

#### 10-301/10-601 Introduction to Machine Learning

Machine Learning Department School of Computer Science Carnegie Mellon University

# Principal Component Analysis (PCA)

Matt Gormley, Henry Chai, Hoda Heidari Lecture 23 Apr. 10, 2024

## Reminders

- Homework 8: Deep RL
  - Out: Mon, Apr. 8
  - Due: Fri, Apr. 19 at 11:59pm

### **DIMENSIONALITY REDUCTION**

Examples of high dimensional data:

- High resolution images (millions of pixels)







Examples of high dimensional data:

– Multilingual News Stories

(vocabulary of hundreds of thousands of words)

<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	Prints the Sens	juo ari	bune. We Lad	+
<text><text><text><text><text><text></text></text></text></text></text></text>	Ym.m.	HILO, RAWAIL HAWAIIAN REGANDS, TUN	DAY, FEBRUARY at, 1905. No.	
<text><text><text><text><text><text></text></text></text></text></text></text>	lpe gilo Cetibune, Norteager's Intention	Antifer of KILAUFA CRATER	CONTINUES IN FRUPT	ON
<text><text><text><text><text><text></text></text></text></text></text></text>	Televise Auror Televise Auror Televise Auror Auroration Company, Lin Auroration Company, Lin	FOUR DAYS OF ACTIVITY	- HONOLULU BETS APPROPRIATION	\$
<text><text><text><text><text><text></text></text></text></text></text></text>		ANOTHER OUTBREAK	OCCURS FOR HARBOR AND LIGHTHO	IUSE
<text><text><text><text><text><text></text></text></text></text></text></text>	And the second s	and which said such from productions for the first here constrained where and on to active assign, its columnities the first here are neared and summary in the detailed for present-back, wonding tests of active is to back, and waves into and principal with length forgetting of additional to the second second active and principal with length forgettings of additional which evaluates the second second second second second second second which evaluates the second second second second second second which evaluates the second second second second second second second which evaluates the second sec	The first days The first days a high patrolin (The segmentation repress formula) as appropriation of \$40,000 main and Malana (The segmentation relates relates to Homitals hasher only, and as	dan da dar hele
<text><text><text><text><text><text></text></text></text></text></text></text>	rs. Grace and Irwin at his his -	at 2015, Linkind, die henrie the cambine of standing, standing, denking will be hand sample gamilie the dark familiense herizer, and an analiteasement has been of the familie the families must his sparker write denk of darksame, while Pitter hald warp in the molecule standard of darksame st	thing have some of the operation to all the blanch. This can will be authore the array as any the large the some [] the address the addres	nie No
<text><text><text><text><text><text></text></text></text></text></text></text>	Another particular and and and and and	a given that all and blockings. For the scene, when standing on the local meanings and here, in the scene of the flower of Fary, one residence of meanings meanings of the scene of the scene world that an help meanings of the scene of the scene world that within the scene of	ing older of the spectrum for the Barton Manufact barbar parent Honor. Konon Konon Konon Konon	
<text><text><text><text><text><text></text></text></text></text></text></text>	Chass, M. Lelilond	of the Court Brance in Sig is consistency with of long particular difficulture, while the minute of Brance, we a figs of Brance, we of Brance, we of Brance, we of Brance, we of Brance, we of Brance, we of Brance, we a figs of Brance, we a figs of Brance, we be of Brance, we a figs of Brance, we a figs of Brance, we be a brance Brance,	their offerences and an flow read serverity the filters built. In carries glopolations, including at lines for earrory Bill servery Bill servery Bill	Racher is brank
<page-header><text><text><text><text><text><text></text></text></text></text></text></text></page-header>	Altern and Annual Strengthers     Altern Annual Strengthers	LATER-Just as the TRIP, NE goes to p phone monage from the Volume Hears by a condition of the second	Postmaster Pratt Confirmed.	
<page-header><text><text><text><text><text><text></text></text></text></text></text></text></page-header>	idgway & Ridgway	the next stress of a Parties reserving from the trainer report that have a first through the handlike which checked the and a mogniture from this of the second to be a first the stress of the second to be a first t	Latest Sugar Quotations.	5mm ay.
<page-header><text><text><text><text><text><text></text></text></text></text></text></text></page-header>	Mart of Participant States of State of State of State of State of State of States		18 Manforth Bosts, upt allofa; per ten Bauli Ma	
	Beauty Colors to China Brane Brane China State of State o	the loss of the second	Fire & New Orleans.	
	Being from a class tree for a class of the	A the start and the formation of the start and the start a	Fire At New Orleans.	
	Republic to the second state of the second sta		Fire & New Orleans.	
	The second			
	00		5yet 799	
АУХА BRACA Потитории и потитории и потитори и потитори и потитори и потитори и потитори и потитори и	la		5yet 799	
	la		A day in the second sec	
<ul> <li>And Antin And Antin Anti</li></ul>	liar		And the second s	
<ul> <li>A starting to the starting to the</li></ul>	laar oran in an		And the second s	
mand many density given region and the second secon	laar oran in an		A state of the sta	
information and an and a second secon	laar oran in an		June     June       June     June <t< td=""><td></td></t<>	
The man is a second sec	laar oran in an		Image: State	
арадната и то таки отнати полнати и	laar oran in an	<section-header><section-header></section-header></section-header>		
нати при развил на при тото. правља даја и при тото. Паради при нати на при тото на при	laar oran in an	<section-header><section-header></section-header></section-header>	<text><text><text><text><text></text></text></text></text></text>	
a point of most range and		<section-header><section-header></section-header></section-header>	<page-header><text><text><text><text><text></text></text></text></text></text></page-header>	
рая было пропродъематерисания на нали на продата на на на на продата на на на на продата на на на продата на	Carl	<section-header><section-header></section-header></section-header>	<page-header><text><text><text><text><text><text><text></text></text></text></text></text></text></text></page-header>	



