CSE 590
Data Science Fundamentals

Similarity and Distances

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern miming	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

Introduction

Please also refer to the statistics foundations lecture

- many distance metrics were discussed there already
- Euclidian, cosine, correlation, …

Topics discussed now

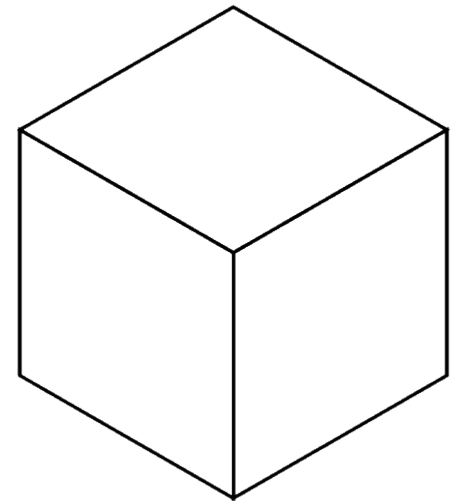
- curse of dimensionality
- structural similarity distance

High-D Space is Tricky

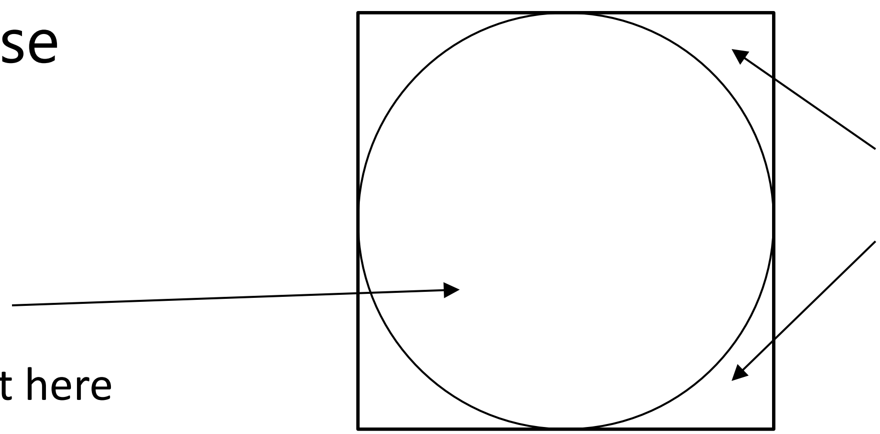
The curse of dimensionality

As $n \to \infty$

- Cube: side length l, diagonal d, volume V
- $V \to \infty$ for $l > 1$
- $V \to 0$ for $l < 1$
- $V = 0$ for $l = 1$
- $d \to \infty$

