Introduction, Notation, and Overview

-Applied Multivariate Analysis & Statistical Learning-

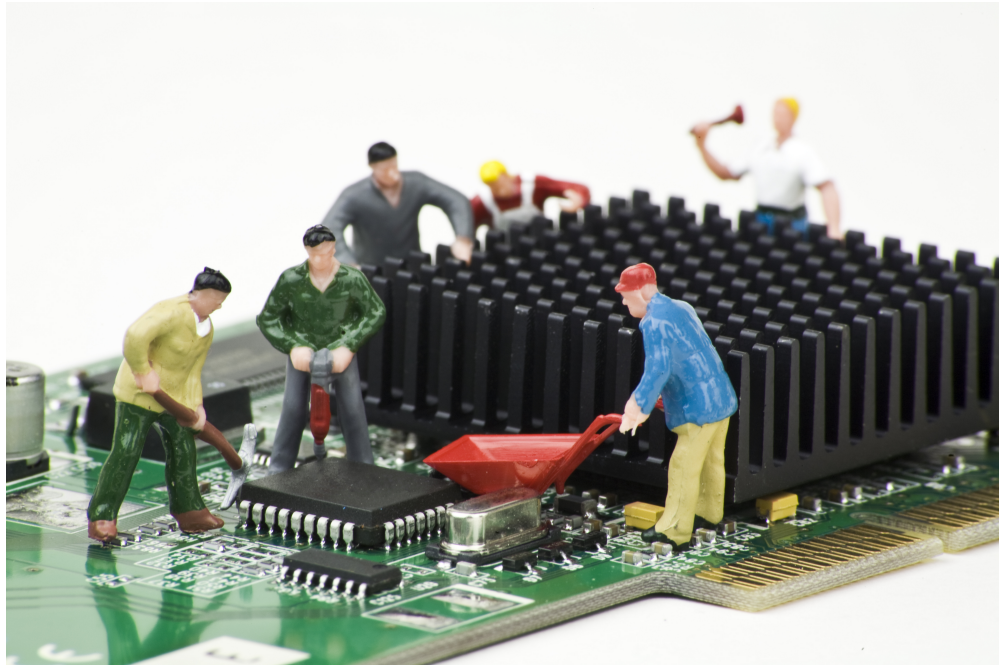
MMA chapters 2, 3 & ISL chapter 2.1

Lecturer: Darren Homrighausen, PhD

Preamble:

- Give an overview of the course
- Go over terminology and introduce notation
- Discuss some main themes for the class
- Cover the relevant topics in linear algebra & probability up to singular value decomposition (SVD)

Class Overview



Statistical learning is about using data to ...

- ... discover structure
- ... glean non-obvious insights into a problem
- ... make predictions about unknown quantities

Class Overview

Practically speaking, this means we seek to:

- find relationships in data that give good predictive performance
- reduce the size of the group of variables for scientific, statistical, or computational purposes

and, perhaps most importantly..

Know the techniques, how they work, when they apply, and how to implement them

Class Outline

Over the next semester we will cover

1. Unsupervised methods like clustering & dimension reduction
2. Supervised methods ranging from linear regression/classification to more advanced methods
3. Kernelization and dimension expansion
4. Algorithms for large data analysis

   (In particular, the fundamental notion of convex vs. non-convex optimization)

This course will emphasize methods over applications & theory

(This means we won't be proving results but we will be giving detailed motivations behind the methods and approaches discussed in this class)

References:

Main references:

- *Methods of Multivariate Analysis* (Rencher & Christensen)
- *An Introduction to Statistical Learning with Applications in R* (James, Witten, Hastie, & Tibshirani)

Secondary references:

- *The Elements of Statistical Learning* (Hastie, Tibshirani, & Friedman)
- Topic specific notes or lectures

(We will refer to these as MMA, ISL, and ESL, respectively)

General orientation to the field

Statistical learning

A harrowing aspect of this field: it's an academic chimera

Hence, statistical (machine) learning goes by many names:

- (business) analytics
- data science
- statistics
- data mining
- artificial intelligence
- others?