報功

#### Examples of high dimensional data: – Brain Imaging Data (100s of MBs per scan)

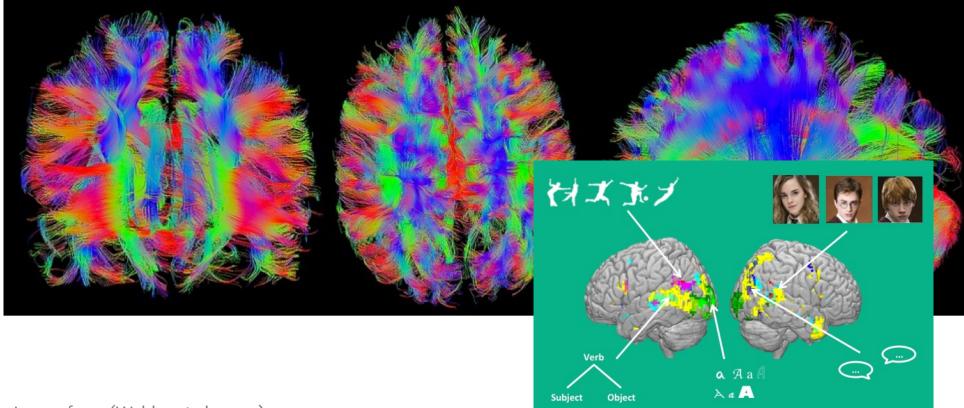
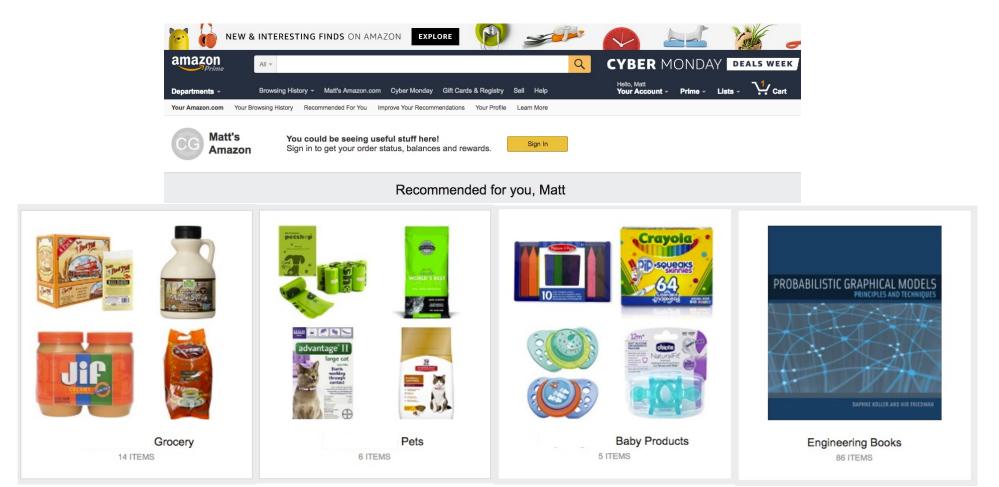


Image from (Wehbe et al., 2014)

Image from https://pixabay.com/en/brain-mrt-magnetic-resonance-imaging-1728449/

#### Examples of high dimensional data:

#### – Customer Purchase Data



# Learning Representations

#### **Dimensionality Reduction Algorithms:**

Powerful (often unsupervised) learning techniques for extracting hidden (potentially lower dimensional) structure from high dimensional datasets.