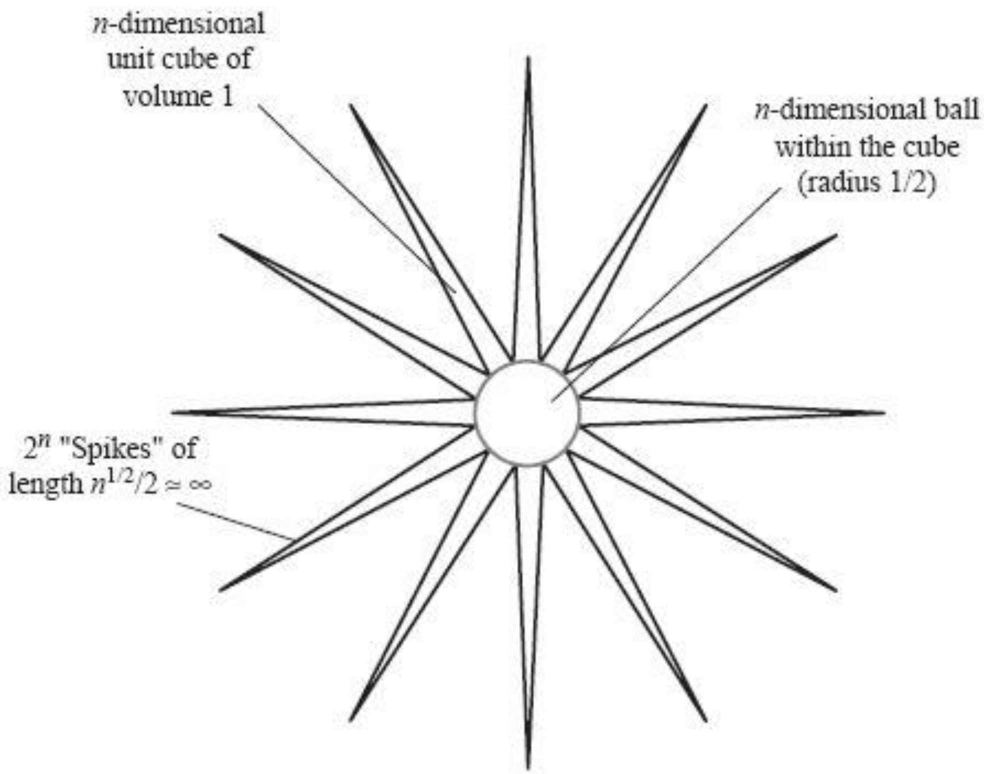


and very sparse



High-D Space is Tricky

Essentially hypercube is like a "hedgehog"



Curse of Dimensionality

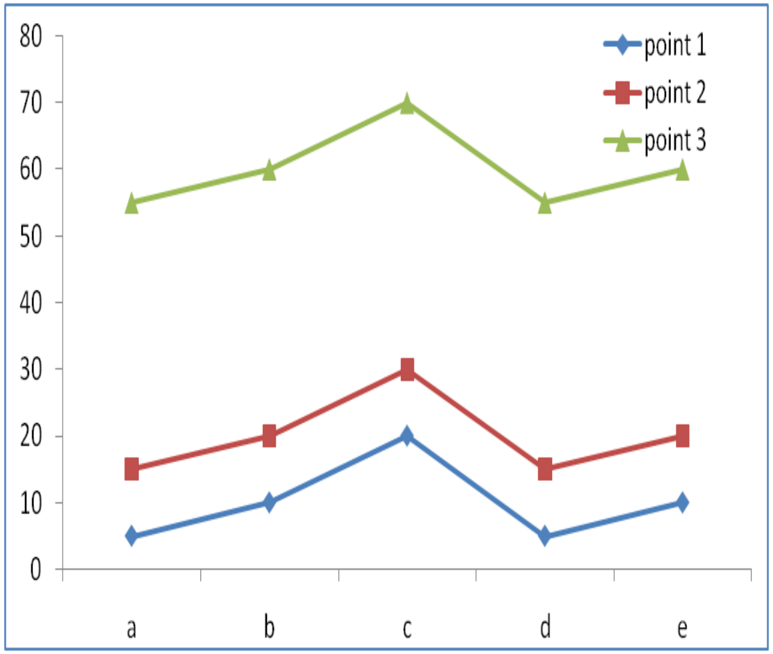
Points are all at about the same distance from one another

- concentration of distances
- fundamental equation (Bellman, ‘61)

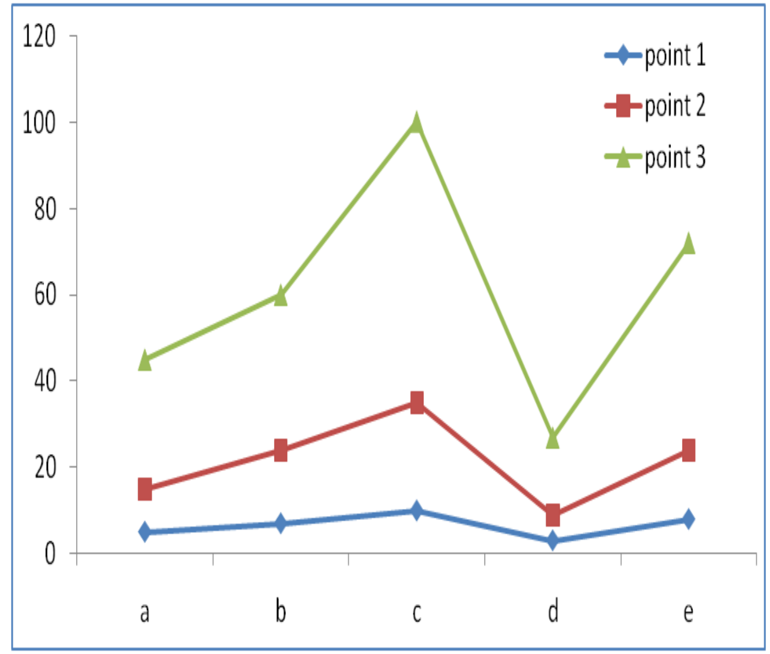
$$\lim_{n \to \infty} \frac{Dist_{\max} - Dist_{\min}}{Dist_{\min}} \to 0$$

