithm to is precisely the secant set of points $D_t \subset \mathbb{R}^n$ for each $t \in \mathbb{Z}_{\geq 0}$. This captures

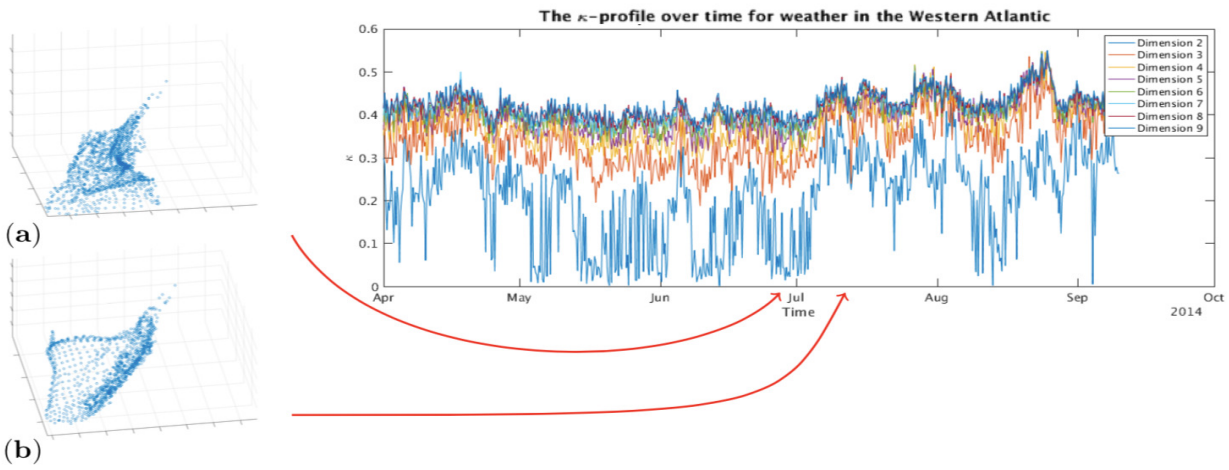

Fig. 3. The κ-profile for weather data from the Western Atlantic in 2014 from the grid ($20-33°$N, $44-69°$W). The first hurricane of the year, Hurricane Arthur, approached the grid in the beginning of July, but because of the time-delayed embedding, the effect of the storm should be seen near the end of June. We suspect the drop in the κ-value corresponding to projection into 2 dimensions which occurs around the end of June is probably attributable to this. Similarly, the drop in the same κ-value in mid-August and early September are possibly related to Hurricane Cristobal and Hurricane Edouard, respectively. (a) and (b) show projections of points from two sets from $\{D_{A,t}\}$ as indicated by the arrows.

geometric properties of $\{D_t\}$ as t changes. When we also want a single κ-profile to capture temporal changes in the data, an alternative approach is to construct a new sequence of datasets $\{D_t^{(\ell)}\}$ such that

$$D_t^{(\ell)} := \{[x, f_t(x), f_{t+1}(x), \ldots, f_{t+\ell}(x)] \in \mathbb{R}^{n\ell} \mid x \in D_t\}$$

where the brackets indicate vector concatenation. This type of construction is often known as a *time-delay embedding*.

Motivation for this approach is given by Takens' theorem [15]. Given a discrete dynamical system where the state space is an m-dimensional compact, smooth manifold M, evolution of the system can be defined as a smooth map $f : M \to M$. Suppose that we only have access to a single smooth measurement function $\varphi : M \to \mathbb{R}$. Takens' theorem says that under suitably general conditions, we can construct a diffeomorphic copy of the manifold M in $\mathbb{R}^k$ via a length $k \geq 2m+1$ time-delay embedding of φ with respect to f, i.e. we define the embedding by sending x to $[\varphi(x), f(\varphi(x)), \ldots, f^k(\varphi(x))]$.

