Dimensionality Reduction
PCA

Aarti Singh

Machine Learning 10-315
Nov 16, 2020

Slides Courtesy: Tom Mitchell, Eric Xing, Lawrence Saul

High-Dimensional data

- High-Dimensions = Lot of Features

Document classification

Features per document =

thousands of words/unigrams

millions of bigrams, contextual

information

Surveys - Netflix

480189 users x 17770 movies

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

High-Dimensional data

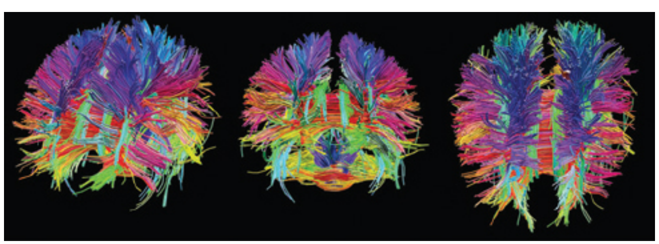
- High-Dimensions = Lot of Features

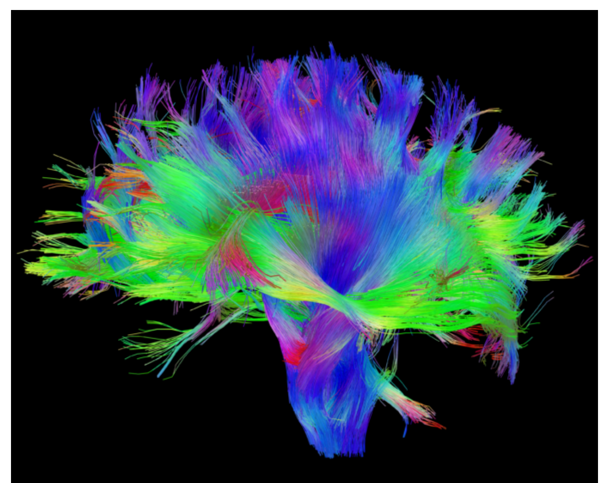
High resolution images

millions of pixels

Diffusion scans of Brain

300,000 brain fibers



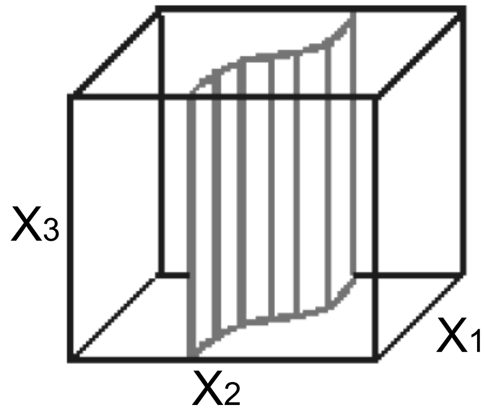


Curse of Dimensionality

- Why are more features bad?
  - Redundant features (not all words are useful to classify a document) more noise added than signal
  - Hard to interpret and visualize
  - Hard to store and process data (computationally challenging)
  - Complexity of decision rule tends to grow with # features. Hard to learn complex rules as it needs more data (statistically challenging)

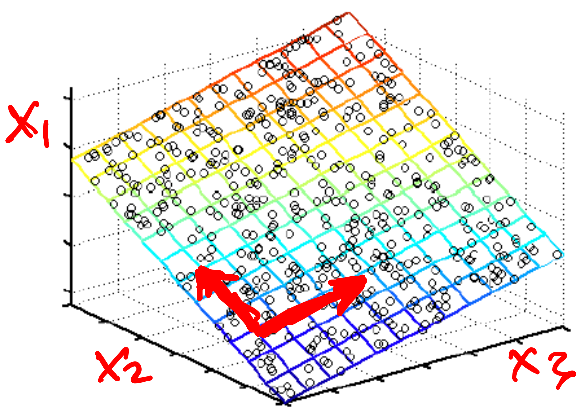
Dimensionality Reduction

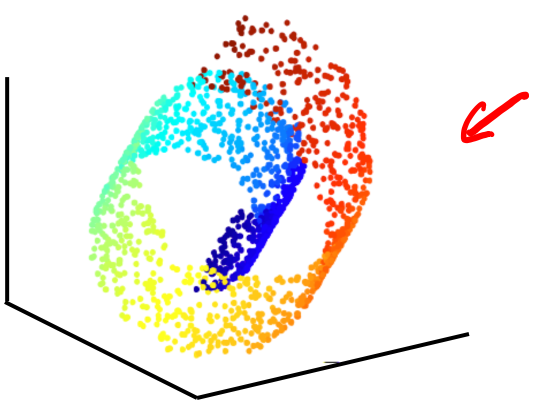
- Feature Selection – Only a few features are relevant to the learning task



X3 - Irrelevant

- Latent features – Some linear/nonlinear combination of features provides a more efficient representation than observed features



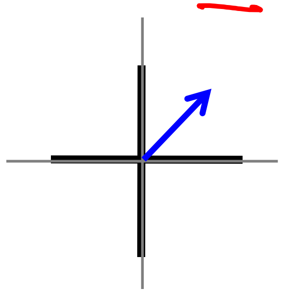


Feature Selection

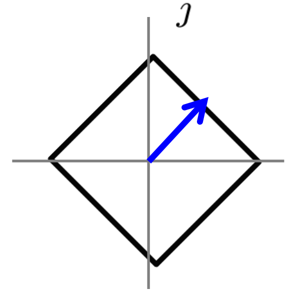
- One Approach: **Regularization (MAP)**
  Integrate feature selection into learning objective by penalizing number of features with non-zero weights

$$\widehat{W} = \arg\min_{W} \underbrace{\sum_{i=1}^{n} - \log P(Y_i|X_i; W)}_{\text{-ve log likelihood}} + \underbrace{\lambda \|W\|_{0 \nearrow 1}}_{\text{penalty}}$$