- so as *n* increases, it is impossible to distinguish two points by (Euclidian) distance
  - unless these points are in the same cluster of points
  - need to use other distance metrics

Similarity of N-D Points



Same pattern, with offset



Same pattern, with scaling

Distance in High-D Space

MDS optimization function:

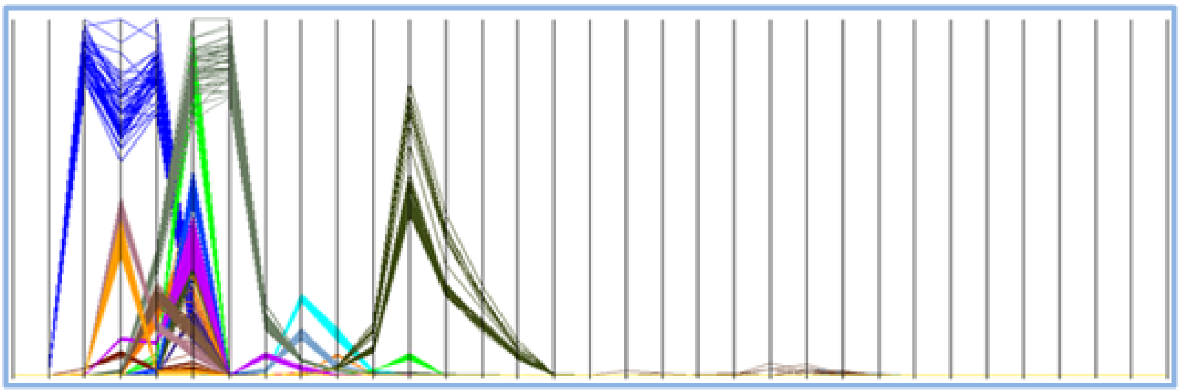
$$\min(y_1, .., y_n) \sum_{i<j} \left(\left\| y_i - y_j \right\| - \delta_{ij} \right)^2$$

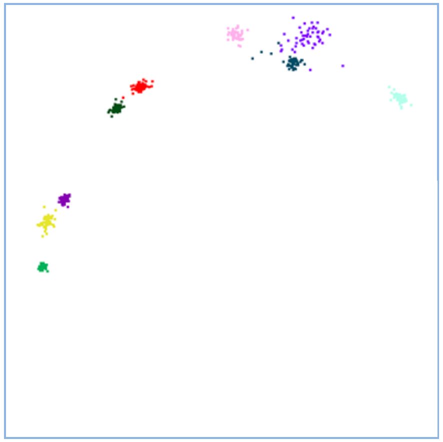
- Euclidian distance
- measures point-pair error
- sums all

distance in ND

distance in 2D

Why is The Euclidian Distance less ideal?





