

# Introduction to Data Compression

## A Linear Algebra Approach

Layan Ansari, Adam Pearl, Iñaki Arango

December 11, 2022

# Cover Letter

(We will write something here for the final product). Reflect on challenges and successes in the cover letter; what you planned to accomplish but couldn't; show that you put in effort and that you took a lot away from this project; reflect on what you learned while doing this project and how it may have helped you discover a new passion or delve into a passion you already had; thank whoever did your peer review.

# Contents

<b>Chapter</b>	<b>Introduction</b>	<b>Page 4</b>
<b>Chapter</b>	<b>Tools for Compression</b>	<b>Page 5</b>
2.1	Probability & Statistics	5
	Sample Space — 5 • Events — 6 • Probability — 6 • Random Variables — 6 • Expectation — 7 • Variance — 7 • Covariance — 8	
2.2	Information Theory (maybe could be skipped)	8
	Strictly required math/algebra information theory background — 8	
<b>Chapter</b>	<b>Start of the Story: What do we want to compress?</b>	<b>Page 9</b>
3.1	Handwriting Recognition	9
<b>Chapter</b>	<b>Principal Component Analysis</b>	<b>Page 11</b>
4.1	Extract “high information” regions in the data space	11
4.2	Quantify the intuitive findings from the information theory subsection	11
4.3	Formalizing the calculation of “high information” eigenvectors	11
4.4	Coding example perhaps?	16
<b>Chapter</b>	<b>Singular Value Decomposition</b>	<b>Page 17</b>
5.1	Explain how an alternative way to look at this is through SVD	17
<b>Chapter</b>	<b>Uses with Compression</b>	<b>Page 18</b>
6.1	XYZ storage	18
6.2	images (medical, handwiting, S XYZ), sound (voice recognition)	18

<b>Chapter</b>	<b>Uses to optimize and enable new kinds of algorithms (fingerprint detection)</b>	<b>Page 19</b>
7.1	XYZ recognition	19
7.2	Facial and handwritng detection algorithms, voice recognition algorithms	19
<b>Chapter</b>	<b>Discussion</b>	<b>Page 20</b>
8.1	Is this software Anglo-centric?	20
<b>Chapter</b>		<b>Page 21</b>
9.1	Random Examples	21
9.2	Random	22

# Chapter 1

## Introduction

Data flows everywhere, it is in every product that we interact with daily. More than 2.5 quintillion bytes are created every day (that is 2.5 followed by 18 zeros) [Ale15]. For the next 5 years data is expected to grow by 40% every year [Mar22].

We need, and will continue to need, more powerful and faster computers capable of processing this increasing amount of information, and better storage systems to safekeep it.

There has been a lot of improvement hardware-wise in the last couple of decades, which increased the density of our storage systems. However, what if we could improve our information storage density through other, non-physical, means? What if we could store the same amount of “information” but with fewer “bits”? This would allow us to work in parallel with scientists researching physical systems.

This is where data compression comes in. It allows us to store the same amount of “information” in less bits, digits, characters, or whatever the most basic unit of storage is in our medium of choice. An early example of compression work is the Morse Code, which was invented in 1838 for use in the telegraph industry. Telegraph bandwidth was scarce, so shorter code words were assigned for common letters in the English alphabet, such as “e” and “t”, that would be sent more often [Wol02].

Compression can also be useful for analysis and recognition of data patterns. Suppose that given a telgraph system, we had decided to make a picture for every single letter in the English alphabet, and instead of sending Morse code over a telegraph, we had sent the all the pixel information for the letter images every time we wanted to transmit a sentence across the wire. Not only would that have taken so much longer and wasted bandwidth (related to compression), but it would also have been way harder to recognize and would have been more error prone (people’s caligraphy is different from one another). One would have to record all the pixel information for every letter, reconstruct it, and then decide on which letter it is: a computationally hard and intensive task for both, a computer and a human.

Morse code, an example of compression, shows how crucial it is to find reliable ways of storing and sharing our information in a way that minimizes space used and maximizes reliability.

# Chapter 2

## Tools for Compression

There are various tools at our disposal that we can use to study and perform data compression and analysis. Many of the common techniques that are used today are the outcome of collaboration between scientists and engineers focused on different branches of Math, such as Probability, Statistics, Information Theory, and Linear Algebra.