Hence when considering the problem of monitoring a large collection of data that is changing over time, it is useful to use time-delay embeddings so that we can monitor both the geometry and dynamics of the data. Of course in the real world, we rarely know the dimension of the manifold M. It is thus an important problem to estimate the value $2m+1$, or the shortest time-delay embedding that will fully capture the dynamics. Algorithms which address this problem include false nearest neighbors [16], and saturation methods [17]. In the spirit of the latter, the κ-profile should provide a new tool to approach this problem. We intend to explore this idea in future research.

III. Example: Weather Data

Statistics arising from measurements of the Earth's weather are one of the most fertile sources of big data available. In this section we provide an example of how large-scale regional weather patterns are reflected in the κ-profile of weather datasets. We obtained historical weather model data from the National Center for Environmental Prediction (NCEP) Climate Forecast System (CFS), version 2 [18]. We took two rectangular grids, with data points at each .5° of latitude/longitude:

- The first is in the Western Atlantic ($20-33°$N, $44-69°$W), northeast of the Caribbean and in the path of Atlantic hurricane activity, from April 1st, 2014, to October 1st, 2014 with measurements every 6 hours.
- The second is located in the Western Pacific ($18-26°$N, $123-132°$E) east of Taiwan in the line of some of the Pacific typhoon activity, from July 1st, 2015, to December 1st, 2015, with measurements every 6 hours.

The data contains 9 measurements at each grid point, such as pressure, wind speed, air temperature, and precipitation rate. Thus, before time-delay embedding, we consider points in these datasets to be in $\mathbb{R}^9$. We used a 19 step time-delay embedding in order to construct new datasets $\{D_{A,t}\}$ from the Western Atlantic data and $\{D_{P,t}\}$ from the Western Pacific data, from which we produced the κ-profiles that we analyze below. Note that each set $D_{A,t}, D_{P,t} \subset \mathbb{R}^{171}$ corresponds to a 114-hour time window.

In Figures 3 and 4, we provide κ-profiles for $\{D_{A,t}\}$ and $\{D_{P,t}\}$. We see that in both cases, elements of $\{D_{A,t}\}$ and $\{D_{P,t}\}$ fluctuate between being 2 and 3 dimensional. This is further supported by examining projections of the data, obtained by the SAP algorithm (see Figure 3 for example). Rough comparisons of these figures with the record of storms

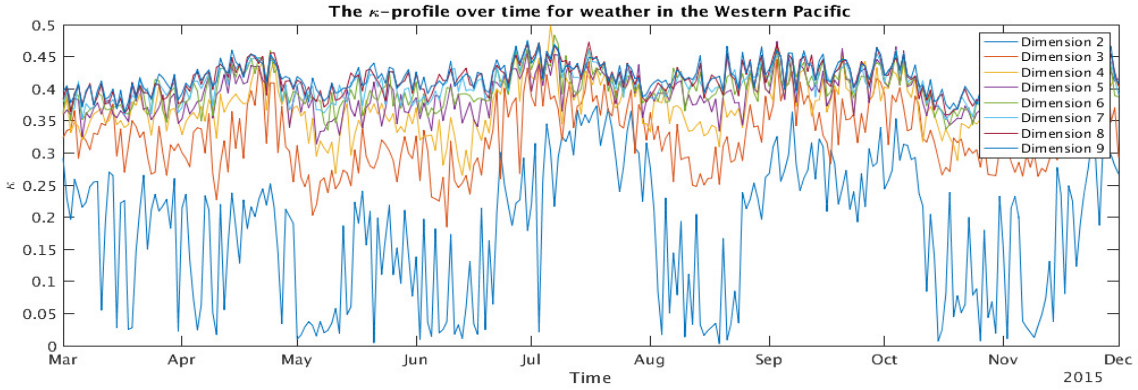

Fig. 4. The κ-profile for weather data taken from the rectangle $(18-26^{\circ}\text{N}, 123-132^{\circ}\text{E})$ in the Western Pacific, east of Taiwan. We suspect that the sustained drop in the κ-value associated with projection into 2 dimensions beginning in May may be related to Typhoon Dolphin, which moved through the grid during this time (taking into account the time-delay embedding). The drop in κ-value in early August may be attributable to Typhoon Soudelor.