#### **Examples:**

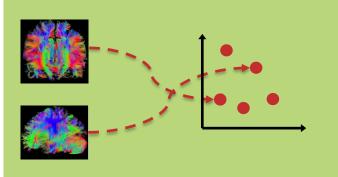
PCA, Kernel PCA, ICA, CCA, t-SNE, Autoencoders, Matrix Factorization

#### **Useful for:**

- Visualization
- More efficient use of resources (e.g., time, memory, communication)
- Statistical: fewer dimensions  $\rightarrow$  better generalization
- Noise removal (improving data quality)

#### This section in one slide...

#### 1. Dimensionality reduction:



#### 2. Random Projection:

FJ (1) Randonly sample matrix VERKXM (2) Project down:  $\vec{U}^{(i)} = V\vec{x}^{(i)}$ 

#### 4. Algorithm for PCA:

#### 3. Definition of PCA:

Choose the matrix V that either...

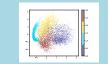
- 1. minimizes reconstruction error
- 2. consists of the K eigenvectors with largest eigenvalue

The above are equivalent definitions.

The option we'll focus on:

Run Singular Value Decomposition (SVD) to obtain all the eigenvectors. Keep just the top-K to form V. Play some tricks to keep things efficient.

5. An Example



# DIMENSIONALITY REDUCTION BY RANDOM PROJECTION

#### **Random Projection**

<u>Goal</u>: project from M-dimensions down to K-dimensions

Data:

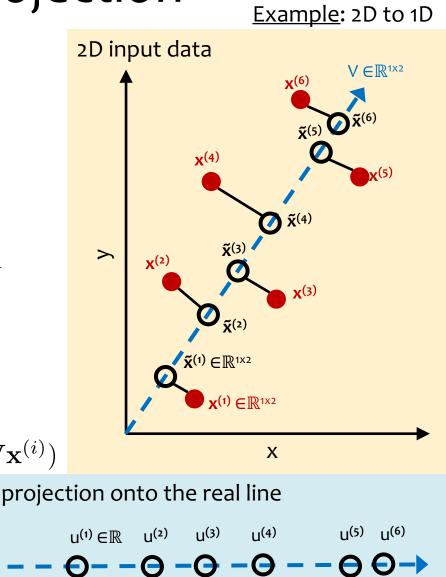
$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$$
 where  $\mathbf{x}^{(i)} \in \mathbb{R}^M$ 

<u>Algorithm</u>:

1. Randomly sample matrix:  $\mathbf{V} \in \mathbb{R}^{K \times M}$  $V_{km} \sim \text{Gaussian}(0, 1)$ 

2. Project down: 
$$\underbrace{\mathbf{u}^{(i)}}_{K \times 1} = \underbrace{\mathbf{V}}_{K \times MM \times 1} \underbrace{\mathbf{x}^{(i)}}_{K \times MM \times 1}$$

3. Project up: 
$$\tilde{\mathbf{x}}^{(i)}_{M \times 1} = \underbrace{\mathbf{V}^T}_{M \times KK \times 1} \mathbf{u}^{(i)} = \mathbf{V}^T (\mathbf{V} \mathbf{x}^{(i)})$$



#### **Random Projection** Example: 2D to 1D 2D input dat V ∈ ℝ<sup>1x2</sup> <u>Goal</u>: project from M-dimensions down to K-dimensions Data: $\mathcal{D} = { \mathbf{x}^{(i)} }_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^{M}$ <u>Algorithm</u>: > 1. Randomly sample matrix: $\mathbf{V} \in \mathbb{R}^{K \times M}$ $V_{km} \sim \text{Gaussian}(0,1)$ 2. Project down: $\mathbf{u}^{(i)} = \mathbf{V} \mathbf{x}^{(i)}$ $K \times 1$ $K \times MM \times T$ 3. Project up: $\mathbf{x}^{(i)} = \mathbf{V}^T \mathbf{u}^{(i)} = \mathbf{V}^T (\mathbf{V} \mathbf{x}^{(i)})$ Х $M \times KK \times 1$

**Problem:** a random projection might give us a poor low dimensional representation of the data

### Johnson-Lindenstrauss Lemma

- **Q:** But how could we ever hope to preserve any useful information by randomly projecting into a low-dimensional space?
- A: Even random projection enjoys some surprisingly impressive properties. In fact, a standard of the J-L lemma starts by assuming we have a random linear projection obtained by sampling each matrix entry from a Gaussian(0,1).