While we have focused on Linear Algebra this semester, it is necessary to know some introductory concepts from these other branches to understand basic data compression. In this section we will introduce these branches and explain some of their basic concepts.

### 2.1 Probability & Statistics

Probability is the study of the the likelihood of events, independently and given the occurrence of other events (e.g., given that it is cloudy, what is the probability that it will rain today?). Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data [Wik22].

#### 2.1.1 Sample Space

The most basic concept in probability is that of the **sample space** and **events**.

##### Definition 2.1.1: Sample Space

The **sample space** is the set of all possible outcomes that an experiment can have.

##### Example 2.1.1 (Flipping a coin)

If we flip a coin then the sample space is  $S = \{\text{Heads}, \text{Tails}\}$ .

##### Example 2.1.2 (Rolling a die)

If we roll a single 6-sided die the sample space is  $S = \{1, 2, 3, 4, 5, 6\}$ .

## 2.1.2 Events

### Definition 2.1.2: Event

**Events** are then mathematically defined as a subset of the sample space.

### Example 2.1.3 (Rolling a die)

For example, in the 6-sided die example let  $M$  be the event we roll a 4 or greater. Then  $M = \{4, 5, 6\} \subseteq S$ .

In this experiment we state that we are rolling the die once, so we cannot make any statements about what happens if you first roll something, and the something else. In that case, one would have to redefine the experiment to say that each die roll has possibilities  $R = \{1, 2, 3, 4, 5, 6\}$  and the total experiment with  $n$  roles has sample space  $S = R^n$ .

## 2.1.3 Probability

### Definition 2.1.3: Probability

The **probability** of an event (denote  $P(M)$  for an event  $M$ ) is the measure of how likely the event is to occur as the outcome of a random experiment. It is the sum of the probabilities of each element of the set being the outcome of the performed experiment.

Probability is measured out of 1, which means that events certain to occur have a probability of 1 and events that will never occur have a probability of 0.

Since the sample space contains all the possible outcomes for the experiment, it is guaranteed that one the sample space's elements will be the outcome of the experiment, so

$$P(S) = 1$$

## 2.1.4 Random Variables

### Definition 2.1.4: Random Variables

**Random variables** (also r.v.'s) are functions that map the outcome of a random experiment (in the sample space) to a measurable quantity (e.g., real number) or a mathematical object (e.g., matrix).

When the experiment is performed, and the outcome is mapped to a value, we say the random variable “crystallizes” to that value. The set of all values a random variable crystallizes to is called the support, denoted  $R_X$  for a r.v.  $X$ . A specified random variable crystallizing to a certain value is an event.

We normally denote random variables with capital letters and their crystallized values with lower case letters.

### Example 2.1.4 (Flipping a coin)

Let us flip a fair coin with equal probabilities of landing heads or tails. The sample space is

then  $S = \{H, T\}$  with  $P(\{H\}) = P(\{T\}) = 0.5$ .

Let  $X$  be a random variable that crystallizes to -1 if the coin lands Heads and to 1 if the coin lands Tails. Then  $(X = -1)$  and  $(X = 1)$  are events and we write  $P(X = -1) = P(\{H\})$  and  $P(X = 1) = P(\{T\})$ .

If we were to write the probability function for  $X$ , where  $x$  is the crystallized value that  $X$  takes, we would get

$$P(X = x) = \begin{cases} 0.5 & \text{if } x = -1 \\ 0.5 & \text{if } x = 1 \end{cases}$$

### 2.1.5 Expectation

#### Definition 2.1.5: Expectation

The **expectation** (also called the mean) of a random variable  $X$ , denoted  $\mathbb{E}(X)$ , is the weighted average of the values that  $X$  can take. The weight for a value  $x$  is determined by the probability of  $X$  crystallizing to that value.

We also write the variable with a bar on top to represent its mean. Its formula is

$$\bar{X} = \mathbb{E}(X) = \sum_{x \in R_X} x \cdot P(X = x)$$

#### Example 2.1.5 (Rolling a die)

Expanding on the previous example of rolling a 6-sided die once. Let  $X$  be a random variable that crystallizes to the number rolled. Then  $R_X = \{1, 2, 3, 4, 5, 6\}$  and  $P(X = x) = \frac{1}{6}$  for all  $x \in R_X$ , since each number has the same probability of being rolled.