(Caveat: there are slight philosphical & application differences)

Statistical learning

It is sometimes expressed as an equation such as

$$\text{Statistical learning} = \text{Applied Statistics} + \text{Tech field}$$

or what it isn't

$$\text{Data science} \neq \text{Statistical Learning} + \text{Databases}$$

(Caveat: I don't necessarily believe either of these statements)

My training is in statistical (machine) learning hence I'll be talking mainly from that perspective

Terminology & notation

The set-up

We observe n pairs of data $(X_1^\top, Y_1)^\top, \ldots, (X_n^\top, Y_n)^\top$

Let[1] $Z_i^\top = (X_i^\top, Y_i) \in \mathbb{R}^p \times \mathbb{R}$

We'll refer to the training data as $\mathcal{D} = \{Z_1, \ldots, Z_n\}$

- Y_i is the supervisor or response
  (NOT DEPENDENT VARIABLE)
- $X_i \in \mathbb{R}^p$ is the feature or covariate (vector)
  (or explanatory variables or predictors. NOT INDEPENDENT VARIABLES)

Example: Y_i is whether a threat is detected in an image and the X_{ij} is the value at the j^{th} pixel of an image (p might be $1024^2 = 1048576$)

[1] These transposes get tiredsome. We'll get a bit sloppy and drop them selectively in what follows.

Introduction

Some common tasks we will encounter:

Regression: predict $Y \in \mathbb{R}$ from covariates or features X

Classification: predict $Y \in \{1, 2, \ldots, G\}$ from X

(Here, the labels or classes of Y are arbitrary)

Finding structure:

- Finding groups or clusters in the data
- Dimension reduction
- Independence relationships

Notation

We will concatenate the features into the design or feature matrix $\mathbb{X}$, and the supervisors into the supervisor vector $\mathbb{Y}$

$$\mathbb{X} = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix} = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p} \quad \text{and} \quad \mathbb{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$$

Commonly, a column $x_0^\top = \underbrace{(1, \ldots, 1)}_{n \text{ times}}$ is included

(This encodes an intercept term, with intercept parameter β_0)

In words: The features (columns) will be lower case letters and the observations (rows) will be upper case letters

We will refer to the entry in the i^{th} row, j^{th} column as X_{ij}

Notation summary

- We have data $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$
  (The training data)
- $X \in \mathbb{R}^p$ is a vector of measurements for each subject
  (Example: $X_i = [1, \text{income}_i, \text{education}_i]^\top$)
- $x \in \mathbb{R}^n$ is a vector of subjects for each measurement
  (Example: $x_j = [\text{income}_1, \text{income}_2, \ldots, \text{income}_n]^\top$)
- X_{ij} is the j^{th} measurement on the i^{th} subject
  (Example: $X_{ij} = \text{income}_i$)

ISL chapter 1. Note that MMA, ISL, and my notes have slightly different notation in terms of capitalization & indexes

Example: Biometrics

Example

Suppose we have 4 subjects in an experiment & we record

- BMI
- minutes spent exercising in the last 7 days

We want to predict each subject's resting heart rate

The classic linear model would model the regression function as

$$f_*(X) = \beta_0 + \beta^\top X = \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{exercise}$$

where

$$f(X) = \mathbb{E}[\text{resting heart rate}|X]$$
$$X = [\text{BMI}, \text{exercise}]$$

(Note: we could write $f_*(X) = \beta^\top X$ and $X = [1, \text{BMI}, \text{exercise}]$ instead)

ISL equation (2.1)

MMA chapter 3.1. The definition of expected value $\mathbb{E}$

Example

Under this model, the feature matrix and supervisor vector look like

$$\mathbb{X} = \begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 21 & 92 \\ 1 & 17 & 12 \\ 1 & 29 & 306 \\ 1 & 25 & 53 \end{bmatrix}}_{\text{int. BMI exercise}} \in \mathbb{R}^{4\times 3}$$

and

$$\mathbb{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_4 \end{bmatrix} = \begin{bmatrix} 72 \\ 47 \\ 82 \\ 64 \end{bmatrix} \in \mathbb{R}^4$$

