Statistics 185 - Introduction to Dimension Reduction

Course Introduction and Overview

Instructor: Alex Young

Tuesday September 3

Examples of High-Dimensional data

Ubiquitous in the modern data-driven economy

- Static Images
  - Handwritten digits, signature
  - Medical images
- Movies/Audio/Signals
- Genetic data
  - Personalized medicine
  - Genotype/phenotype connections
- Textual data
  - Spam detection
- Patient and consumer data
  - Who is likely to buy (or die)?

What is high-dimensional?

Depends on the number of samples?

Consider the d-dimensional box $[0,1]^d$ divided into bins by cutting 10 slabs in each direction.

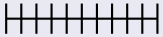


Figure: 1D $\rightarrow$ 10 bins

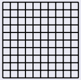


Figure: 2D $\rightarrow$ 100 bins

What is high-dimensional?

Depends on the number of samples?

Consider the d-dimensional box $[0,1]^d$ divided into bins by cutting 10 slabs in each direction.

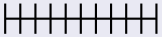


Figure: 1D $\rightarrow$ 10 bins

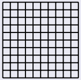


Figure: 2D $\rightarrow$ 100 bins

Exponential in the number of samples

- If we want to make one observation in each bin, we need 10^d samples.
- A 19×19 pixel image translates to **one** 19^2-dimensional datum.
  - Need 10^{381} samples
  - Estimated 10^{82} atoms in the universe!
- Assumes uniform distribution but the idea generalizes.

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model $+$ High-dimensional data $=$ Overfitting (often)

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model $+$ High-dimensional data $=$ Overfitting (often)
- Model $+$ Reduced dimensions $=$ Generalizability (hopefully)

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model + High-dimensional data = Overfitting (often)
- Model + Reduced dimensions = Generalizability (hopefully)

Reasons for Dimension Reduction

- Feature extraction or selection

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model + High-dimensional data = Overfitting (often)
- Model + Reduced dimensions = Generalizability (hopefully)

Reasons for Dimension Reduction

- Feature extraction or selection
- Learn geometric features from data that cannot be visualized

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model + High-dimensional data = Overfitting (often)
- Model + Reduced dimensions = Generalizability (hopefully)

Reasons for Dimension Reduction

- Feature extraction or selection
- Learn geometric features from data that cannot be visualized
- Data compression

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model + High-dimensional data = Overfitting (often)
- Model + Reduced dimensions = Generalizability (hopefully)

Reasons for Dimension Reduction

- Feature extraction or selection
- Learn geometric features from data that cannot be visualized
- Data compression
- Reduced computation cost

One name, many goals

Curse of Dimensionality (Bellman)

- High dimension $\rightarrow$ sparse data
  - Need exponentially increasing (in dimension samples), otherwise ...
  - Model + High-dimensional data = Overfitting (often)
- Model + Reduced dimensions = Generalizability (hopefully)

Reasons for Dimension Reduction

- Feature extraction or selection
- Learn geometric features from data that cannot be visualized
- Data compression
- Reduced computation cost
- Removal of redundant information

Example - Nonnegative Matrix Factorization

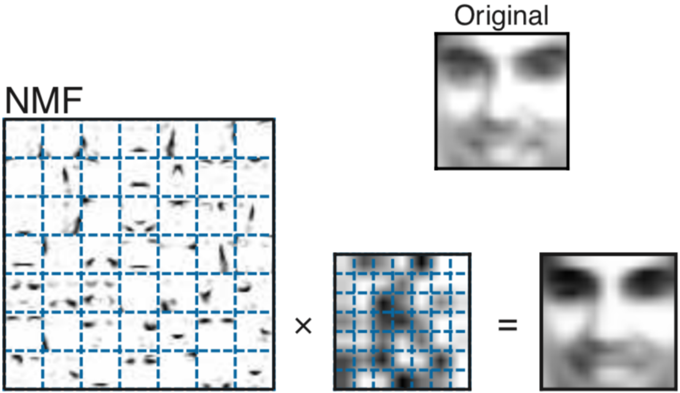


Figure: Image taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, originally from Lee and Seung (1999)

Example - Vector Quantization/k-means

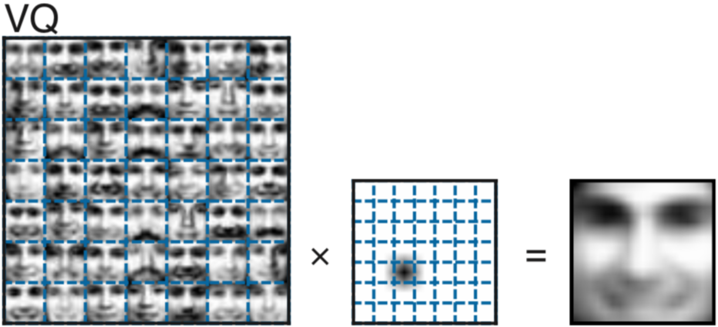


Figure: Image taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, originally from Lee and Seung (1999)