# An Elementary Proof of a Theorem of Johnson and Lindenstrauss

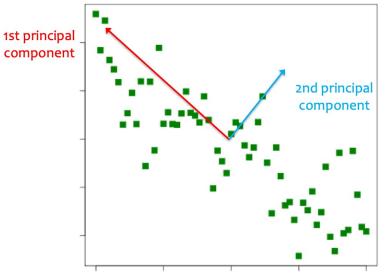
Sanjoy Dasgupta,<sup>1</sup> Anupam Gupta<sup>2</sup>

**ABSTRACT:** A result of Johnson and Lindenstrauss [13] shows that a set of *n* points in high dimensional Euclidean space can be mapped into an  $O(\log n/\epsilon^2)$ -dimensional Euclidean space such that the distance between any two points changes by only a factor of  $(1 \pm \epsilon)$ . In this note, we prove this theorem using elementary probabilistic techniques. © 2003 Wiley Periodicals, Inc. Random Struct. Alg., 22: 60-65, 2002

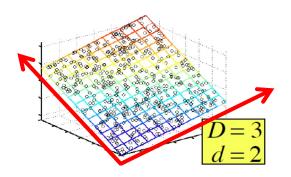
# DEFINITION OF PRINCIPAL COMPONENT ANALYSIS (PCA)

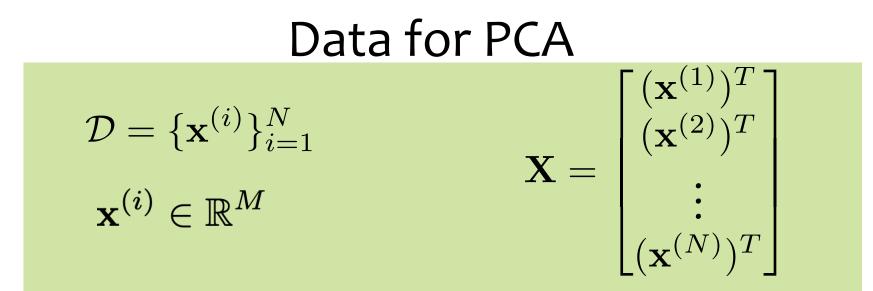
# Principal Component Analysis (PCA)

- Assumption: the data lies on a low Kdimensional linear subspace
- **Goal:** identify the axes of that subspace, and project each point onto hyperplane
- Algorithm: find the K eigenvectors with largest eigenvalue using classic matrix decomposition tools



PCA Example: 2D Gaussian Data





We assume the data is **centered**, i.e. the **sample mean** is zero

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} = \mathbf{0}$$

**Q:** What if your data is **not** centered?

A: Subtract off the sample mean  $ilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}, orall i$ 

### Background: Sample Variance

Suppose we have a sequence of random samples  $\{x^{(1)}, \ldots, x^{(N)}\}$  from a random variable X.

The (biased) sample variance  $\hat{\sigma}^2$  is given by:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \hat{\mu})^2$$

where  $\hat{\mu}$  is the sample mean.

#### Sample Covariance Matrix

The sample covariance matrix  $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$  is given by:

$$\Sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} (x_j^{(i)} - \mu_j) (x_k^{(i)} - \mu_k)$$

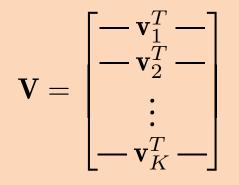
Since the data matrix is centered, we rewrite as:

$$\mathbf{\Sigma} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

# Principal Component Analysis (PCA)

Linear Projection: Given KxM matrix V, and Mx1 vector  $\mathbf{x}^{(i)}$  we obtain the Kx1 projection  $\mathbf{u}^{(i)}$  by:  $\mathbf{u}^{(i)} = \mathbf{V} \mathbf{x}^{(i)}$ 