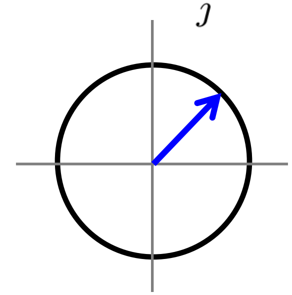
$$\|W\|_0 = \#\{W_j > 0\} \qquad \|W\|_1 = \sum_j |W_j| \qquad \|W\|_2 = \sum_j W_j^2$$



Minimizes # features chosen



Convex compromise



Small weights of features chosen

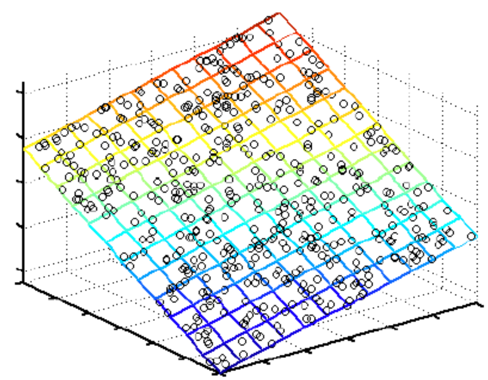
Latent Features

Combinations of observed features provide more efficient representation, and capture underlying relations that govern the data

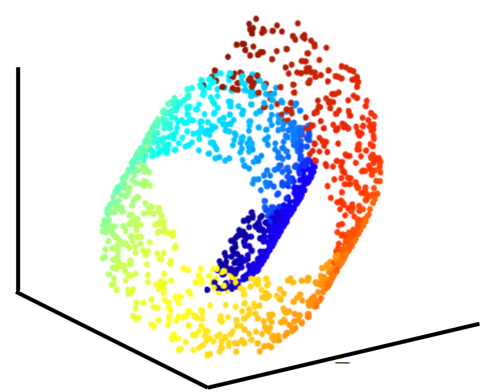
E.g. Ego, personality and intelligence are hidden attributes that characterize human behavior instead of survey questions

Topics (sports, science, news, etc.) instead of documents

Often may not have physical meaning

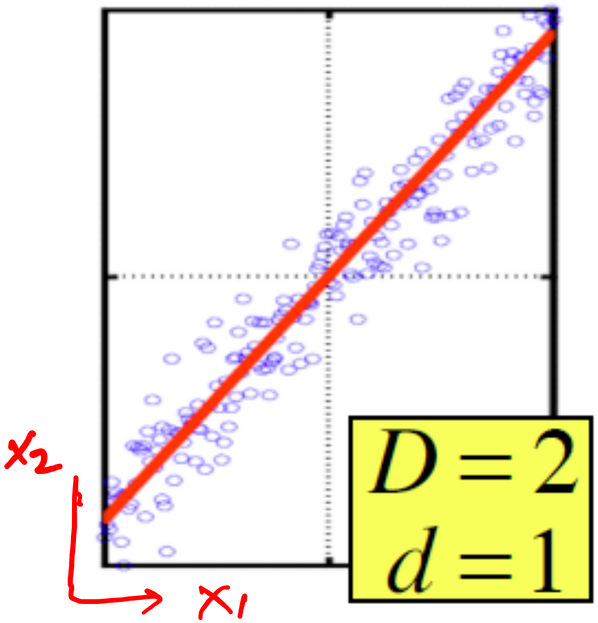


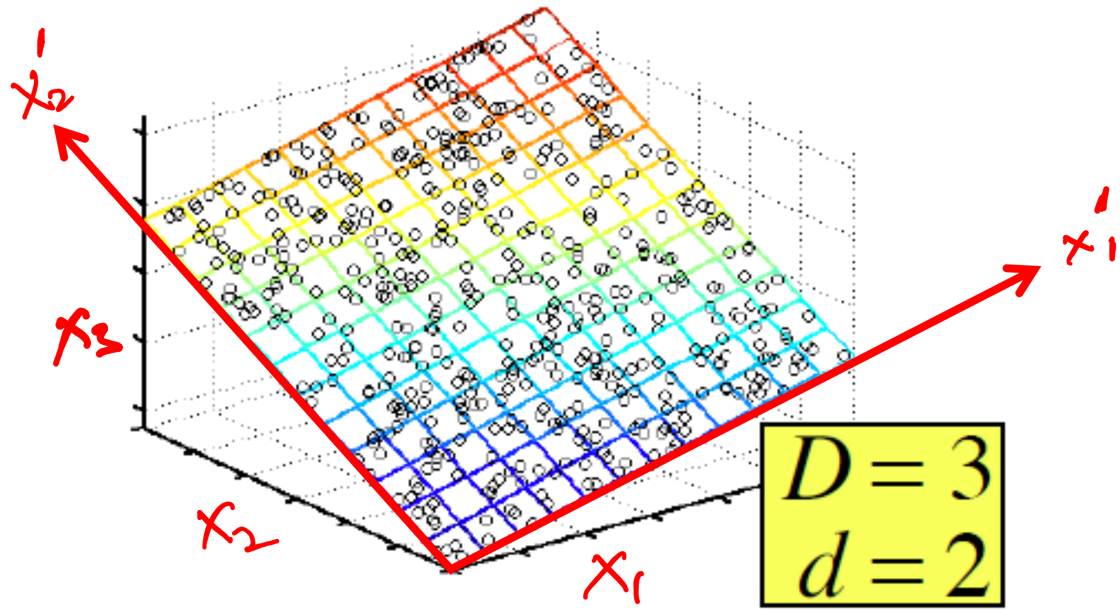
- Linear
  - **Principal Component Analysis (PCA)**
  - Factor Analysis
  - Independent Component Analysis (ICA)



