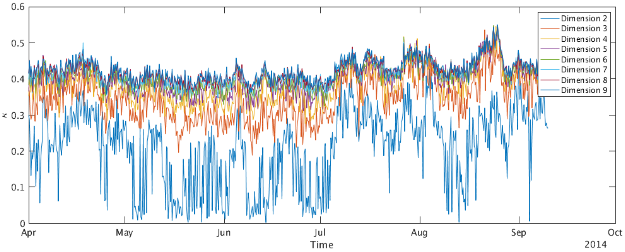
Monitoring the shape of weather, soundscapes, and dynamical systems: a new statistic for dimension-driven data analysis on large datasets.

Henry Kvinge* Elin Farnell Michael Kirby Chris Peterson

Colorado State University, Fort Collins, CO



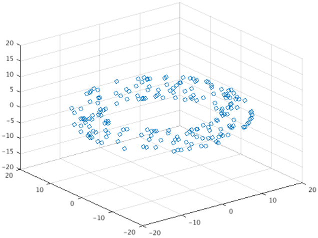
Dimensionality reduction

Dimensionality reduction is a key tool for extracting information from high dimensional data sets.

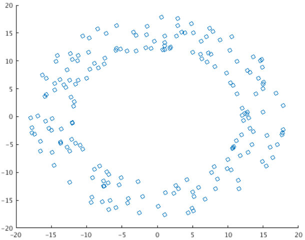
When our data points correspond to elements of $\mathbb{R}^n$ we can think of dimensionality reduction as the process of mapping:

$$\text{points in } \mathbb{R}^n \quad \mapsto \quad \text{points in } \mathbb{R}^m$$

for $m < n$.



$\longrightarrow$



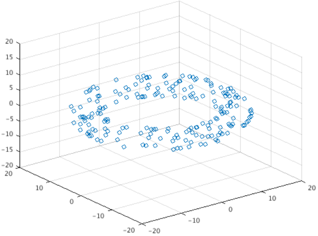
$n = 3$ $\qquad$ $k = 2$

Dimension driven statistics

Some dimensionality reduction algorithms naturally provide statistics that indicate how much structure was lost during reduction.

These statistics give a picture of the intrinsic dimension of the data.

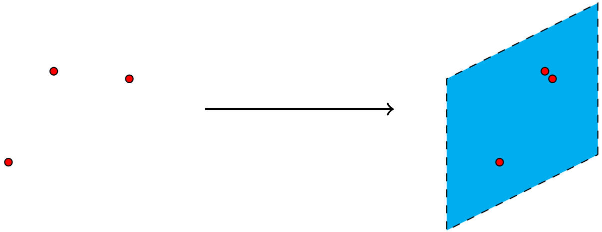
Classic example: The singular values in PCA (or eigenvalues in multidimensional scaling) suggest the number of dimensions required to capture variance of data.



Secant-based approach to dimensionality reduction

Dimensionality reduction can be framed in terms of secant sets.

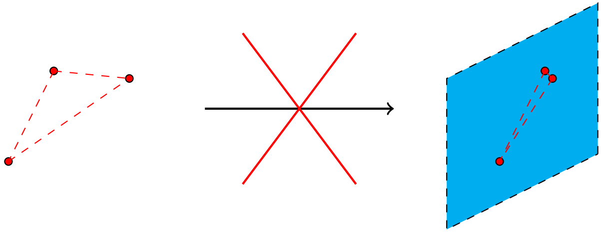
Example: In many cases we do not want to collapse points in $\mathbb{R}^n$ onto each other in $\mathbb{R}^m$.



Secant-based approach to dimensionality reduction

Dimensionality reduction can be framed in terms of secant sets.

Example: In many cases we do not want to collapse points in $\mathbb{R}^n$ onto each other in $\mathbb{R}^m$.



Secant-based dimensionality reduction

What is another way of saying that we want to preserve the distances between points?

For a data set $D \in \mathbb{R}^n$ the normalized *secant set* S is

$$S := \left\{ \frac{x-y}{||x-y||} \mid x, y \in D, \text{ with } x \neq y \right\}.$$

Secant-based dimensionality reduction algorithms work under the principle that we should look for dimension reducing transformations which **preserve the secant set of our data set**.

SAP algorithm outline

In this project we were interested in solving the optimization problem

$$\underset{P \in \mathrm{Proj}(\mathbb{R}^n, \mathbb{R}^m)}{\arg\max} \left(\min_{s \in S} ||P^T s|| \right)$$

where $\mathrm{Proj}(\mathbb{R}^n, \mathbb{R}^m)$ consists of all $n \times m$ matrices whose columns are orthonormal vectors in $\mathbb{R}^n$.

How to solve: Several methods exist, we used an algorithm that we developed called *secant-avoidance projection (SAP) algorithm*.