Example

Adding a quadratic polynomial transformation

$$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \sum_{j \le j'}^{p} x_j x_{j'} \beta_{jj'}$$
$$= \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{exercise} + \beta_{11} \text{BMI}^2 + \beta_{22} \text{exercise}^2$$
$$+ \beta_{12} \text{BMIexercise}$$

Under this model, the feature matrix looks like

$$\mathbb{X} = \begin{bmatrix} x_0 & x_1 & \cdots & x_5 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 21 & 92 & 21^2 & 92^2 & 21 * 92 \\ 1 & 17 & 12 & 17^2 & 12^2 & 17 * 12 \\ 1 & 29 & 306 & 29^2 & 306^2 & 29 * 306 \\ 1 & 25 & 53 & 25^2 & 53^2 & 25 * 53 \end{bmatrix}}_{\text{int.} \quad \text{BMI} \quad \text{exercise} \quad \text{BMI}^2 \quad \text{exercise}^2 \quad \text{BMI}*\text{exercise}}$$

($\mathbb{Y}$ is the same)

Main themes

Some main themes: Assumptions

What assumptions are needed to motivate the method or guarantee some property?

Example: Suppose I observe some data $Y_1, \ldots, Y_n \in \mathbb{R}$

I want to make a prediction about a new observation Y_{n+1}

I could use $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

I'm (implicitly?) assuming that the ...

- ...expectations are all (nearly) the same:

$$\mathbb{E}Y_1 \approx \mathbb{E}Y_2 \approx \cdots \approx \mathbb{E}Y_n$$

- ...observations have covariance (nearly) equal to zero
- ...observations all have (nearly) the same variance
- ...probability of Y_i being far from $\mathbb{E}Y_i$ is small

Some main themes: Convexity

Convex problems can be solved efficiently. If necessary, we try to approximate nonconvex problems with convex ones

- Convex set: A set B is convex if for any $\beta, \beta' \in B$ and any $\tau \in [0, 1]$

$$\tau\beta + (1 - \tau)\beta' \in B$$

- Convex function: A function ℓ is convex if the "area" above the function is a convex set

  (This area is formally known as the epigraph of ℓ)

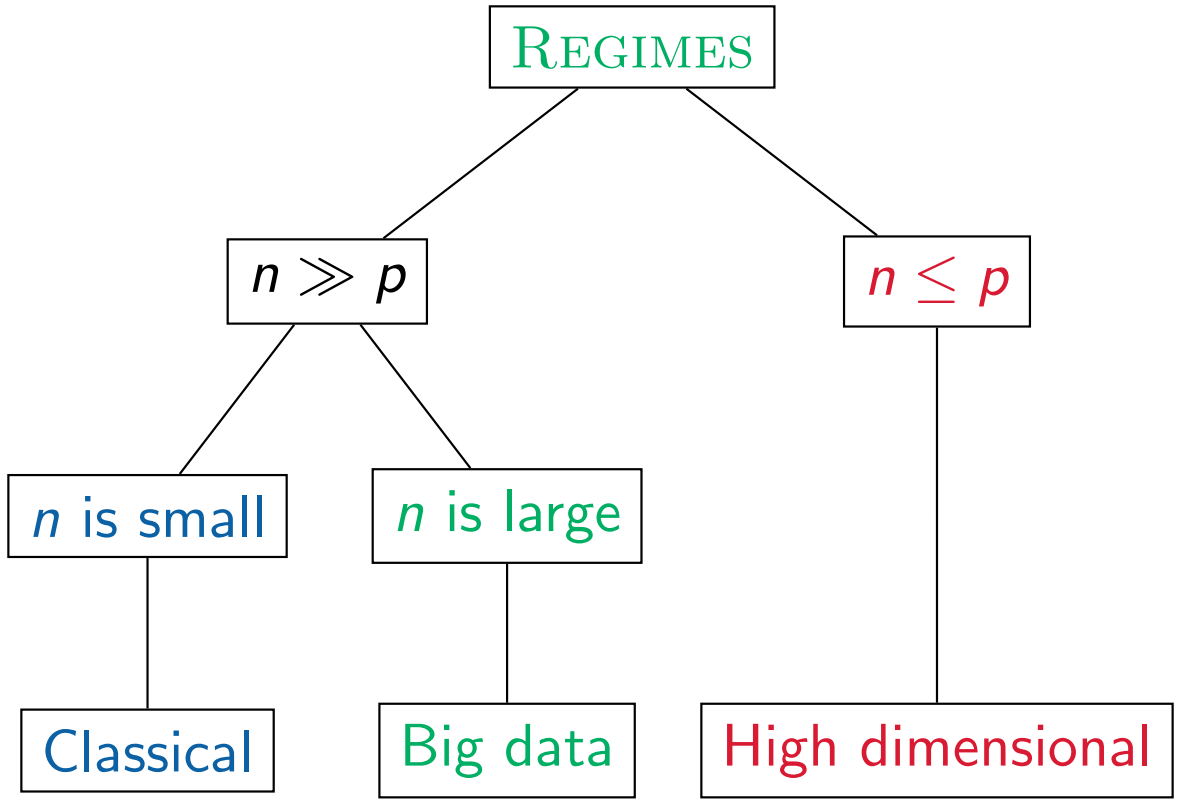
  This is the same as: for any $\beta, \beta' \in \mathbb{R}^p$ and any $\tau \in [0, 1]$