(hurricanes and typhoons, respectively) suggests that in general the κ-values decrease as a storm approaches (indicating an increase in the dimension of the data), and increase again when no major weather events are occurring. We hypothesize, for example, that the drop in κ-values in Figure 3 in late-June/early-July is related to Hurricane Arthur which passed by the grid near this time (taking into account the time-delay embedding).

While the κ-profiles shown in Figures 3 and 4 seem to capture historical weather activity in these areas during the given time range, the correlation is not perfect. For example, one would expect a drop in κ-values in Figure 4 during Typhoon Dujuan which passed through the grid in mid-September. It seems likely that such discrepancies could be understood if a direct connection between the physics of weather and changes in dimension of the corresponding datasets was known. This is an avenue for further research.

IV. Example: soundscapes

All environments produce ambient noise which serves as a continuous stream of information that reflects the state of the environment. We ran the SAP algorithm on audio recordings from five different locations/states: (1) an area experiencing heavy rain (2) a forest, (3) a London street, (4) a train station, and (5) a sea shore [19]. The audiofiles were originally recorded at a sampling rate of 48,000 with 2 channels. We resampled these at 1/100 this rate and took a point to be a length 5,000 window. For each environment then, the corresponding dataset D consists of some number of points in $\mathbb{R}^{10,000}$. We selected overlapping windows in order to capture a maximum amount of temporal information about the soundscape.

We used the SAP algorithm to calculate the κ-profile for each of these soundscape datasets (Figure 5). We see that the heavy rain and sea shore soundscape appear to be quite high-dimensional. This is perhaps unsurprising, as both of these soundscapes consist of relatively uniform, incoherent noise. On the other hand, the datasets for soundscapes corresponding to a train station and a London street appear to have lower dimension. This fits with the nature of urban ambient noise which generally has more structure.

While we did not do it here, it would be easy to monitor a continuous sound recording in a manner similar to that found in Section III. As indicated by Figure 5, such a set-up should capture fundamental changes in the soundscape through the κ-profile. For example, a sudden downpour in the London street environment would correspond to a sharp drop in κ-values in the κ-profile.

V. Example: the Kuramoto-Sivashinsky equation

The Kuramoto-Sivashinsky (KS) equation [20] in one spatial dimension is the partial differential equation

$$u_t + vu_{xxxx} + u_{xx} + \frac{1}{2}(u_x)^2 = 0,$$

where v is a positive constant and $u : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a function that satisfies an L-periodic initial condition $u(x+L,0) = u(x,0)$. See [21] and [22] for motivation and applications. For $u(x,0)$ satisfying the initial condition, a unique solution exists and is bounded. These solutions have coherent spatial structure but exhibit temporal chaos. For different initial conditions and periodicity values L, the solution manifold has varying dimension. This makes data collected from the KS equation ideal for studying algorithms which seek to capture dimensionality and related statistics.

Following [20], the version of the KS equation that we use to generate data is

$$u_t + 4u_{xxxx} + \alpha\Big(u_{xx} + \frac{1}{2}(u_x)^2\Big) = 0,$$

where the periodicity has been set at 2π, $v = 4$, and a bifurcation parameter α has been introduced. The stable solution manifolds for various values of α were described via numerical experiment in [20].

We now consider some examples with varying values of α and hence dimension. In each case, we generate a dataset from the solution manifold consisting of 10,000 points in $\mathbb{R}^{32}$. In several of these examples, we also provide a comparison with the PCA approach to dimensionality-reduction for context.