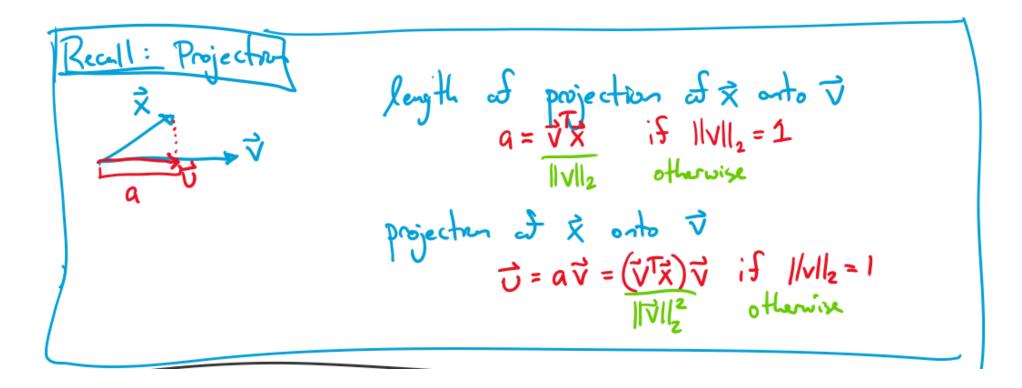
#### **Definition of PCA:**

**PCA** repeatedly chooses a next vector  $\mathbf{v}_j$  that minimizes the reconstruction error s.t.  $\mathbf{v}_j$  is orthogonal to  $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{j-1}$ .

Vector  $\mathbf{v}_{i}$  is called the **jth principal component**.

Notice: Two vectors **a** and **b** are **orthogonal** if  $\mathbf{a}^T \mathbf{b} = 0$ . The K-dimensions in PCA are uncorrelated

#### **Vector Projection**



## Objectives for PCA

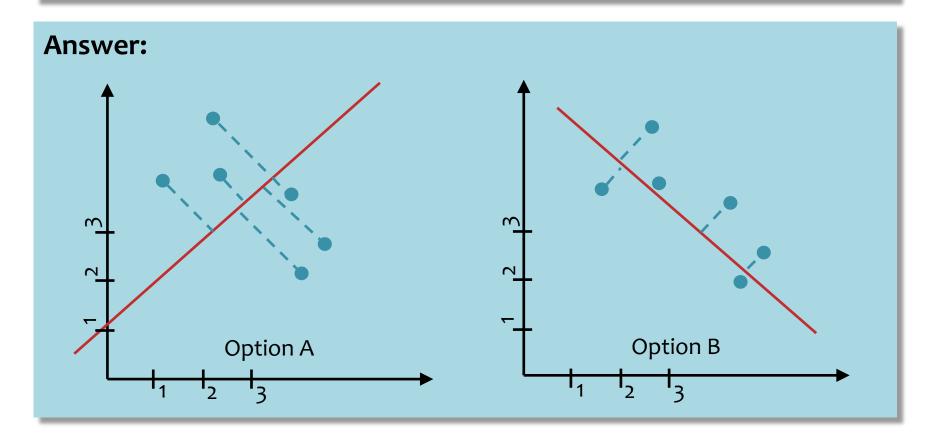
Minimize the Reconstruction Error Maximize the Variance

### Projection Example

#### **Question:**

Below are two plots of the same dataset D. Consider the two projections shown.

- 1. Which maximizes the variance?
- 2. Which minimizes the reconstruction error?



#### **PCA Objective Functions**

What is the first principal component  $v_1$  chosen by PCA?

Option 1: The vector that minimizes the reconstruction error

$$\mathbf{v}_{1} = \underset{\mathbf{v}:||\mathbf{v}||^{2}=1}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - (\mathbf{v}^{T} \mathbf{x}^{(i)}) \mathbf{v}||^{2}$$

Option 2: The vector that maximizes the variance

$$\mathbf{v}_1 = \operatorname*{argmax}_{\mathbf{v}:||\mathbf{v}||^2 = 1} \frac{1}{N} \sum_{i=1}^N (\mathbf{v}^T \mathbf{x}^{(i)})^2$$

Equivalence of Maximizing Variance and Minimizing Reconstruction Error

**Claim:** Minimizing the reconstruction error is equivalent to maximizing the variance.

**Proof:** First, note that:

$$||\mathbf{x}^{(i)} - (\mathbf{v}^T \mathbf{x}^{(i)})\mathbf{v}||^2 = ||\mathbf{x}^{(i)}||^2 - (\mathbf{v}^T \mathbf{x}^{(i)})^2$$
 (1)

since  $\mathbf{v}^T \mathbf{v} = ||\mathbf{v}||^2 = 1$ .

Substituting into the minimization problem, and removing the extraneous terms, we obtain the maximization problem.

$$\mathbf{v}^{*} = \underset{\mathbf{v}:||\mathbf{v}||^{2}=1}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - (\mathbf{v}^{T} \mathbf{x}^{(i)}) \mathbf{v}||^{2}$$
(2)

$$= \underset{\mathbf{v}:||\mathbf{v}||^{2}=1}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)}||^{2} - (\mathbf{v}^{T} \mathbf{x}^{(i)})^{2}$$
(3)

$$= \underset{\mathbf{v}:||\mathbf{v}||^2=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)})^2$$
(4)

### **PCA Objective Functions**

What is the first principal component  $v_1$  chosen by PCA?