- Nonlinear
  - **Kernel PCA** – HW4!
  - Laplacian Eigenmaps
  - ISOMAP, Local Linear Embedding (LLE)

Principal Component Analysis (PCA)





When data lies on or near a low d-dimensional linear subspace, axes of this subspace are an effective representation of the data

Identifying the axes is known as Principal Components Analysis, and can be obtained by Eigen or Singular value decomposition

Data for PCA

$d < D$

Data X = $[x_1, x_2, \ldots, x_n]$ where each data point x_i is D-dimensional vector

X is D x n matrix

Assume data are centered i.e. sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$ $\hat{\mu}$

What if data is not centered?

Subtract off sample mean from each data point $x_i - \frac{1}{n}\sum_{j=1}^{n} x_j \to x_i$

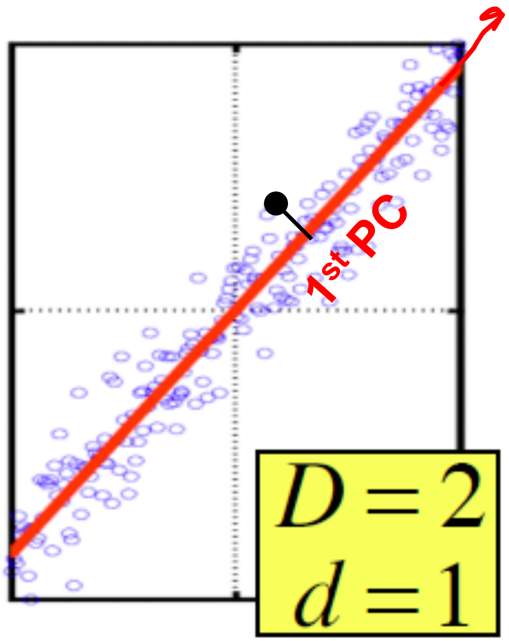
Since data matrix is centered, sample covariance matrix can be written as

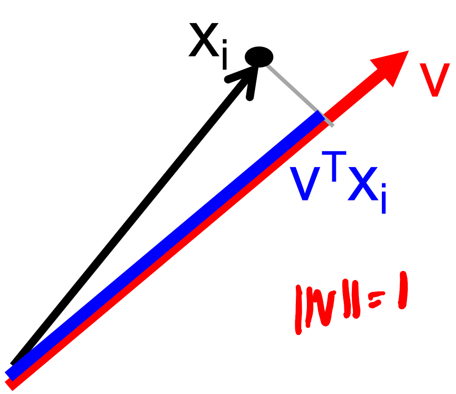
$$\underbrace{S = \frac{1}{n} X X^\top}$$

$S_{ij} = \frac{1}{n} [x_1(i)\ x_2(i) \ldots x_n(i)] \begin{bmatrix} x_1(j) \\ x_2(j) \\ \vdots \\ x_n(j) \end{bmatrix}$

$= \frac{1}{n}\sum_{k=1}^{n} x_k(i) x_k(j)$

Principal Component Analysis (PCA)





Principal Components (PC) are orthogonal directions that capture most of the variance in the data

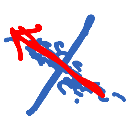
1st PC – direction of greatest variability in data

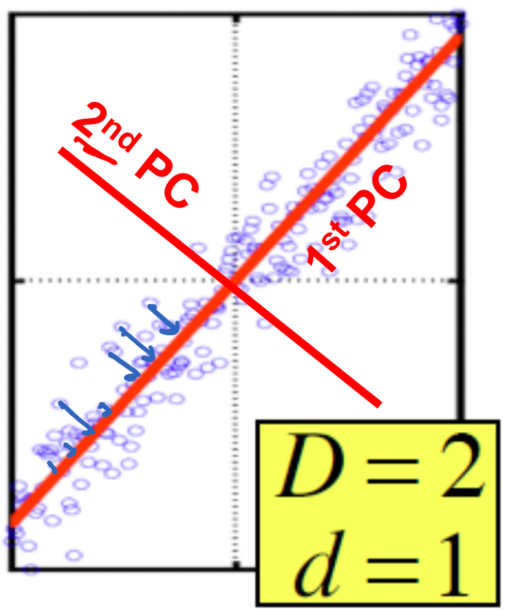
Projection of data points along 1st PC discriminate the data most along any one direction

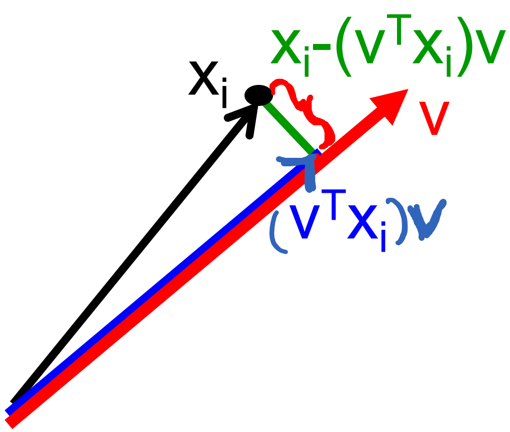
Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