The expected value of  $X$  is then

$$\mathbb{E}(X) = \sum_{x \in R_X} x \cdot P(X = x) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = 3.5$$

### 2.1.6 Variance

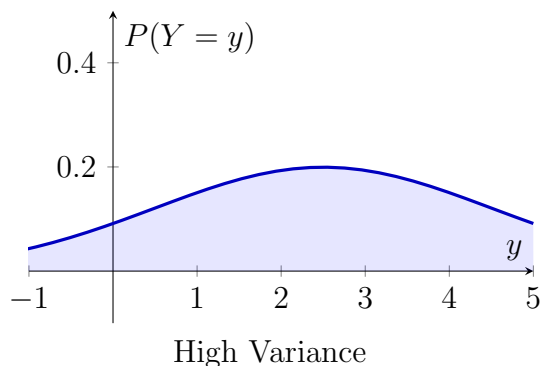
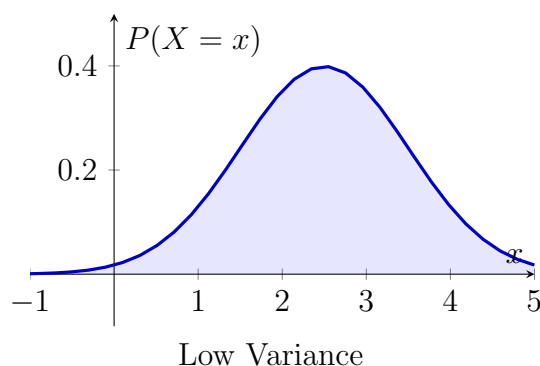
#### Definition 2.1.6: Variance

**Variance** is a measure of spread. It measures how far the values in the support of an r.v. are to the mean. For a r.v.  $X$  we denote its variance as  $\text{Var}(X)$ . The formula for variance is

$$\text{Var}(X) = \mathbb{E}[(X - \bar{X})^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Below is a graphical comparison of between the probability function of a low variance r.v.  $X$  and the probability function of a high variance r.v.  $Y$ . The horizontal axis indicates a value in the support of the random variable and the vertical axis indicates the probability of the random variable taking that value.





## 2.1.7 Covariance

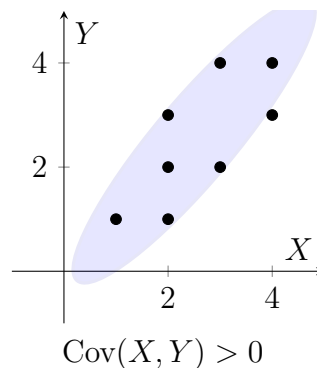
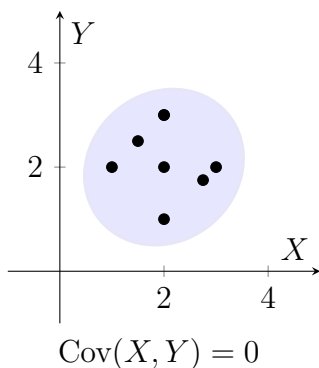
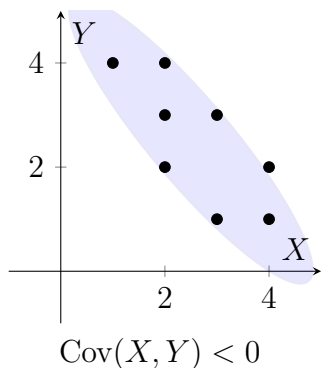
### Definition 2.1.7: Covariance

While the variance is the measure of spread for a single variables, the **covariance** examines the directional relationship between two variables [Ric07]. Its formula is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \bar{X})(Y - \bar{Y})] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

If greater values of one variable correspond to greater values of the other variable covariance is positive. If greater values of one variable correspond to lesser values of the other variable, covariance is negative.

Below is a graphical comparison of what negative, zero and positive covariance looks like between two random variables. Each point encodes one instance of the two variables crystallized together.



If covariance is 0, we say that variables are “uncorrelated”, and we interpret it as there being no linear relationship between the two.