Option 1: The vector that minimizes the reconstruction error

$$\mathbf{v}_{1} = \operatorname*{argmin}_{\mathbf{v}:||\mathbf{v}||^{2}=1} \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - (\mathbf{v}^{T} \mathbf{x}^{(i)})\mathbf{v}||^{2}$$

Option 2: The vector that maximizes the variance

$$\mathbf{v}_1 = \operatorname*{argmax}_{\mathbf{v}:||\mathbf{v}||^2 = 1} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)})^2$$

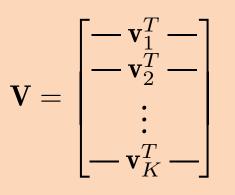
#### **Question:**

Why can't we just use gradient descent to find the minimum of the PCA optimization problem?

#### **Answer:**

# Principal Component Analysis (PCA)

Linear Projection: Given KxM matrix V, and Mx1 vector  $\mathbf{x}^{(i)}$  we obtain the Kx1 projection  $\mathbf{u}^{(i)}$  by:  $\mathbf{u}^{(i)} = \mathbf{V} \mathbf{x}^{(i)}$ 



#### **Question:**

Why can't we just use gradient descent to find the minimum of the PCA optimization problem?

#### **Definition of PCA:**

**PCA** repeatedly chooses a next vector  $\mathbf{v}_j$  that minimizes the reconstruction error s.t.  $\mathbf{v}_j$  is orthogonal to  $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{j-1}$ .

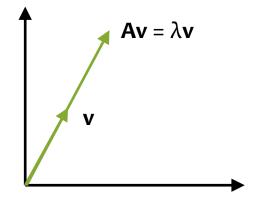
Vector  $\mathbf{v}_{j}$  is called the **jth principal component**.

Notice: Two vectors **a** and **b** are **orthogonal** if  $\mathbf{a}^T \mathbf{b} = 0$ . The K-dimensions in PCA are uncorrelated

#### **Answer:**

# Background: Eigenvectors & Eigenvalues For a square matrix A (n x n matrix), the vector v (n x 1 matrix) is an eigenvector iff there exists eigenvalue $\lambda$ (scalar) such that:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$



The linear transformation **A** is only stretching vector **v**.

That is,  $\lambda \mathbf{v}$  is a scalar multiple of  $\mathbf{v}$ .

# Background: Eigenvectors & Eigenvalues

Fact #1: The eigenvectors of a **symmetric matrix** are **orthogonal** to each other.

Fact #2: The **covariance matrix Σ** is **symmetric**.

#### The First Principal Component

**Claim:** The vector that maximizes the variances is the eigenvector of  $\Sigma$  with largest eigenvalue.

**Proof Sketch:** To find the first principal component, we wish to solve the following constrained optimization problem (variance minimization).

$$\mathbf{v}_1 = \operatorname*{argmax}_{\mathbf{v}:||\mathbf{v}||^2 = 1} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}$$
(1)

So we turn to the method of Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(\mathbf{v},\lambda) = \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)$$
(2)

Taking the derivative of the Lagrangian and setting to zero gives:

$$\frac{d}{d\mathbf{v}} \left( \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} - \lambda (\mathbf{v}^T \mathbf{v} - 1) \right) = 0$$
 (3)

$$\mathbf{\Sigma}\mathbf{v} - \lambda\mathbf{v} = 0 \tag{4}$$

$$\Sigma \mathbf{v} = \lambda \mathbf{v}$$
 (5)

Recall: For a square matrix **A**, the vector **v** is an **eigenvector** iff there exists **eigenvalue**  $\lambda$  such that:

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v} \tag{6}$$

Rewriting the objective of the maximization shows that not only will the optimal vector  $v_1$  be an eigenvector, it will be one with maximal eigenvalue.

$$\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} = \mathbf{v}^T \lambda \mathbf{v} \tag{7}$$

$$=\lambda \mathbf{v}^T \mathbf{v} \tag{8}$$

$$=\lambda ||\mathbf{v}||^2 \tag{9}$$

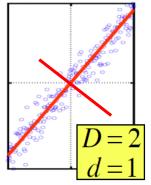
$$=\lambda$$
 (10)

#### Principal Component Analysis (PCA)

 $(X X^T)v = \lambda v$ , so v (the first PC) is the eigenvector of sample correlation/covariance matrix  $X X^T$ 

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

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



#### Eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$

- The 1<sup>st</sup> PC  $v_1$  is the the eigenvector of the sample covariance matrix  $X X^T$  associated with the largest eigenvalue
- The 2nd PC  $v_2$  is the the eigenvector of the sample covariance matrix  $X X^T$  associated with the second largest eigenvalue
- And so on ...