Principal Component Analysis (PCA)







Principal Components (PC) are orthogonal unit norm directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

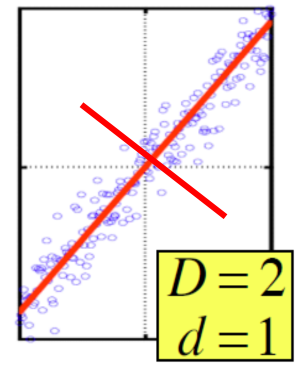
2nd PC – Next orthogonal (uncorrelated) direction of greatest variability

(remove all variability in first direction, then find next direction of greatest variability)

And so on …

Principal Component Analysis (PCA)

Let $v_1, v_2, \ldots, v_d$ denote the principal components $d \le D$

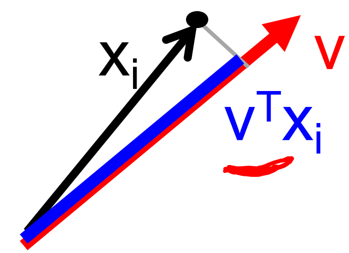


Orthogonal and unit norm $\quad v_i^T v_j = 0 \quad i \neq j$

$\|v_i\|^2 = v_i^T v_i = 1$

Find vector that maximizes sample variance of projection

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\underbrace{\mathbf{X}\mathbf{X}^T}_{D \times D}\mathbf{v}$$



$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

Lagrangian: $\max_{\mathbf{v}} \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} - \lambda\mathbf{v}^T\mathbf{v}$

Wrap constraints into the objective function

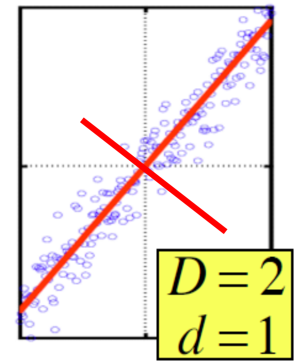
$2XX^Tv - 2\lambda v = 0$

$\lambda(v^Tv - 1) = \lambda v^T v - \lambda$

$$\partial/\partial\mathbf{v} = 0 \quad \Rightarrow (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0 \quad \Rightarrow \boxed{(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v}}$$

var of proj points $= v^T X X^T v = v^T(\lambda v) = \lambda v^T v = \lambda$

Principal Component Analysis (PCA)



$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v}$$

Therefore, v is the eigenvector of sample covariance matrix XX^T $\frac{1}{n}$

Sample variance of projection $= \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

Eigenvalues $\lambda_1 > \lambda_2 > \lambda_3 > \ldots$

The 1st Principal component v_1 is the eigenvector of the sample covariance matrix XX^T associated with the largest eigenvalue λ_1

The 2nd Principal component v_2 is the eigenvector of the sample covariance matrix XX^T associated with the second largest eigenvalue λ_2

And so on …

XX^T

$\underbrace{D\times n \;\; n\times D}_{D\times D}$

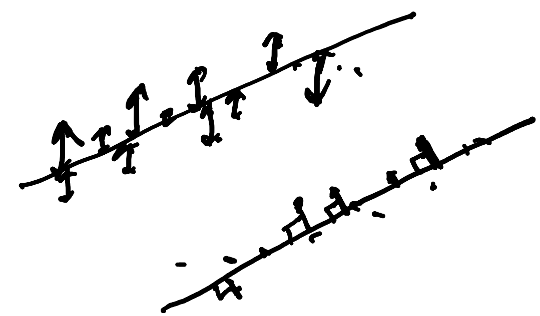
D eigenvectors

eval$(XX^T) = 0 \Rightarrow$ proj of X onto corresp v has no variance

Another interpretation

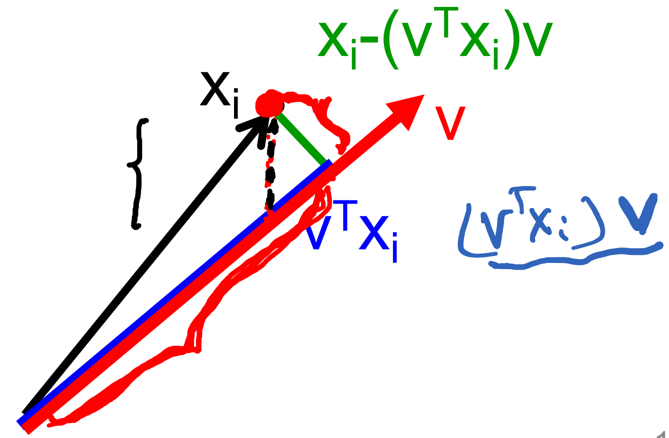
Maximum Variance Subspace: PCA finds vectors v such that projections on to the vectors capture maximum variance in the data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$



Minimum Reconstruction Error: PCA finds vectors v such that projection on to the vectors yields minimum MSE reconstruction

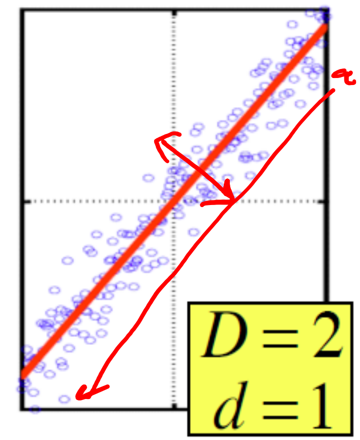
$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \underbrace{(\mathbf{v}^T\mathbf{x}_i)\mathbf{v}}_{\text{new representation}}\|^2$$



➢ Is this same as linear regression?

Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension.



Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

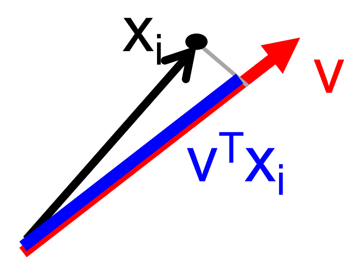
Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \ldots, v_d$ where $d = \text{rank}(XX^T)$

Original Representation
data point

$x_i = [x_i^1, x_i^2, \ldots. x_i^D]^T$
(D-dimensional vector)

Transformed representation
projections

$[v_1^T x_i, v_2^T x_i, \ldots\ v_d^T x_i]$
(d-dimensional vector)

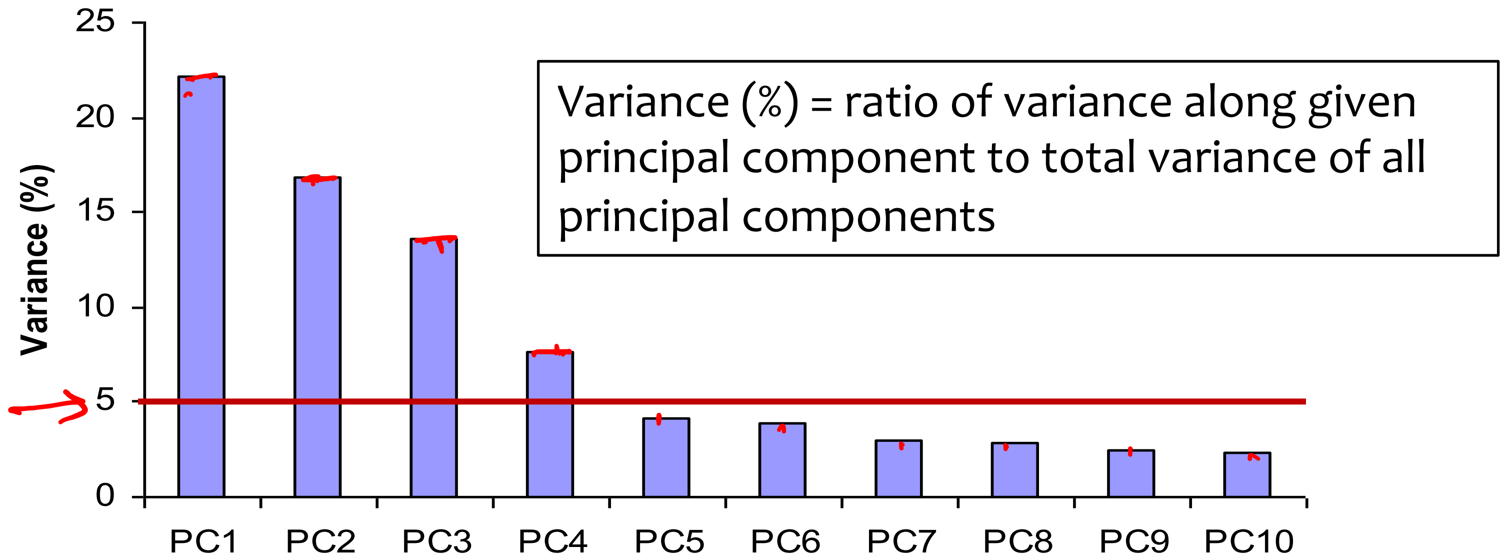


Dimensionality Reduction using PCA

In high-dimensional problem, data usually lies near a linear subspace, as noise introduces small variability

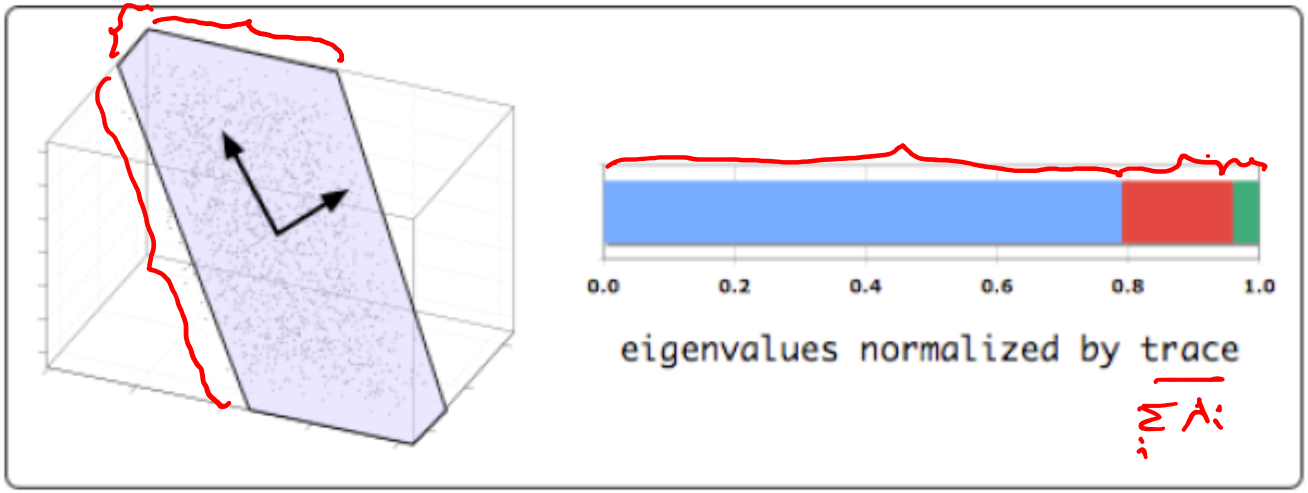
Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of lesser significance.