Perceptual (dis)similarity is not gauged by a Euclidian metric

- our cognitive faculties look for *pattern* similarity
- poly lines with similar pattern signature are deemed *closer*
- the equivalent points need to also be *closer* in the MDS plot
- need a new perceptual distance metric that gauges this pattern similarity

Structural Similarity Index (SSIM)

$$\mathrm{SSIM}(x,y) = \left[\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\right]^{\alpha} \cdot \left[\frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}\right]^{\beta} \cdot \left[\frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}\right]^{\gamma}$$

luminance   contrast   structure

x

y

Structural Similarity Index (SSIM)

$$\mathrm{SSIM}(x,y)=\left[\frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1}\right]^{\alpha}\cdot\left[\frac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2}\right]^{\beta}\cdot\left[\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}\right]^{\gamma}$$

luminance　　contrast　　structure

frequently pooled over a 11×11 sliding window

$$SSIM_{pooled}=\frac{1}{n_w}\sum_{i=1}^{n_w}SSIM(x_i,y_i)$$

Example from Image Processing

original

contrast stretched

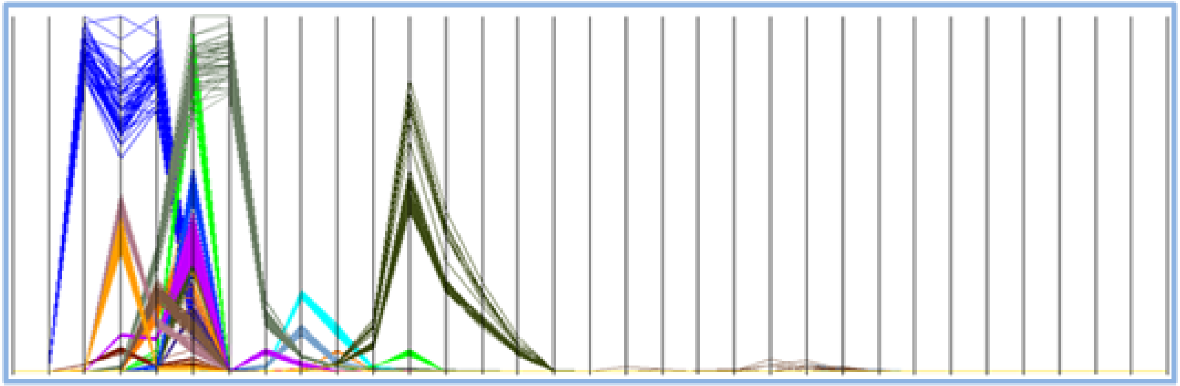
blurred

salt + pepper

Both images have the same MSE but different SSIM

- SSIM = 0.91, 0.71, 0.77
- Euclidian distance expression is similar to that of MSE
- hence it can be expected that SSIM will do better when ported