## 2.2 Information Theory (maybe could be skipped)

### 2.2.1 Strictly required math/algebra information theory background

**Iñaki:**  
Fill  
out  
with  
defini-  
tions  
or

# Chapter 3

## Start of the Story: What do we want to compress?

### 3.1 Handwriting Recognition

Communication is an essential part of relating to people, and one of the oldest and most accessible methods of communication within a given language is writing by hand. In spite of the major effort that has been expended to bring about a paper-free society, a very large number of paper-based documents are processed daily by computers all over the world in order to handle, retrieve, and store information. The problem is that the manual process used to enter the data from these documents into computers demands a great deal of time and money (Bortolozzi, de Souza Britto Jr., Oliveira and Morita, n.d.). These documents may need to be processed for a number of reasons, among them historical documentation (e.g. digitally documenting culturally and historically significant documents and scripts, which until recently were more often than not handwritten or on print paper), recognition for medical prescriptions, or for tablet soft-wares to convert user's handwriting into digital text.

**Layan:**  
put in  
cita-  
tion  
later

Thus, the task of handwriting recognition is the transcription of handwritten data into a digital format, and this task obviously benefits from data compression. The goal is to process handwritten data electronically with the same or nearly the same accuracy as humans (Gunter, n.d.).

**Kai:**  
Here  
you  
write  
that  
these  
docu-  
ments  
may  
need  
to be  
pro-  
cessed  
for a  
num-  
ber of  
rea-  
sons,  
but  
go on

Basically, handwriting can be divided into two categories, cursive script and printed handwriting. Accuracy is the main problem in handwriting recognition for both categories because of the similar strokes and shapes some letters may possess. The software may have an inaccurate recognition of the letter, considering the possibility of the handwriting being illegible or some other factors (). One notable problem that makes this task difficult especially in cursive handwriting recognition is the fact that there may be no obvious character boundaries (the start and end of a character); compared to printed handwriting, it does not have gaps or spaces between each letter to know the start and stop of recognition per character (). This issue is compounded for languages like Arabic, where cursive is the only form of script and there exist “shortcuts” to further simplify the cursive script and make writing more fluid (e.g. removing the “dots”/accent marks that exist above/below certain letters).

This is where the data compression/dimension reduction and eigenface technique comes into play! In essence, if we have a large data set that consists of thousands or even millions of images of words (probably limited to one language), we want to find a way to recognize patterns in these images. From these patterns, we can then determine which ones have the most “importance” and attempt to express these images as a weighted combination of these most important patterns (and thus in a lower dimension).

# Chapter 4

## Principal Component Analysis

### 4.1 Extract “high information” regions in the data space

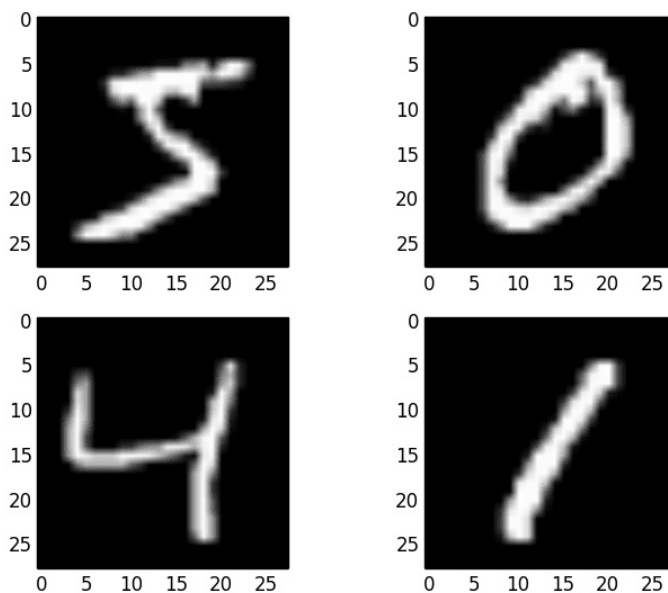


Figure 4.1: Samples of the numbers 5, 0, 4, and 1 from the MNIST handwriting dataset.

### 4.2 Quantify the intuitive findings from the information theory subsection

### 4.3 Formalizing the calculation of “high information” eigenvectors

**Iñaki:**

Use our variance knowledge from example image to say that some sections of the image have less information (the constant ones) and some have more information (the ones

The motivation for using principal component analysis is to find what Turk and Pentland referred to as “face space,” and which in our case we can conceptualize as the data space of letters. The face images, as previously explained, live as vectors in large dimensional spaces (65,536-dimensional space for a typical image). The goal at this stage is to reduce our large input vector space to a lower dimensional subspace that can be described by an eigenbasis.