You might lose some information, but if the eigenvalues are small, you don't lose much

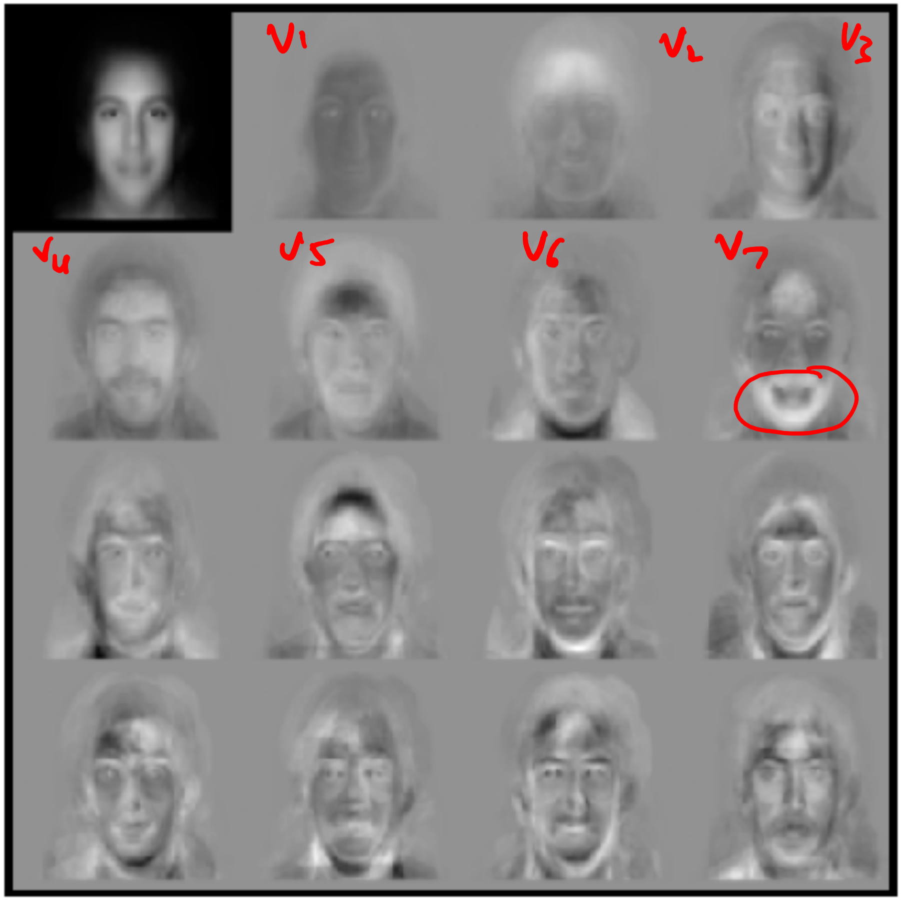
Example of PCA



Eigenvectors and eigenvalues of covariance matrix for $n=1600$ inputs in $d=3$ dimensions.

$x_i = \begin{bmatrix} \\ \\ \end{bmatrix}$ D-dim $\quad XX^T \xrightarrow{evec} v_1 \dots v_{15}$ D-dim

Example: faces



Eigenfaces from 7562 images:

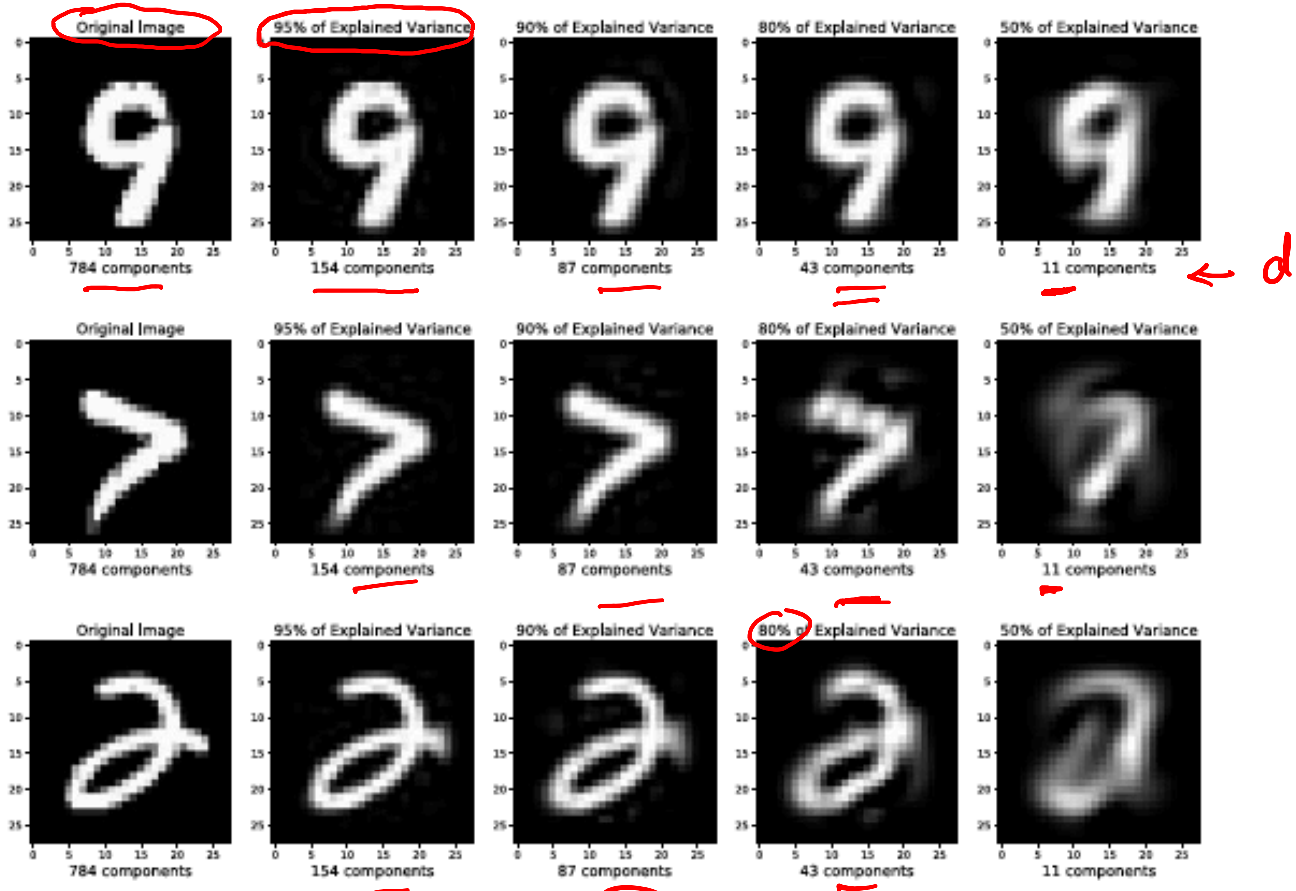
top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)