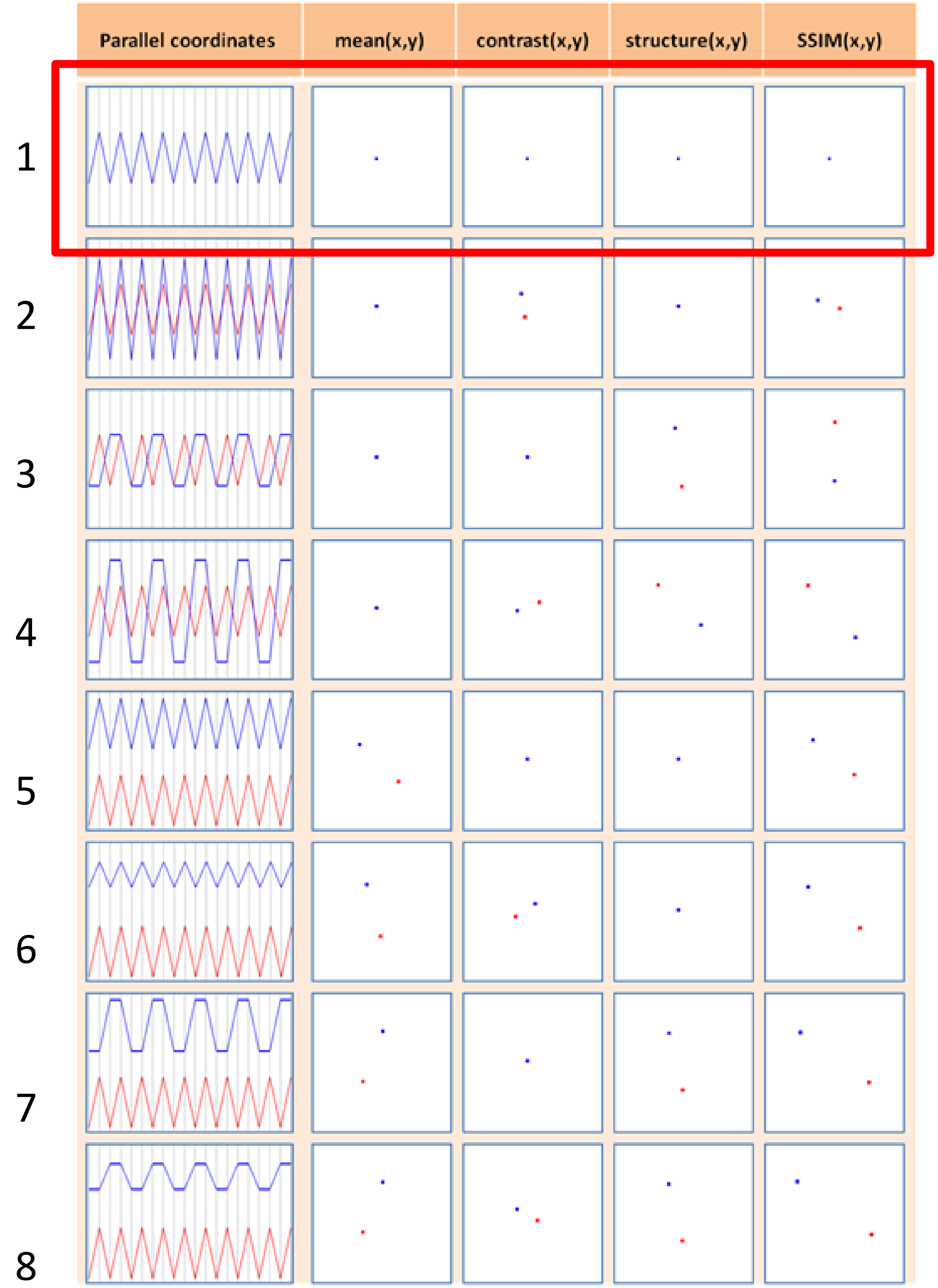
Analogy of SSIM to ND Distance



Just like images, polylines have

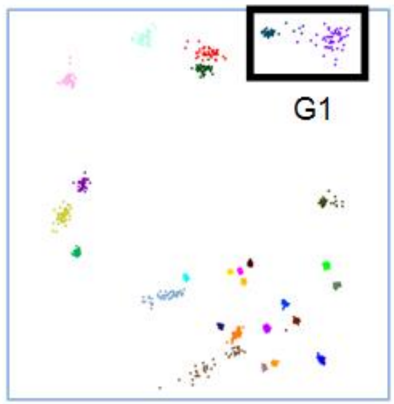
- luminance → mean
- contrast
- structure ← evaluates the structural similarity after the differences in mean and contrast have been accounted for

Cases

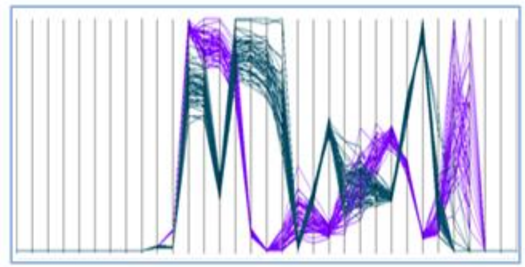

Parallel coordinates
mean(x,y)
contrast(x,y)
structure(x,y)
SSIM(x,y)
1
2
3
4
5
6
7
8