SAP searches for the projection such that the most shrunken secant is maximized.

The κ-profile

Suppose we have used SAP (or a similar method) to find $\bar{P}$ that gives an approximate solution to

$$\bar{P} \approx \underset{P \in \text{Proj}(\mathbb{R}^n, \mathbb{R}^m)}{\arg\max} \left(\min_{s \in S} ||P^T s|| \right).$$

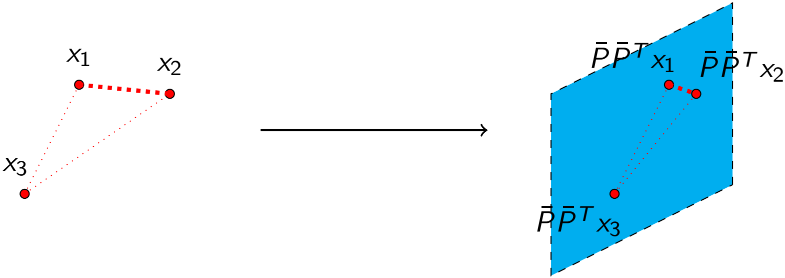
Then we can calculate κ_m,

$$\kappa_m := \min_{s \in S} ||P^T s||.$$

The value $\kappa_m \in [0,1]$ captures the projected length of the secant **least well preserved by $\bar{P}$.**

The κ-profile

Example: If $\bar{P} : \mathbb{R}^3 \to \mathbb{R}^2$ maps



Then

$$\kappa_2 = \frac{||\bar{P}\bar{P}^T x_1 - \bar{P}\bar{P}^T x_2||}{||x_1 - x_2||} \approx .2$$

The κ-profile

We define the κ-**profile** to be the n-tuple

$$(\kappa_1, \kappa_2, \ldots, \kappa_n).$$

The κ-profile gives a measure of how well the data can be projected into dimensions $1, 2, \ldots, n$.

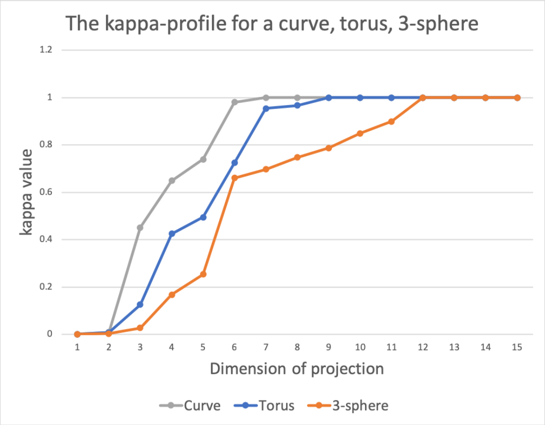
We can give some bounds to the intrinsic dimension of D via $(\kappa_1, \ldots, \kappa_n)$ and Whitney's Embedding Theorem.

The κ-profile

We plot the κ-profile as a curve where

- The x-axis is the dimension of the projection
- The y-axis is the corresponding κ-value

The κ-profile often correlates closely with the dimension of a data set



A comparison to information obtained from PCA

The κ-profile gives information which is distinct from that provided by the singular values in PCA.

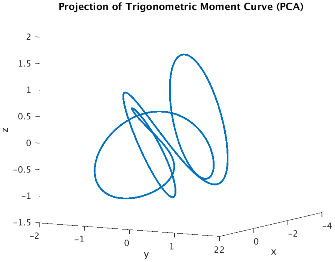
This follows from the fact that the underlying optimization problems are different.

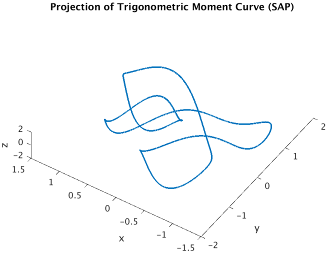
- PCA solves
$$\underset{P \in \mathrm{Proj}(\mathbb{R}^n, \mathbb{R}^m)}{\arg\max} \sum_{s \in S} ||P^T s||$$
- SAP solves
$$\underset{P \in \mathrm{Proj}(\mathbb{R}^n, \mathbb{R}^m)}{\arg\max} \left(\min_{s \in S} ||P^T s|| \right)$$

A comparison to information obtained from PCA

As a consequence:

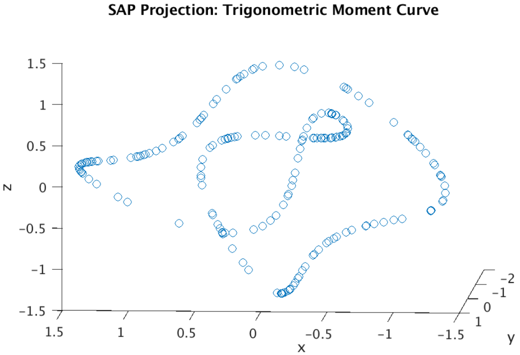
- PCA “tries” to minimize the extent to which all secants are shrunk in projection,
- while SAP “focuses” on making sure no particular secant is shrunk too much.