Example - Principal Component Analysis

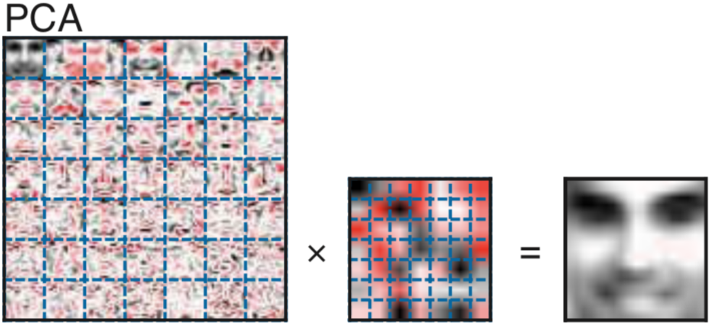
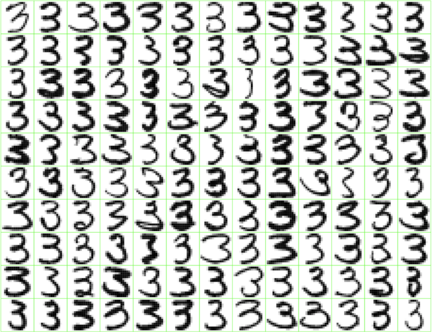


Figure: Image taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, originally from Lee and Seung (1999)

Example - Principal Component Analysis



$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \tag{14.55}$$

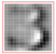
$=$  $+ \lambda_1 \cdot$  $+ \lambda_2 \cdot$  .

Figure: Images and equation taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman

Hierarchical Clustering

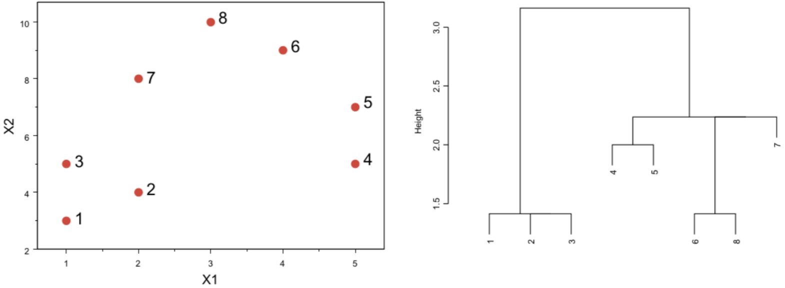


Figure: Image taken from *Modern Multivariate Statistical Techniques* by Izenman

Hierarchical Clustering

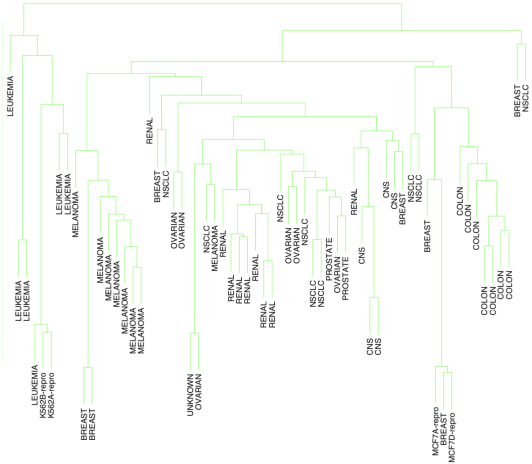


Figure: Image of Hierarchical Clustering for tumor microarray data taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman

Spectral Clustering

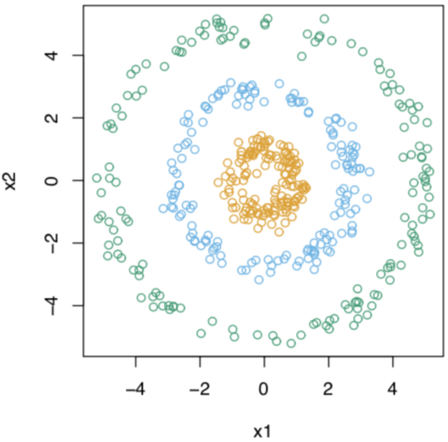


Figure: Image taken from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman

Taxonomy of Dimension Reduction (by Convexity)