Effect of Windowing

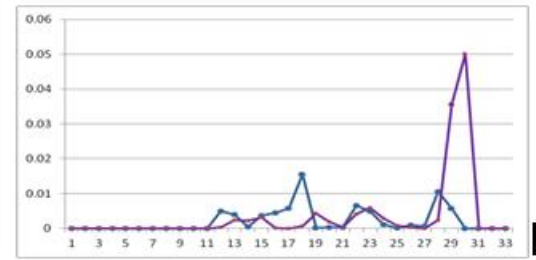
not windowed



windowed



clusters in G1



local variance

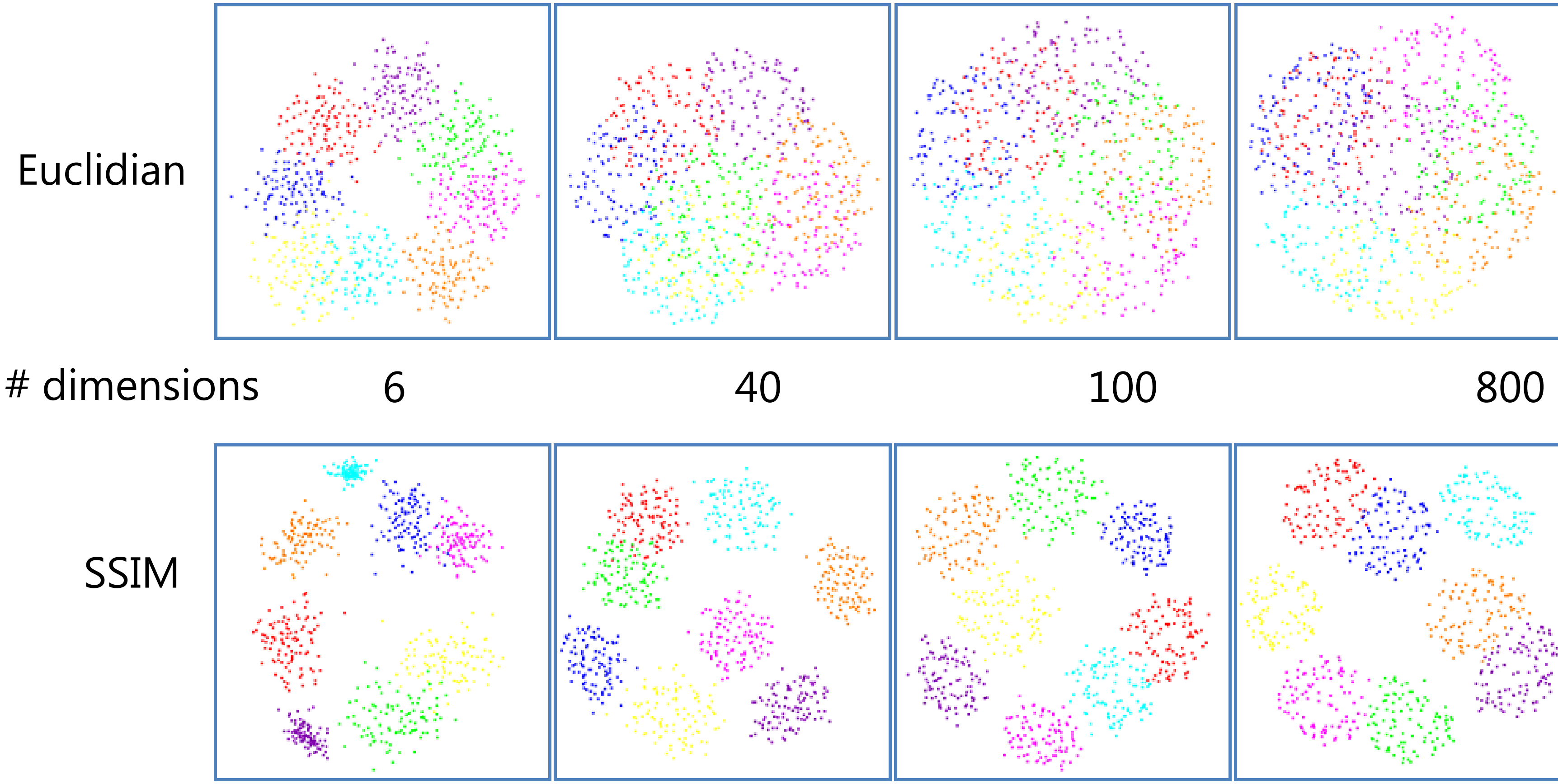
Procedure

- order dimensions such that sum of pairwise correlations is maximized
- use 11-point window