Trigonometric moment curve $\phi : \mathbb{R} \to \mathbb{R}^{10}$,
$\phi(t) := (\cos(t), \sin(t), \cos(2t), \ldots, \cos(5t), \sin(5t))$.

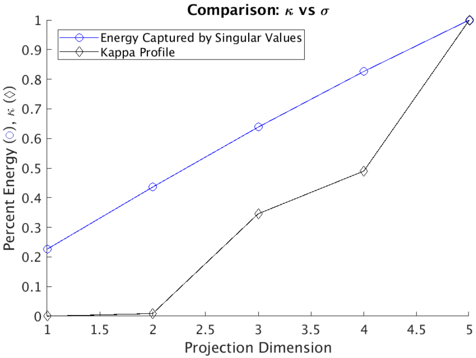
A comparison to information obtained from PCA



Trigonometric moment curve $\phi : \mathbb{R} \to \mathbb{R}^5$,

$$\phi(t) := (\cos(t), \sin(t), \cos(2t), \dots, \sin(4t), \cos(5t)).$$

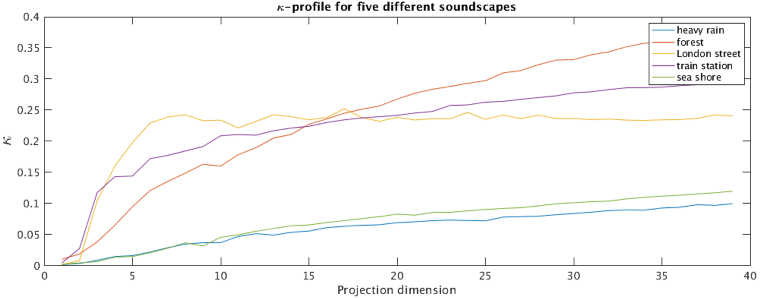
A comparison to information obtained from PCA



- From singular values it appears that data is 5-dimensional (its actually 1-dimensional).
- But the κ-profile suggests that data is somewhere from 1 to 3-dimensional.

Examples of the κ-profile

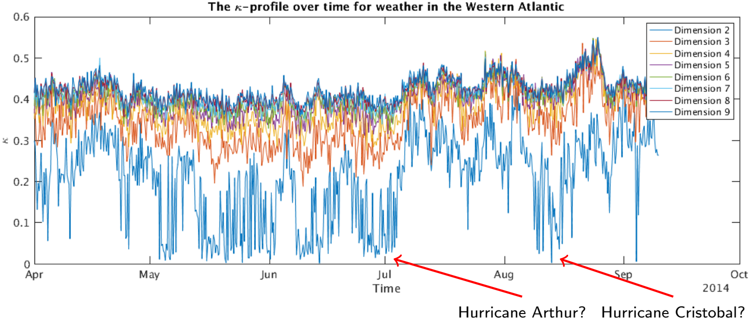
The κ-profile for soundscapes.



The κ-profile suggests that soundscapes with more random, incoherent noise are unsurprisingly higher dimensional.

The κ-profile

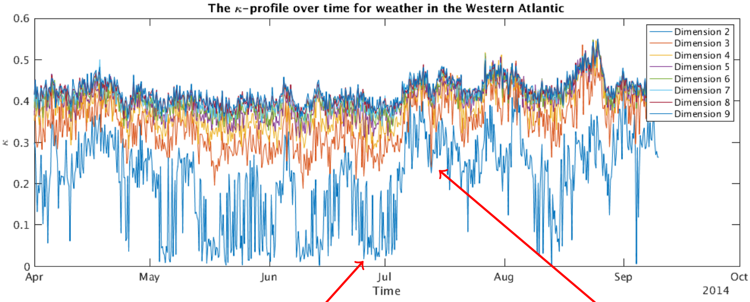
The κ-profile for dimensions $2, 3, \dots, 9$ as a function of time, for a large 2015 weather data set taken from a grid in the Western Atlantic.

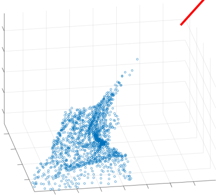


Storm activity seems to roughly correlate with jumps in κ_2.

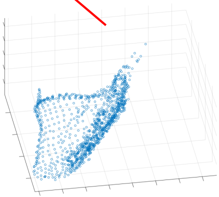
Stormy weather corresponds to the data becoming more 3-dimensional?

The κ-profile





projection into $\mathbb{R}^3$



projection into $\mathbb{R}^3$

Thank you!

(I am currently on the job market.)