Slide from Nina Balcan

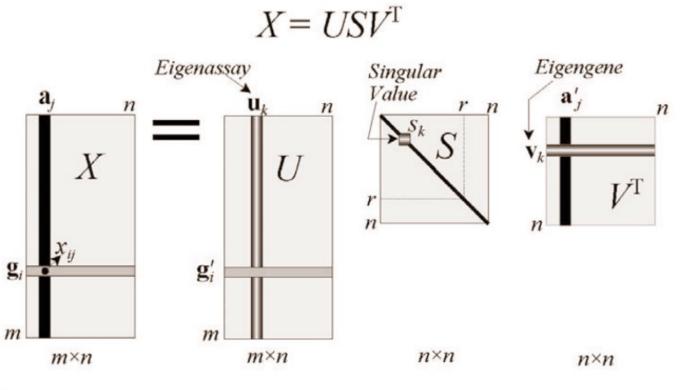
### **ALGORITHMS FOR PCA**

# Algorithms for PCA

How do we find principal components (i.e. eigenvectors)?

- Power iteration (aka. Von Mises iteration)
  - finds each principal component one at a time in order
- Singular Value Decomposition (SVD)
  - finds all the principal components at once
  - two options:
    - Option A: run SVD on  $X^T X$
    - Option B: run SVD on X (not obvious why Option B should work...)
- Stochastic Methods (approximate)
  - very efficient for high dimensional datasets with lots of points

SVD



Data X, one row per data point

US gives coordinates of rows of X in the space of principle components S is diagonal,  $S_k > S_{k+l}$ ,  $S_k^2$  is kth largest eigenvalue Rows of  $V^T$  are unit length eigenvectors of  $X^T X$ 

If cols of X have zero mean, then  $X^T X = c \Sigma$ and eigenvects are the Principle Components

#### **Singular Value Decomposition**

To generate principle components:

- Subtract mean  $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$  from each data point, to create zero-centered data
- Create matrix X with one row vector per (zero centered) data point
- Solve SVD:  $X = USV^T$
- Output Principle components: columns of V (= rows of  $V^T$ )
  - Eigenvectors in V are sorted from largest to smallest eigenvalues
  - S is diagonal, with  $s_k^2$  giving eigenvalue for kth eigenvector

#### **Singular Value Decomposition**

To project a point (column vector x) into PC coordinates:  $V^T x$ 

If  $x_i$  is i<sup>th</sup> row of data matrix X, then

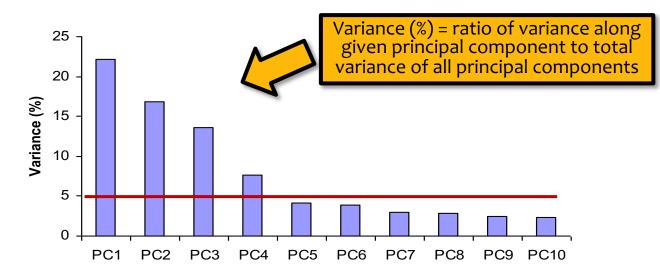
• (i<sup>th</sup> row of US) =  $V^T x_i^T$ 

• 
$$(US)^T = V^T X^T$$

To project a column vector x to M dim Principle Components subspace, take just the first M coordinates of  $V^T x$ 

### How Many PCs?

- For M original dimensions, sample covariance matrix is MxM, and has up to M eigenvectors. So M principal components (PCs).
- Where does dimensionality reduction come from? Can *ignore* the components of lesser significance.



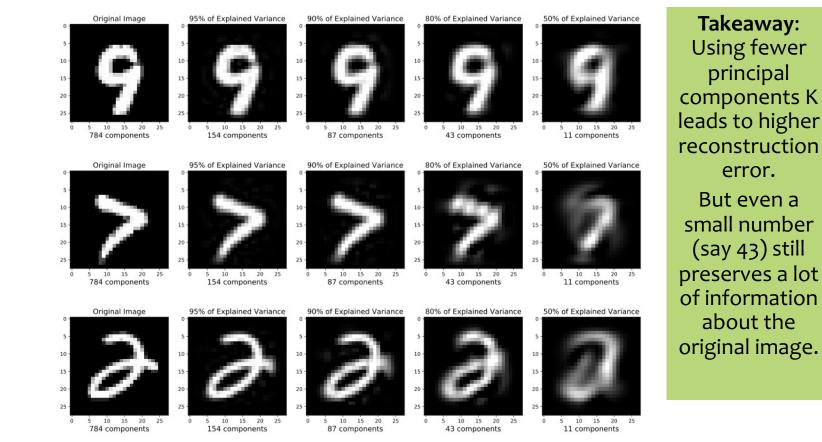
- You do lose some information, but if the eigenvalues are small, you don't lose much
  - M dimensions in original data
  - calculate M eigenvectors and eigenvalues
  - choose only the first D eigenvectors, based on their eigenvalues
  - final data set has only D dimensions