Observations

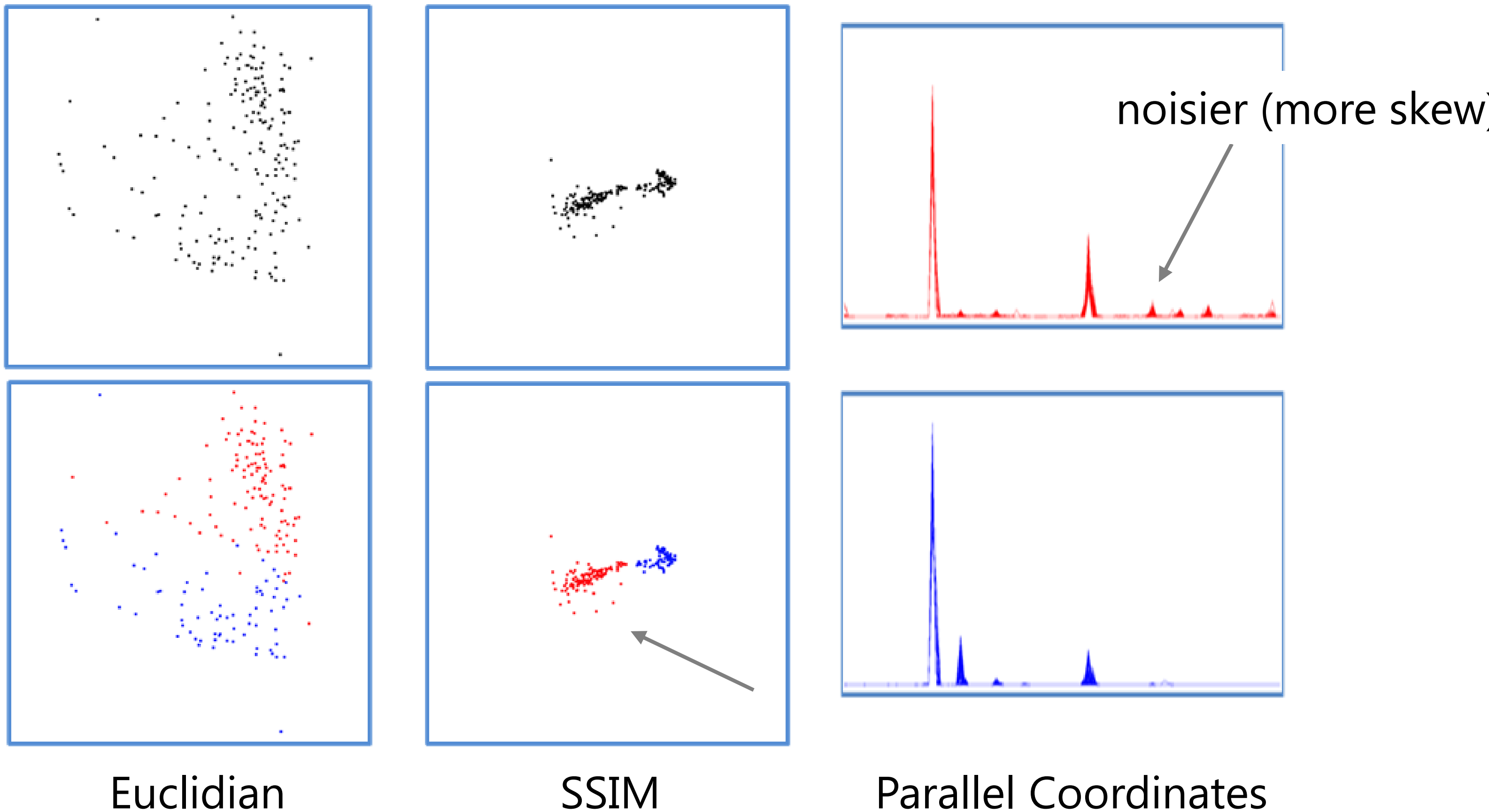
- purple cluster has higher local variance than blue
- with windowing it has an equivalent spread in the MDS plot
- without windowing this effect is averaged out and likewise in MDS