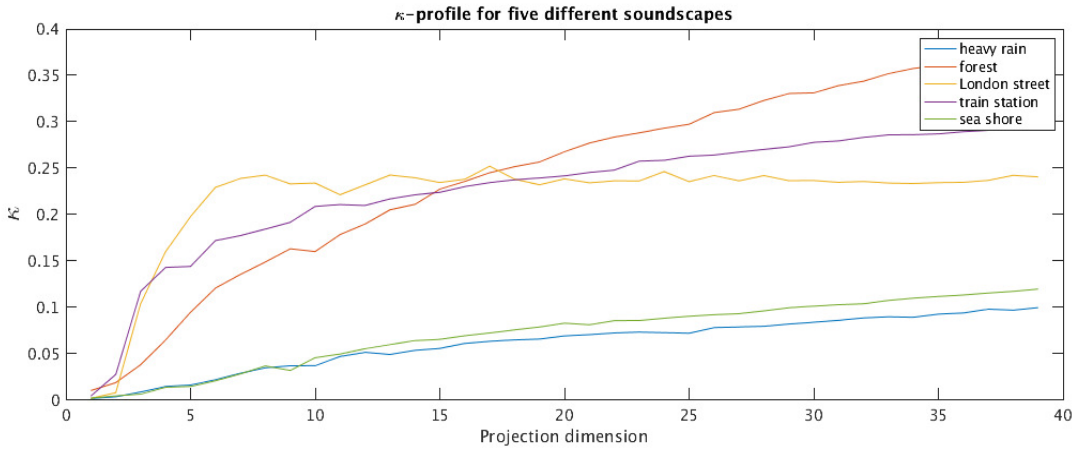

Fig. 5. The κ-profile of 5 soundscapes. The dimension of heavy rain and the sea shore appear to be much higher dimensional than the other environments.

Let us begin with $\alpha = 19$; the solution is a periodic orbit [20]. In Figure 6, we show the κ-profile for projection dimensions 1 to 20 for this value of α as well as several others. In this case, we can reasonably embed the data in $\mathbb{R}^1$. It is worth noting as well that the projection into $\mathbb{R}^7$ is nearly isometric on the normalized secant set since the value of κ is very close to one.

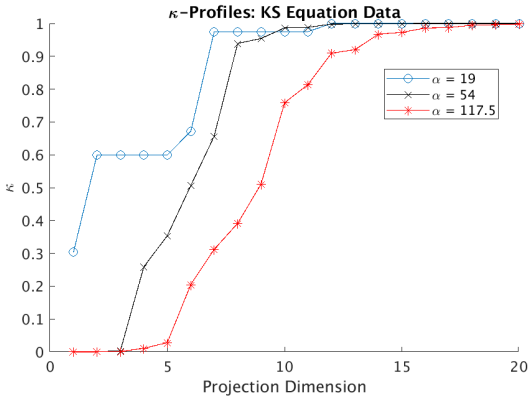

Fig. 6. κ-profiles for the KS equation data with $\alpha = 19, \alpha = 54$, and $\alpha = 117.5$ for projection dimensions 1 to 20.

For comparison, consider the singular values of the dataset (in this example and throughout the paper, data is not mean-subtracted), shown in Figure 7. From the singular values for $\alpha = 19$, we infer similar information to that contained in the κ-profile: the data can be projected into $\mathbb{R}^1$ or $\mathbb{R}^2$ without much loss of information. The two vectors that form a basis for the 2-dimensional subspace into which the data is projected for PCA and SAP are shown in Figure 8; while they capture similar information, they provide distinct projections. For each embedding dimension, we consider the geodesic distance between the subspace that defines the PCA projection and that of the SAP projection; see Figure 9. Note that the subspaces are distinct for projections to $\mathbb{R}^m$ with $m < 7$.