Example: MNIST digits

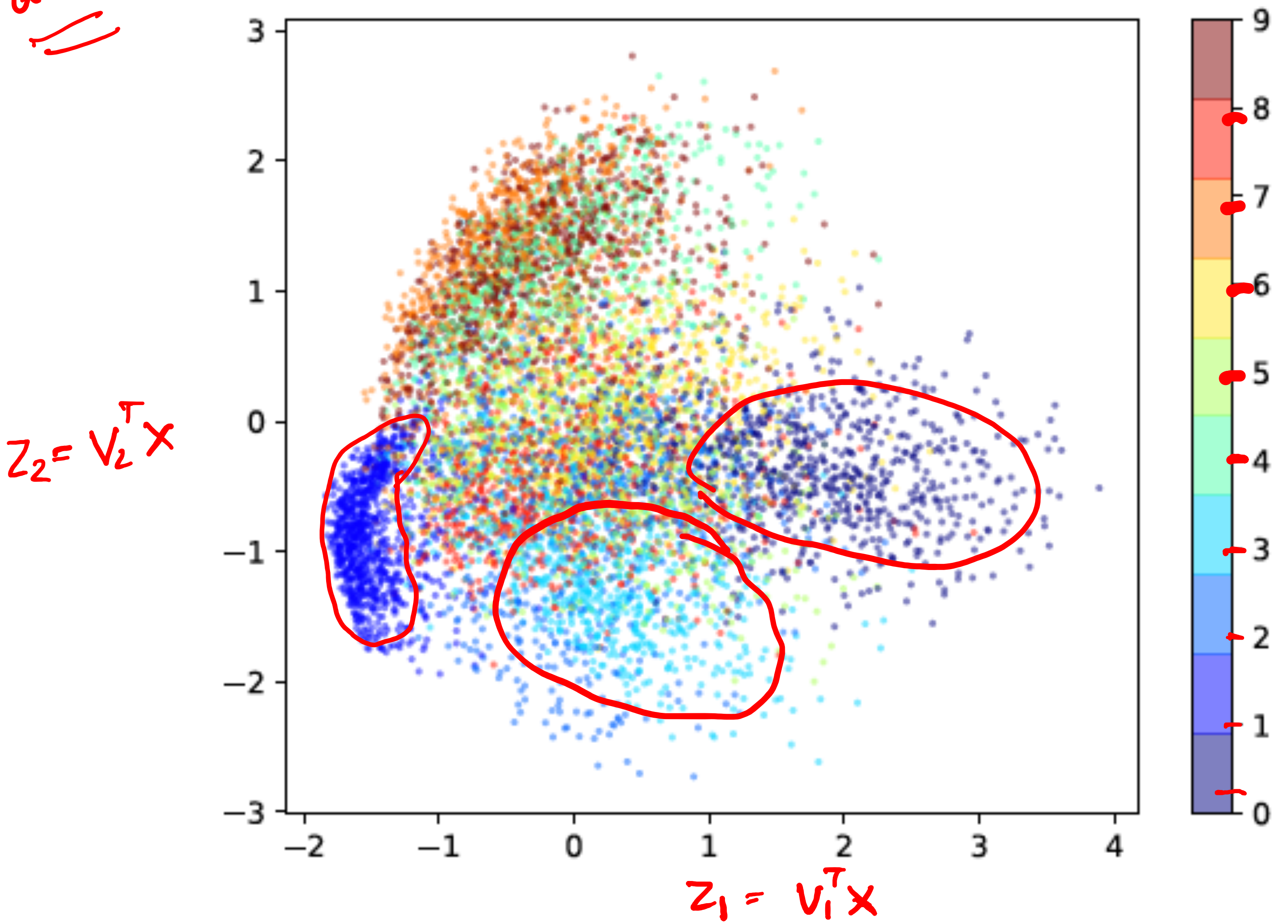
$D = 784$

- 28x28 images = 784 PCA vectors
- Project to K dimensional space and then project back up