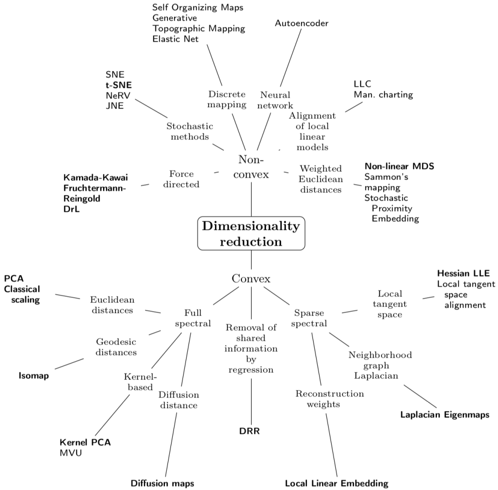


Image from *dimRed and coRanking—Unifying Dimensionality Reduction in R* by Guido Kraemer, Markus Reichstein, and Miguel D. Mahecha

Taxonomy of Dimension Reduction (by Linearity)

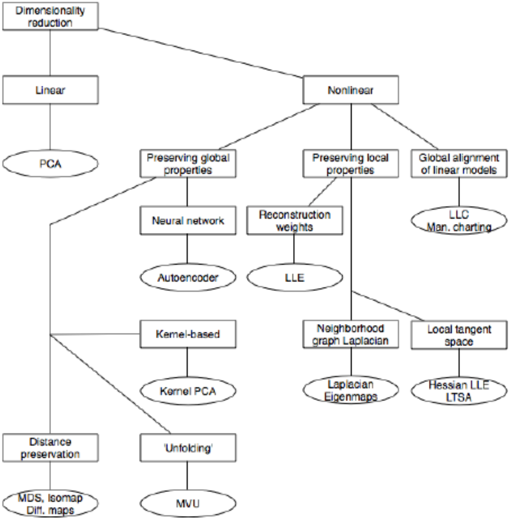


Image from *Dimensionality Reduction: A Comparative Review* by Laurens van der Maaten, Eric Postmam, and H. Jaap Van Den Herik

Course Objectives

Recent trends in Dimension Reduction Research

- Google Scholar results for 'Dimension Reduction' search
  - 2017 - 88,000 articles
  - 2018 - 72,700 articles
  - 2019 - 40,000 articles
- Expanded into more specialized subfields, i.e. manifold learning

Course Objectives

Recent trends in Dimension Reduction Research

- Google Scholar results for 'Dimension Reduction' search
  - 2017 - 88,000 articles
  - 2018 - 72,700 articles
  - 2019 - 40,000 articles
- Expanded into more specialized subfields, i.e. manifold learning

An Enormous and Active Field

- Many missing topics from previous slides
  - Specialized (variants of existing) algorithms being developed for specific needs
  - Primary focus of listed techniques is on continuous data
  - Graphical and/or discrete data receiving increased interest
- Implementation for massive datasets drives parallel research in computation
  - Messy landscape of competing codes, packages, and applications

Course Objectives

An introduction to the field

- Cannot give an exhaustive summary
  - Field is still evolving $\rightarrow$ ever-expanding list of topics
  - Best computational tool(s) is a topic of much debate
    - Computation alone can miss important insight
- High-dimensional geometry can be ~~challenging~~ ~~counterintuitive~~ weird
  - Cannot directly visualize to detect patterns in data
  - Different techniques $\rightarrow$ different behaviors/goals $\rightarrow$ different insights
    - Manifold Learning, Clustering/Grouping, Classification, etc

Course Objectives

An introduction to the field

- Cannot give an exhaustive summary
  - Field is still evolving $\rightarrow$ ever-expanding list of topics
  - Best computational tool(s) is a topic of much debate
    - Computation alone can miss important insight
- High-dimensional geometry can be ~~challenging~~ ~~counterintuitive~~ weird
  - Cannot directly visualize to detect patterns in data
  - Different techniques $\rightarrow$ different behaviors/goals $\rightarrow$ different insights
    - Manifold Learning, Clustering/Grouping, Classification, etc

Goal: Learn to critique

- Demonstrate the practice of using careful analysis to explore (classic) techniques with the aim of learning:
  - What behavior(s) they aim to capture
  - Their strengths, weaknesses, limitations, and challenges
  - When they will work and when they are likely to fail

Commonalities across Methods

Data $\rightarrow$ Vectors

- Observe d variables from N subjects, i.e. N vectors in $\mathbb{R}^d$
- Mean $\rightarrow$ vector; Covariance $\rightarrow$ Matrix
  - *Multivariate probability/statistics*

Commonalities across Methods

Data $\rightarrow$ Vectors

- Observe d variables from N subjects, i.e. N vectors in $\mathbb{R}^d$
- Mean $\rightarrow$ vector; Covariance $\rightarrow$ Matrix
  - *Multivariate probability/statistics*

What does the cloud of data points look like?