Achieves Better Cluster Separability



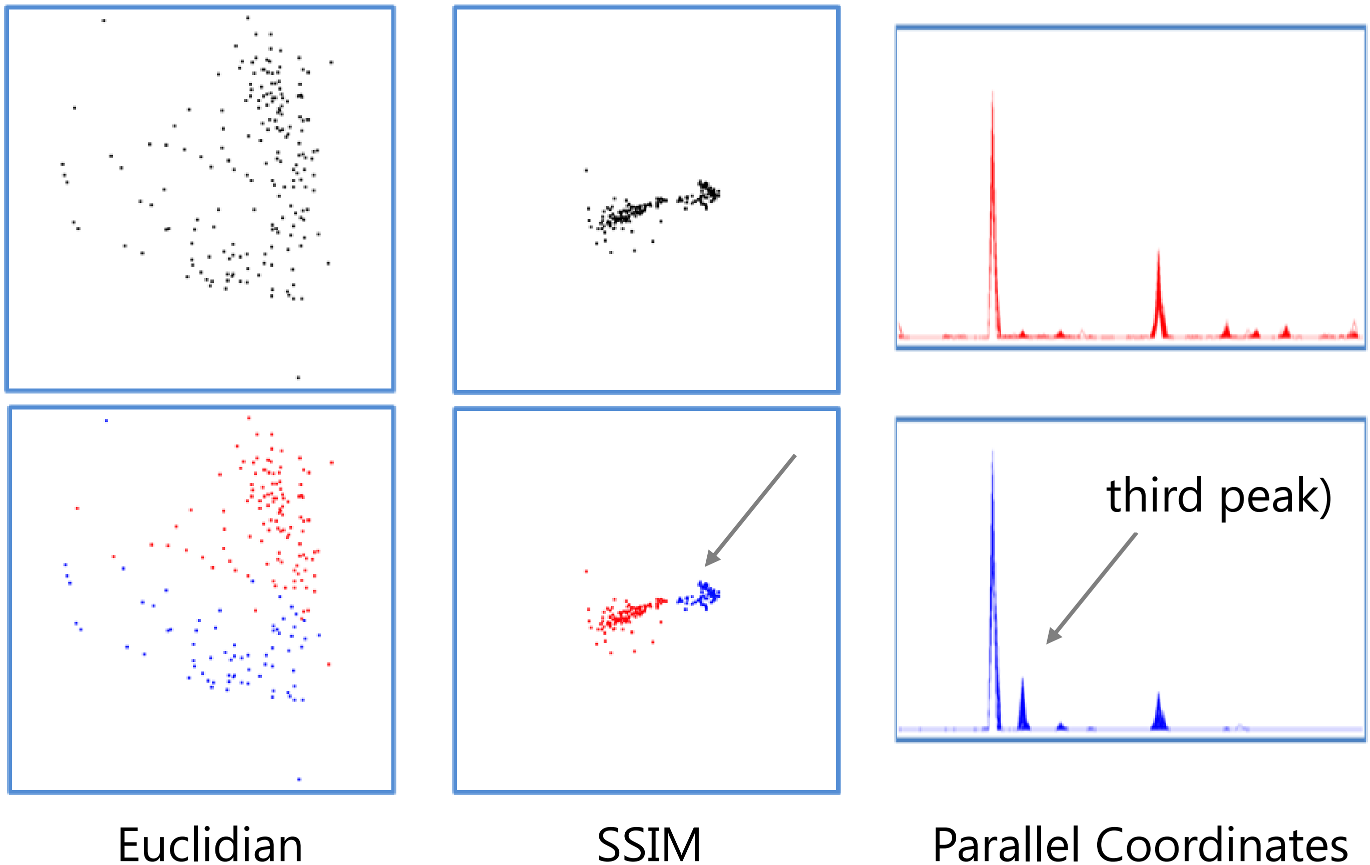
8 clusters, 800 data points

Mass Spectra Data



Mass Spectra of Aerosol Particles, 450 D, 2,000 data points

Mass Spectra Data



Mass Spectra of Aerosol Particles, 450 D, 2,000 data points