$$f(\tau\beta + (1 - \tau)\beta') \leq \tau f(\beta) + (1 - \tau) f(\beta')$$

  (Convex functions cannot have multiple local minima)

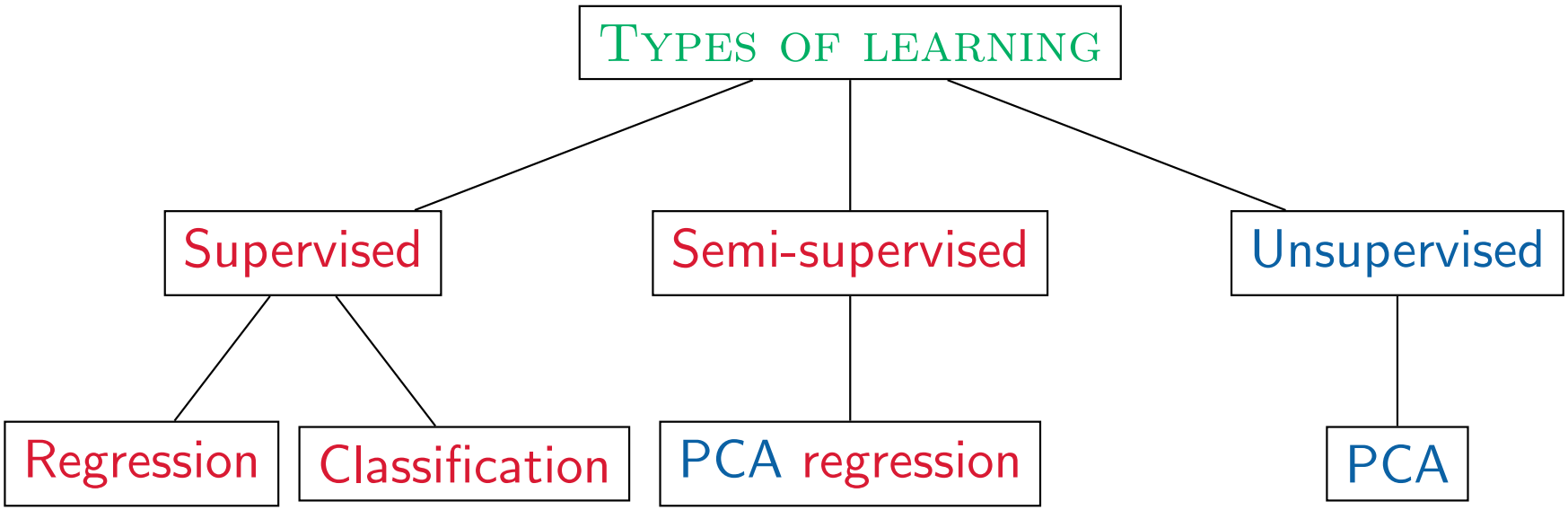
Some main themes: Regimes

There are roughly three regimes of interest, assuming $\mathbb{X} \in \mathbb{R}^{n \times p}$