#### Definition 4.3.1: Covariance

In order to construct... we need to define the **covariance**. Covariance measures the overall correlation between certain variables. The covariance  $\sigma$  of two variables,  $X$  and  $Y$ , is defined

$$\sigma(X, Y) = \frac{1}{M} \sum_{i=1}^M (X_i - \bar{X})(Y_i - \bar{Y})$$

#### Definition 4.3.2: Covariance Matrix

**Covariance matrix**, then, is a matrix that measures the covariance of various variables. For variables  $x_1, \dots, x_p$ , the covariance matrix is defined as follows

$$C = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_{pp} \end{bmatrix}$$

where  $\sigma_{ij}$  represents the covariance between the  $i^{th}$  and  $j^{th}$  variable. In the case of eigenfaces, the  $p$  independent variables correspond to our  $N^2$  bits, and thus  $p$  corresponds to the dimension of our images.

Having thus defined our covariance matrix, we must now discuss why we are taking the eigenvectors of this particular matrix to be our eigenfaces. In order for our eigenfaces to be as efficient as possible, we want each eigenface to encode as much variance as possible.

We take eigenvectors to simplify a relation defined by matrix multiplication to a relation defined by scalar multiplication, so that we may write any face as a linear combination of eigenvectors. Taking an eigenvector  $e$ ,  $Ce = \lambda e$ . If  $\lambda$  is large, then the covariance between  $e$  and all many other faces is high. If  $\lambda$  is small, then the covariance is less. Therefore, we will see later that we want to prioritize large eigenvalues when constructing our eigenbasis.

Before we discuss the eigenvalues, however, we must derive some more definitions of  $C$ .

Firstly, given the large size of  $N^2$ , it is convenient for us to find a more condensed description of the covariance matrix

#### Theorem 4.3.1

**Kai:**  
I think more explanation of face space might help understanding

**Kai:**  
We certainly do not want to work with that!

**Kai:**  
is this also kinda defined in chapter 2?

**Kai:**  
Is this N previously defined

Let  $C$  be a  $p \times p$  covariance matrix where each element  $c_{ij} = \sigma_{ij}$ . Let  $\Phi_a = P_a - \bar{P}$ , where

$$P_a = \begin{bmatrix} x_{1a} \\ x_{2a} \\ \vdots \\ x_{pa} \end{bmatrix}$$

with  $a$  from 1 to  $M$ .

Then  $C = \frac{1}{M} \sum_{a=1}^M \Phi_a \Phi_a^T$ .

**Proof:** We can prove this theorem starting with

$$C = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_{pp} \end{bmatrix}$$

From the definition of covariance, each entry  $c_{ij} = \frac{1}{M} \sum_{i=1}^M (x_{ia} - \bar{x}_i)(x_{ia} - \bar{x}_i)$ . Since matrices add linearly, we can factor out the summation  $\frac{1}{M} \sum_{i=1}^M$ . Thus we can write

$$C = \frac{1}{M} \sum_{i=1}^M \begin{bmatrix} (x_{1a} - \bar{x}_1)(x_{1a} - \bar{x}_1) & \cdots & (x_{1a} - \bar{x}_1)(x_{pa} - \bar{x}_p) \\ \vdots & \ddots & \\ (x_{pa} - \bar{x}_p)(x_{1a} - \bar{x}_1) & \cdots & (x_{pa} - \bar{x}_p)(x_{pa} - \bar{x}_p) \end{bmatrix}$$

Now consider the term  $\Phi_a \Phi_a^T = (P_i - \bar{P})(P_i - \bar{P})^T$ , as defined in the theorem. Since the transpose operator is distributive,

$$\begin{aligned} (P_i - \bar{P})(P_i - \bar{P})^T &= \left( \begin{bmatrix} x_{1a} \\ x_{2a} \\ \vdots \\ x_{pa} \end{bmatrix} - \bar{P} \right) \left( \begin{bmatrix} x_{1i} & x_{2i} & \cdots & x_{pi} \end{bmatrix} - \bar{P}^T \right) \\ &= \left( \begin{bmatrix} x_{1a} \\ x_{2a} \\ \vdots \\ x_{pa} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \right) \left( \begin{bmatrix} x_{1a} & x_{2a} & \cdots & x_{pa} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} \right) \\ &= \begin{bmatrix} x_{1a} - \bar{x}_1 \\ x_{2a} - \bar{x}_2 \\ \vdots \\ x_{pa} - \bar{x}_p \end{bmatrix} \begin{bmatrix} x_{1a} - \bar{x}_1 & x_{2a} - \bar{x}_2 & \cdots & x_{pa} - \bar{x}_p \end{bmatrix} \end{aligned}$$