Projecting MNIST digits

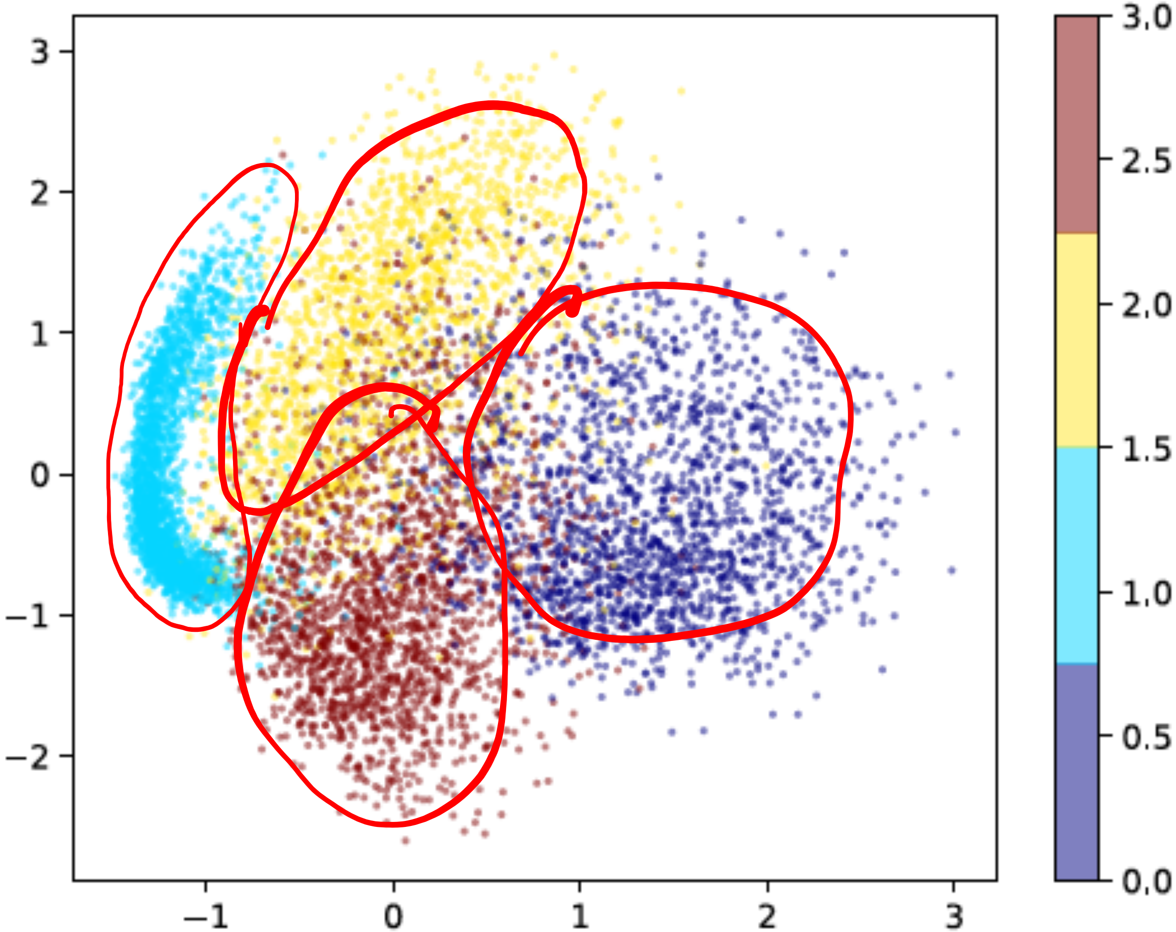
$d=2$

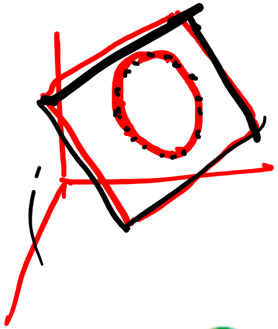


Projecting MNIST digits

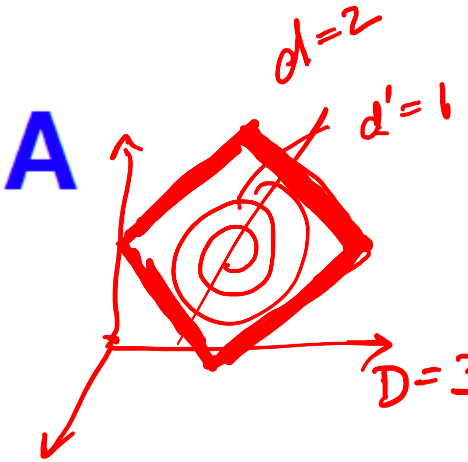
$D=784$

$d=2$

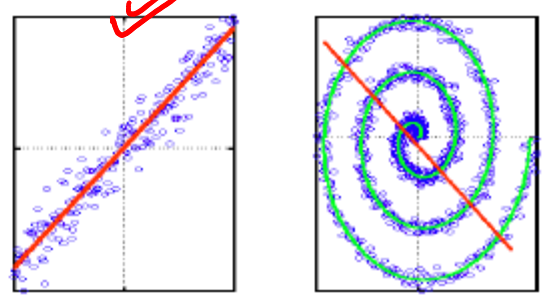




Properties of PCA



- **Strengths**
  - Eigenvector method
  - No tuning parameters
  - Non-iterative
  - No local optima



- **Weaknesses**
  - Limited to second order statistics
  - Limited to linear projections

XX^T

$x \rightarrow \phi(x)$