- If we could view the points, are they concentrated on/near some shape of lower dimension?
  - e.g. Observations in $\mathbb{R}^3$ near a smooth curve/line
  - e.g. Observations in $\mathbb{R}^{381}$ near a 71-dimensional affine subspace?
- If we cannot visualize, how can we make sense of the geometric features?
  - *Linear algebra*

Commonalities across Methods

Central Assumption

- There is some lower dimensional structure inherent to the data!

Commonalities across Methods

Central Assumption

- There is some lower dimensional structure inherent to the data!

Goal: Minimal 'information' loss

- Replace observation in $\mathbb{R}^d$ with compressed representation in $\mathbb{R}^k$ where $k \ll d$
- Different notions of 'information', but maximization/minimization is common aim
  - *Optimization, Multivariable Calculus*

$$w^* = \arg\max_{\|w\|=1} \|Xw\|$$

$$(W^*, H^*) = \arg\max \sum \sum [x_{ij} \log(WH)_{ij} - (WH)_{ij}]$$

Course Overview

Tentative Schedule

The proposed schedule could changed depending upon a myriad of factors.

Week	Date	Topic	Date	Topic
1	9/3	Course Overview (HW#1 out)	9/5	Review: Multivar. Stats.
2	9/10	PCA I (HW#1 due)	9/12	PCA II
3	9/17	PCA III (HW#2 out)	9/19	CCA I
4	9/24	CCA II (HW#2 due)	9/26	NMF I (HW#3 out)
5	10/1	NMF II	10/3	NMF III (HW#3 due)
6	10/8	MDS I (HW#4 out)	10/10	MDS II
7	10/15	MDS III (HW#4 due)	10/17	Midterm (MT coding out)
8	10/22	ISOMAP	10/24	LLE (MT coding due)
9	10/29	Probability Review(HW#5 out)	10/31	Johnson-Lindenstrauss
10	11/5	Compressed Sensing (HW#5 due)	11/7	Hierarchical Clustering (HW#6 out)
11	11/12	Center-based clustering	11/14	Spectral Clustering (HW#6 due)
12	11/19	Spectral Clustering (HW#7 out)	11/21	Kernel Methods
13	11/26	Google PageRank (HW#7 due)	11/28	Thanksgiving: No class
14	12/3	Variational Autoencoder	12/4	Fall Reading Period: No class
Final Exam Period	12/	Term Papers Due		

Course Layout - Assignments/Grading

Homework (40%)

- 7 total assignments, lowest score dropped
  - Homework #1 available on Canvas (due next Tuesday)
- Each assignment will include some (guided) coding portions in R
  - Experience in R a plus but not required
  - Interface through RStudio
- RMarkdown files, complete and submit through Canvas as pdf/html
  - Write answers in Markdown (similar to LaTeX)
  - Good preparation for writing/formatting term paper (more on this shortly)
- Encouraged to work together, but must submit solutions written in your own words

Course Layout - Assignments/Grading

Midterm (20%)

- Written portion (10%)
  - In-class on Thursday October 17th
  - To cover PCA through MDS (see calendar)
  - Similar content to homework sets
  - Detailed discussion on layout/content of the exam on Tuesday October 15th
- Coding portion (10%)
  - Released Thursday October 17th; Due Thursday October 24th
  - RMarkdown file similar to homework
  - No collaboration allowed but Office Hours encouraged!

Course Layout - Assignments/Grading

Term Paper (40%)

- Review the formulation, goals, and limitations of a dimension reduction technique with (at least) one real data application
  - Optional: extension to different setting, rigorous proof of mathematical/statistical properties
- Select topic by November 10th (via email)
  - Recommended topics in syllabus
    - Some examples: Diffusion Maps, MVU, ICA, t-SNE and Spherelets
  - Students welcome to propose their own topics too
  - *No more than two students per topic preferably*
- RMarkdown and LaTeX templates will be available on Canvas
- All mathematical expressions must be appropriately typeset
  - "f_x(x) = int f(x,y) dy = d/dx F(x)" not acceptable
- Due Saturday December 14th at 2:00 PM (final exam time)
  - Office hours will be added during reading period to discuss papers
  - Submit via canvas as pdf/html

Final thought - textbooks

Course notes will be sufficient but ...

Freely available e-references

- *Foundations of Data Science* by Blum, Hopcroft, Kannan
- *The Elements of Statistical Learning* by Hastie, Tibrishani, Friedman
- *Modern Multivariate Statistical Analysis* by Izenman
- Links to each reference in syllabus on Canvas