**Iñaki:**

Clarify notation because  $\bar{x}_p$  is not defined anywhere

Therefore,

$$\Phi_a \Phi_a^T = \begin{bmatrix} (x_{1a} - \bar{x}_1)(x_{1a} - \bar{x}_1) & \dots & (x_{1a} - \bar{x}_1)(x_{pa} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{pa} - \bar{x}_p)(x_{1a} - \bar{x}_1) & & (x_{pa} - \bar{x}_p)(x_{pa} - \bar{x}_p) \end{bmatrix}$$

$$C = \frac{1}{M} \sum_{a=1}^M \Phi_a \Phi_a^T$$

☺

By constructing

### REFERENCE PRINCIPAL COMPONENT ANALYSIS

In the eigenface case, we wish to measure the difference between our face vectors  $\Gamma_i$  differ from the average face. We can define the average face  $\bar{\Gamma}$  as follows,

$$\bar{\Gamma} = \frac{1}{M} \sum_{n=1}^M \Gamma_n$$

We can write the difference  $\Phi_i$  between the average face and the  $i^{th}$  face,

$$\Phi_i = \Gamma_i - \bar{\Gamma}$$

Hello

#### Theorem 4.3.2

Let  $C$  be an  $N^2 \times N^2$  covariance matrix, defined as  $C = \frac{1}{M} \sum_{n=1}^M \Phi_n \Phi_n^T$ , where  $M$  is the dimension of the domain  $C$ . Let  $A$  be an  $N^2 \times N^2$  matrix with columns  $[\Phi_1 \ \Phi_2 \ \dots \ \Phi_M]$ . Then  $C = AA^T$ .

**Proof:**

☺

Going back to our matrix  $C$ ,

$$C = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_{pp} \end{bmatrix}$$

We can expand  $C$  according to the definition of  $\sigma$ ,

$$C = \frac{1}{M} \sum_{a=1}^M \begin{bmatrix} \Phi_1 \Phi_1^T & \Phi_1 \Phi_2^T & \dots & \Phi_1 \Phi_p^T \\ \Phi_2 \Phi_1^T & \Phi_2 \Phi_2^T & & \\ \vdots & & \ddots & \\ \Phi_p \Phi_1^T & & & \Phi_p \Phi_p^T \end{bmatrix}$$

Now we can express our eigenvectors  $v_i$  and eigenvalues  $\lambda_i$  as

$$\lambda_i v_i = AA^T v_i$$

However,  $AA^T$  is an  $N^2$  by  $N^2$  matrix (where  $N \times N$  is the resolution of the image), which for images of any discernable resolution will be much larger than desired in order to compute its eigenvectors.

Let

$$S = AA^T = \lambda_i v_i$$

Then,

$$S^T = (AA^T)^T = A^T A$$

We can show that  $S$  and  $S^T$  have the same eigenvalues.

Since the identity matrix is symmetrical, since the transpose operation is distributive, and since  $\det(A) = \det(A^T)$ ,

$$\det(S^T - \lambda_i I) = \det(S - \lambda_i I)^T = \det(S - \lambda_i I)$$

and therefore  $S$  and  $S^T$  have the same eigenvalues.

Applying this to our eigenfaces,

$$S^T = A^T A = \lambda_i u_i$$

The advantage of effectively reordering the  $A$  and  $A^T$  is that our new matrix  $S^T$  is an  $M \times M$  matrix.  $M$  corresponds to the size of the training set, which in most cases is much smaller than the dimension of the resolution of the image. Therefore, it will be much easier to compute the eigenfaces. We can then write each face as a linear combination of the eigenfaces

MEGAREFERENCE

$$\Gamma_i = \sum_{n=1}^M c_{in} u_n$$

We can interpret the eigenvectors of the covariance matrix as vectors that encode the greatest variance in data, and as such are the most suited for reconstructing faces from (ghosts). The corresponding eigenvalues tell how much variance each eigenvector encodes. For this reason, the greater the eigenvalue the more adaptable and better suited the eigen vector. When choosing eigenvectors to form a basis for the eigenspace used for facial recognition, one therefore prioritizes them by eigenvalue from highest to lowest.