For a different choice of the parameter α, we see distinctly

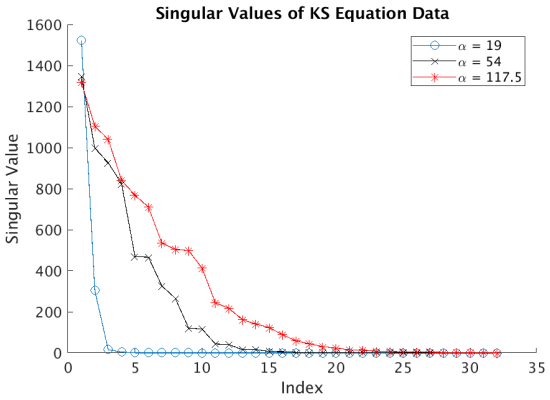

Fig. 7. Singular values for the KS equation with $\alpha = 19, \alpha = 54$, and $\alpha = 117.5$. In the case of $\alpha = 19$, the majority of the energy in the data is captured in the first one to two dimensions. For the other two choices of α, we see singular value decay but a less clear signal regarding the precise dimension of the data.

different behavior. For example, consider $\alpha = 54$, which gives oscillatory and/or chaotic orbits [20]. In this example, we get a good projection of the data starting at dimension 4. For comparison, consider the singular values of the data shown in Figure 7. Here too, we see a change in behavior at dimension 4. However, the dimension estimate inferred from the singular values is not obvious - one could argue that the energy or variance in the data isn't essentially captured until some dimension between 10 and 15. Thus, the added information from the κ-profile is valuable.

As a final example from the KS equation data, we consider $\alpha = 117.5$, which produces chaotic orbits [20]. We show the κ-profile in Figure 6 and the singular values in Figure 7. The more complex behavior of the solution set is reflected in the κ-profile: a good projection arises at dimension 6, and the κ values increase more gradually than in the previous examples with $\alpha = 19$ and $\alpha = 54$. As before, the dimension estimate from the singular values is not obvious, so the κ-profile stands to provide meaningful additional context.

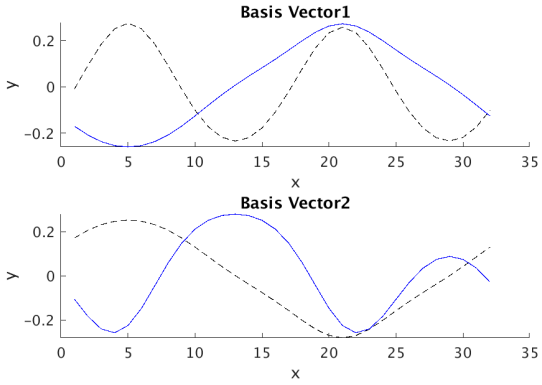

Fig. 8. For $\alpha = 19$: the first two basis vectors for the two dimensionality-reduction techniques. The first and second left singular vectors for PCA are given by the dashed, black lines and the first and second basis vectors from SAP are given by the solid, blue lines.

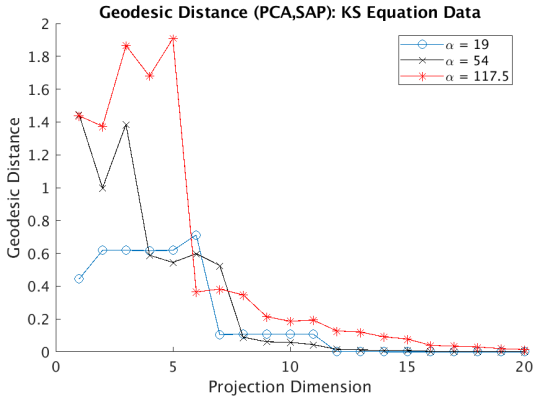

Fig. 9. Geodesic distance between basis of first m left singular vectors and first m SAP basis vectors for projection into $\mathbb{R}^m$ for KS equation with $\alpha = 19, \alpha = 54$, and $\alpha = 117.5$ for $m = 1, \ldots, 20$. The projections are quite distinct for low dimensional projections.

VI. Conclusion

In this paper we present evidence suggesting that the κ-profile is a useful statistic for analyzing and monitoring datasets, particularly those that change over time. While it is in some sense a coarse statistic, it is broadly applicable and carries meaningful information.