(We will return to these in more detail)

Some main themes: Types of learning



Some comments:

- Much more heuristic, unclear what a good solution would be
- Comparing predictions to Y gives a natural notion of prediction accuracy

Linear algebra and probability background

Background

- We will write vectors as

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

  We write this as $z \in \mathbb{R}^n$, which is "z is a member of ar-en."
- We commonly will need to "turn" the vector, which we write as

$$z^\top = \begin{bmatrix} z_1 & z_2 & \dots & z_n \end{bmatrix}$$

MMA Section 2.2.1

Necessary Background: Addition and Multiplication

We will need to extend the ideas of addition and multiplication of numbers to higher dimensional objects (vectors and matrices)

- Suppose $u, v \in \mathbb{R}^q$. Then we write u "times" v as

$$u^\top v = \sum_{i=1}^{q} u_i v_i$$

- This is known as the inner product of u and v

MMA Sections 2.2.3, 2.3.1, 2.3.3

Necessary Background: Lengths

We will need to measure the size of both vectors and matrices.

The most common is the one we use every day Euclidean distance

(Think: the Pythagorean theorem)

$$||x||_2 = \sqrt{\sum_{k=1}^{p} x_k^2} = \sqrt{x^\top x}$$

We call this a norm and refer to this as the "ell two norm"

(Often, it will be written as squared for convenience: $||x||_2^2 = x^\top x$)

MMA equation 2.35

Necessary Background: Lengths

Additionally, we will need the Manhattan distance

$$||x||_1 = \sum_{k=1}^{p} |x_k|$$

We call this the "ell one norm"

Necessary Background: Addition and Multiplication

- Also, for matrices $\mathbb{A} \in \mathbb{R}^{n \times p}, \mathbb{B} \in \mathbb{R}^{p \times r}$,

  (We will refer to the entry in the i^{th} row, j^{th} column of a matrix $\mathbb{A}$ as A_{ij})

$$\mathbb{AB} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ A_{21} & A_{22} & \dots & A_{2p} \\ & & \vdots & \\ A_{n1} & A_{n2} & \dots & A_{np} \end{bmatrix} \cdot \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1r} \\ B_{21} & B_{22} & \dots & B_{2r} \\ & & \vdots & \\ B_{p1} & B_{n2} & \dots & B_{pr} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{p} A_{1j}B_{j1} & \sum_{j=1}^{p} A_{1j}B_{j2} & \dots & \sum_{j=1}^{p} A_{1j}B_{jr} \\ \sum_{j=1}^{p} A_{2j}B_{j1} & \sum_{j=1}^{p} A_{2j}B_{j2} & \dots & \sum_{j=1}^{p} A_{2j}B_{jr} \\ & & \vdots & \\ \sum_{j=1}^{p} A_{nj}B_{j1} & \sum_{j=1}^{p} A_{nj}B_{j2} & \dots & \sum_{j=1}^{p} A_{nj}B_{jr} \end{bmatrix} \in \mathbb{R}^{n \times r}$$

MMA section 2.3.3

Necessary Background: Lengths

For matrices, we will just define something very related to 'length' (but it doesn't technically qualify)

Many times, we are interested in the size of the diagonal of a matrix

This is known as the trace. For matrix $\mathbb{A} \in \mathbb{R}^{p \times p}$

$$\text{trace}(\mathbb{A}) = \sum_{j=1}^{p} A_{jj}$$

That is, the trace is the sum of the diagonal entries.