AWDAWD



## 4.4 Coding example perhaps?

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

## Chapter 5

# Singular Value Decomposition

- 5.1 Explain how an alternative way to look at this is through SVD

# Chapter 6

## Uses with Compression

### 6.1 XYZ storage

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

### 6.2 images (medical, handwiting, S XYZ), sound (voice recognition)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

# Chapter 7

## Uses to optimize and enable new kinds of algorithms (fingerprint detection)

### 7.1 XYZ recognition

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

### 7.2 Facial and handwriting detection algorithms, voice recognition algorithms

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

# Chapter 8

## Discussion

### 8.1 Is this software Anglo-centric?

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

# Chapter 9

## 9.1 Random Examples

### Definition 9.1.1: Limit of Sequence in $\mathbb{R}$

Let  $\{s_n\}$  be a sequence in  $\mathbb{R}$ . We say

$$\lim_{n \rightarrow \infty} s_n = s$$

where  $s \in \mathbb{R}$  if  $\forall$  real numbers  $\epsilon > 0 \exists$  natural number  $N$  such that for  $n > N$

$$s - \epsilon < s_n < s + \epsilon \text{ i.e. } |s - s_n| < \epsilon$$

### Question 1

Is the set  $x\text{-axis} \setminus \{\text{Origin}\}$  a closed set

**Solution:** We have to take its complement and check whether that set is a open set i.e. if it is a union of open balls

### Note:

We will do topology in Normed Linear Space (Mainly  $\mathbb{R}^n$  and occasionally  $\mathbb{C}^n$ ) using the language of Metric Space

### Claim 9.1.1 Topology

Topology is cool

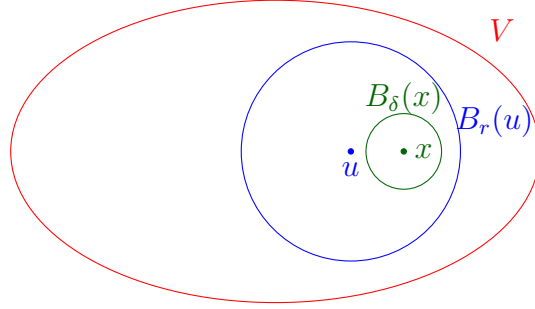
### Example 9.1.1 (Open Set and Close Set)

- Open Set:
- $\phi$
  - $\bigcup_{x \in X} B_r(x)$  (Any  $r > 0$  will do)
  - $B_r(x)$  is open
- Closed Set:
- $X, \phi$
  - $\overline{B_r(x)}$
- $x\text{-axis} \cup y\text{-axis}$

**Theorem 9.1.1**

If  $x \in$  open set  $V$  then  $\exists \delta > 0$  such that  $B_\delta(x) \subset V$

**Proof:** By openness of  $V$ ,  $x \in B_r(u) \subset V$



Given  $x \in B_r(u) \subset V$ , we want  $\delta > 0$  such that  $x \in B_\delta(x) \subset B_r(u) \subset V$ . Let  $d = d(u, x)$ . Choose  $\delta$  such that  $d + \delta < r$  (e.g.  $\delta < \frac{r-d}{2}$ )

If  $y \in B_\delta(x)$  we will be done by showing that  $d(u, y) < r$  but

$$d(u, y) \leq d(u, x) + d(x, y) < d + \delta < r$$

**Corollary 9.1.1**

By the result of the proof, we can then show...

**Lemma 9.1.1**

Suppose  $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$  is subspace of  $\mathbb{R}^n$ .

**Proposition 9.1.1**

$1 + 1 = 2$ .

## 9.2 Random

**Definition 9.2.1: Normed Linear Space and Norm  $\|\cdot\|$** 

Let  $V$  be a vector space over  $\mathbb{R}$  (or  $\mathbb{C}$ ). A norm on  $V$  is function  $\|\cdot\| : V \rightarrow \mathbb{R}_{\geq 0}$  satisfying

- ①  $\|x\| = 0 \iff x = 0 \forall x \in V$
- ②  $\|\lambda x\| = |\lambda| \|x\| \forall \lambda \in \mathbb{R}(\text{or } \mathbb{C}), x \in V$
- ③  $\|x + y\| \leq \|x\| + \|y\| \forall x, y \in V$  (Triangle Inequality/Subadditivity)

And  $V$  is called a normed linear space.