We suggest a few directions for future research. (1) Given that dimension is an intrinsic property of a dataset, it would be interesting to understand whether machine learning algorithms could benefit from inclusion of this as a feature. (2) It would be interesting to understand how the dimension and κ-profile are related to other notions of dataset complexity, such as entropy. (3) As mentioned in Section II-E, in the case where the data can be seen as reflecting an underlying dynamical system, the κ-profile should offer an alternative method of estimating the minimum embedding dimension.

References

[1] 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.

[2] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern recognition*, vol. 36, no. 12, pp. 2945–2954, 2003.

[3] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data Mining and Knowledge Discovery*, vol. 29, no. 2, pp. 534–564, 2015.

[4] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2859–2900, 2015.

[5] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Transactions on Computers*, vol. 100, no. 2, pp. 176–183, 1971.

[6] J. A. Costa and A. O. Hero, "Determining intrinsic dimension and entropy of high-dimensional shape spaces," in *Statistics and Analysis of Shapes*. Springer, 2006, pp. 231–252.

[7] D. Hundley and M. Kirby, "Estimation of topological dimension," in *Proceedings of the Third SIAM International Conference on Data Mining*, San Fransico, 2001, pp. 194–202.

[8] D. S. Broomhead, R. Jones, and G. P. King, "Topological dimension and local coordinates from time series data," *J. Phys. A: Math. Gen*, vol. 20, pp. L563–L569, 1987.

[9] H. Kvinge, E. Farnell, M. Kirby, and C. Peterson, "Too many secants: a hierarchical approach to secant-based dimensionality reduction on large data sets," *arXiv preprint arXiv:1808.01686*, 2018, to appear in the 2018 IEEE High Performance Extreme Computing Conference in Waltham, MA.

[10] M. Anderle, D. Hundley, and M. Kirby, "The bilipschitz criterion for mapping design in data analysis," *Intelligent Data Analysis*, vol. 6, no. 1, pp. 85–104, 2002.

[11] D. Broomhead and M. Kirby, "Large dimensionality reduction using secant-based projection methods: The induced dynamics in projected systems," *Nonlinear Dynamics (Special Issue on Reduced Order Modelling)*, vol. 41, no. 1-3, pp. 47–67, 2005.

[12] K. Falconer, *Fractal geometry*, 2nd ed. John Wiley & Sons, Inc., Hoboken, NJ, 2003, Mathematical foundations and applications.

[13] 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.

[14] ——, "The Whitney reduction network: a method for computing autoassociative graphs," *Neural Computation*, vol. 13, pp. 2595–2616, 2001.

[15] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980 (Coventry, 1979/1980)*, ser. Lecture Notes in Math. Springer, Berlin-New York, 1981, vol. 898, pp. 366–381.

[16] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, pp. 3403–3411, Mar 1992.

[17] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber, "On noise reduction methods for chaotic data," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 3, no. 2, pp. 127–141, 1993.

[18] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y. Chuang, M. Iredell, M. Ek, J. Meng, R. Yang, M. P. Mendez, H. van den Dool, Q. Zhang, W. Wang, M. Chen, and E. Becker, "The NCEP climate forecast system version 2," *Journal of Climate*, vol. 27, no. 6, pp. 2185–2208, 2014.

[19] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, 21/10/2013 2013, pp. 411–412.

[20] J. M. Hyman and B. Nicolaenko, "The Kuramoto-Sivashinsky equation: a bridge between PDE's and dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 18, no. 1-3, pp. 113–126, 1986.

[21] Y. Kuramoto and T. Tsuzuki, "On the formation of dissipative structures in reaction-diffusion systemsreductive perturbation approach," *Progress of Theoretical Physics*, vol. 54, no. 3, pp. 687–699, 1975.

[22] D. Michelson and G. Sivashinsky, "Nonlinear analysis of hydrodynamic instability in laminar flamesii. numerical experiments," *Acta Astronautica*, vol. 4, no. 11, pp. 1207 – 1221, 1977.