#### PCA EXAMPLES

# Projecting MNIST digits

#### Task Setting:

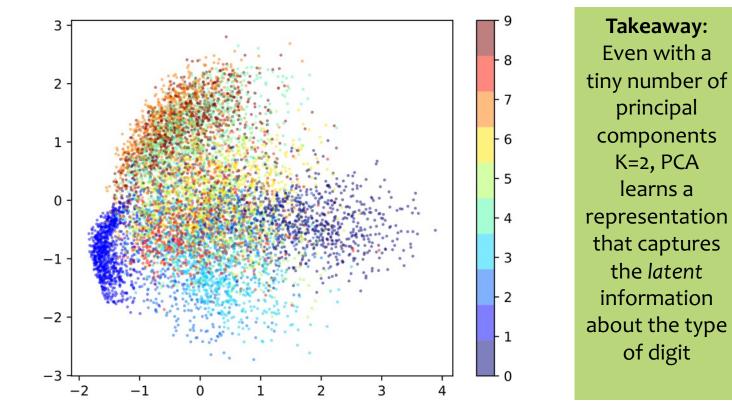
- 1. Take each 28x28 image of a digit (i.e. a vector  $\mathbf{x}^{(i)}$  of length 784) and project it down to K components (i.e. a vector  $\mathbf{u}^{(i)}$ )
- 2. Report percent of variance explained for K components
- 3. Then project back up to 28x28 image (i.e. a vector  $\mathbf{\tilde{x}}^{(i)}$  of length 784) to visualize how much information was preserved



# Projecting MNIST digits

#### Task Setting:

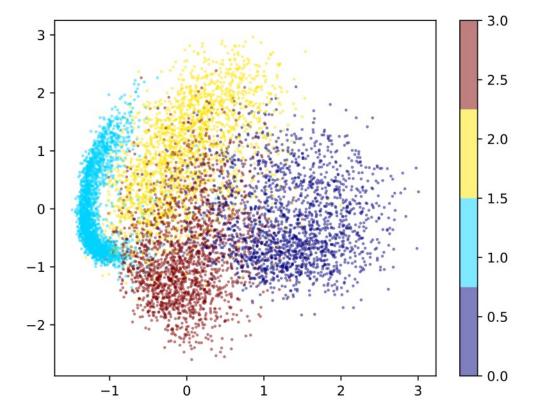
- 1. Take each 28x28 image of a digit (i.e. a vector  $\mathbf{x}^{(i)}$  of length 784) and project it down to K=2 components (i.e. a vector  $\mathbf{u}^{(i)}$ )
- 2. Plot the 2 dimensional points **u**<sup>(i)</sup> and label with the (unknown to PCA) label y<sup>(i)</sup> as the color
- 3. Here we look at all ten digits 0 9



# Projecting MNIST digits

#### Task Setting:

- 1. Take each 28x28 image of a digit (i.e. a vector  $\mathbf{x}^{(i)}$  of length 784) and project it down to K=2 components (i.e. a vector  $\mathbf{u}^{(i)}$ )
- 2. Plot the 2 dimensional points  $\mathbf{u}^{(i)}$  and label with the (unknown to PCA) label  $y^{(i)}$  as the color
- 3. Here we look at just four digits 0, 1, 2, 3



Takeaway: Even with a tiny number of principal components K=2, PCA learns a representation that captures the latent information about the type of digit

# Learning Objectives

#### **Dimensionality Reduction / PCA**

You should be able to...

- 1. Define the sample mean, sample variance, and sample covariance of a vector-valued dataset
- 2. Identify examples of high dimensional data and common use cases for dimensionality reduction
- 3. Draw the principal components of a given toy dataset
- 4. Establish the equivalence of minimization of reconstruction error with maximization of variance
- 5. Given a set of principal components, project from high to low dimensional space and do the reverse to produce a reconstruction
- 6. Explain the connection between PCA, eigenvectors, eigenvalues, and covariance matrix
- 7. Use common methods in linear algebra to obtain the principal components