• Same definition works with  $V$  a vector space over  $\mathbb{C}$  (again  $\|\cdot\| \rightarrow \mathbb{R}_{\geq 0}$ ) where ② becomes  $\|\lambda x\| = |\lambda| \|x\| \forall \lambda \in \mathbb{C}, x \in V$ , where for  $\lambda = a + ib$ ,  $|\lambda| = \sqrt{a^2 + b^2}$

**Example 9.2.1 ( $p$ -Norm)**

$V = \mathbb{R}^m$ ,  $p \in \mathbb{R}_{\geq 0}$ . Define for  $x = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$

$$\|x\|_p = \left( |x_1|^p + |x_2|^p + \dots + |x_m|^p \right)^{\frac{1}{p}}$$

(In school  $p = 2$ )

**Special Case  $p = 1$ :**  $\|x\|_1 = |x_1| + |x_2| + \dots + |x_m|$  is clearly a norm by usual triangle inequality.

**Special Case  $p \rightarrow \infty$  ( $\mathbb{R}^m$  with  $\|\cdot\|_\infty$ ):**  $\|x\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_m|\}$   
For  $m = 1$  these  $p$ -norms are nothing but  $|x|$ . Now exercise

**Question 2**

Prove that triangle inequality is true if  $p \geq 1$  for  $p$ -norms. (What goes wrong for  $p < 1$  ?)

**Solution:** For Property ③ for norm-2

When field is  $\mathbb{R}$  :

We have to show

$$\begin{aligned} \sum_i (x_i + y_i)^2 &\leq \left( \sqrt{\sum_i x_i^2} + \sqrt{\sum_i y_i^2} \right)^2 \\ \Rightarrow \sum_i (x_i^2 + 2x_i y_i + y_i^2) &\leq \sum_i x_i^2 + 2\sqrt{\left[ \sum_i x_i^2 \right] \left[ \sum_i y_i^2 \right]} + \sum_i y_i^2 \\ \Rightarrow \left[ \sum_i x_i y_i \right]^2 &\leq \left[ \sum_i x_i^2 \right] \left[ \sum_i y_i^2 \right] \end{aligned}$$

So in other words prove  $\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$  where

$$\langle x, y \rangle = \sum_i x_i y_i$$

**Note:**

- $\|x\|^2 = \langle x, x \rangle$
- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle \cdot, \cdot \rangle$  is  $\mathbb{R}$ -linear in each slot i.e.

$$\langle rx + x', y \rangle = r\langle x, y \rangle + \langle x', y \rangle \text{ and similarly for second slot}$$

Here in  $\langle x, y \rangle$   $x$  is in first slot and  $y$  is in second slot.



Now the statement is just the Cauchy-Schwartz Inequality. For proof

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$$

expand everything of  $\langle x - \lambda y, x - \lambda y \rangle$  which is going to give a quadratic equation in variable  $\lambda$

$$\begin{aligned} \langle x - \lambda y, x - \lambda y \rangle &= \langle x, x - \lambda y \rangle - \lambda \langle y, x - \lambda y \rangle \\ &= \langle x, x \rangle - \lambda \langle x, y \rangle - \lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle \\ &= \langle x, x \rangle - 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \end{aligned}$$

Now unless  $x = \lambda y$  we have  $\langle x - \lambda y, x - \lambda y \rangle > 0$  Hence the quadratic equation has no root therefore the discriminant is greater than zero.

**When field is  $\mathbb{C}$  :**

Modify the definition by

$$\langle x, y \rangle = \sum_i \overline{x_i} y_i$$

Then we still have  $\langle x, x \rangle \geq 0$

# Bibliography

- [Ale15] Alexaqz. Data is expected to double every two years for the next decade, Aug 2015.
- [Mar22] Bernard Marr. How much data do we create every day? the mind-blowing stats everyone should read, Oct 2022.
- [Ric07] John Rice. *Mathematical Statistics and Data Analysis*. Brooks/Cole Cengage Learning, 2007.
- [Wik22] Wikipedia. Statistics — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Statistics&oldid=1121639148>, 2022. [Online; accessed 29-November-2022].
- [Wol02] Stephen Wolfram. *A New Kind of Science*. Wolfram Media, 2002.