MMA section 2.9

Singular Value Decomposition (SVD)

SVD

Many, many topics in statistics can be computed/interpretted via the singular value decomposition (SVD)

The SVD is a generalization of the eigenvector decomposition

Instead of

$$\mathbb{A} = UDU^\top \longleftarrow \text{eigenvector decomposition}$$

we get

$$\mathbb{A} = UDV^\top \longleftarrow \text{singular value decomposition}$$

This change makes the (unique) SVD always exist

MMA section 2.11.10

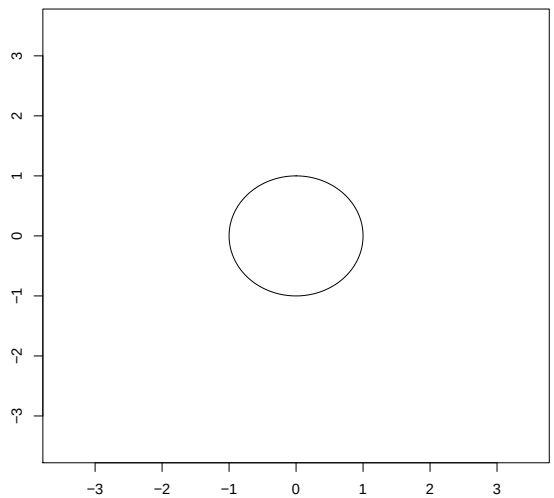
SVD

It turns out we can think of matrix multiplication in terms of circles and ellipsoids

Take a matrix $\mathbb{X}$ and let's look at the set of vectors

$$B = \{\beta : ||\beta||_2 \leq 1\}$$

This is a circle!

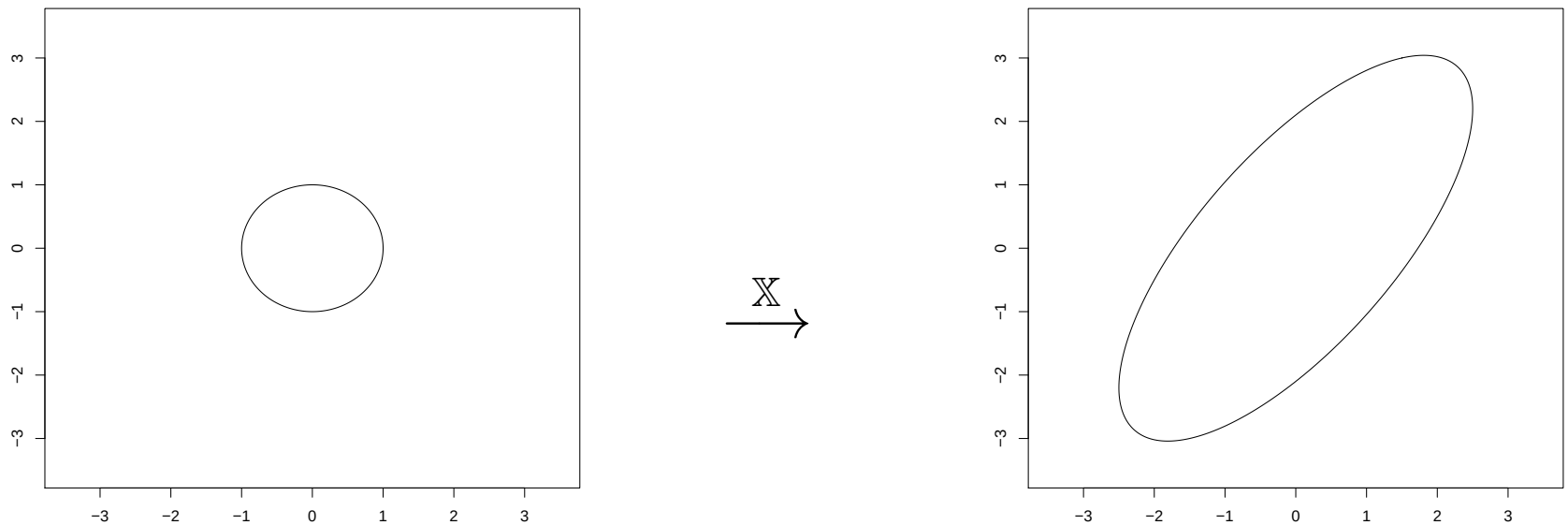


SVD

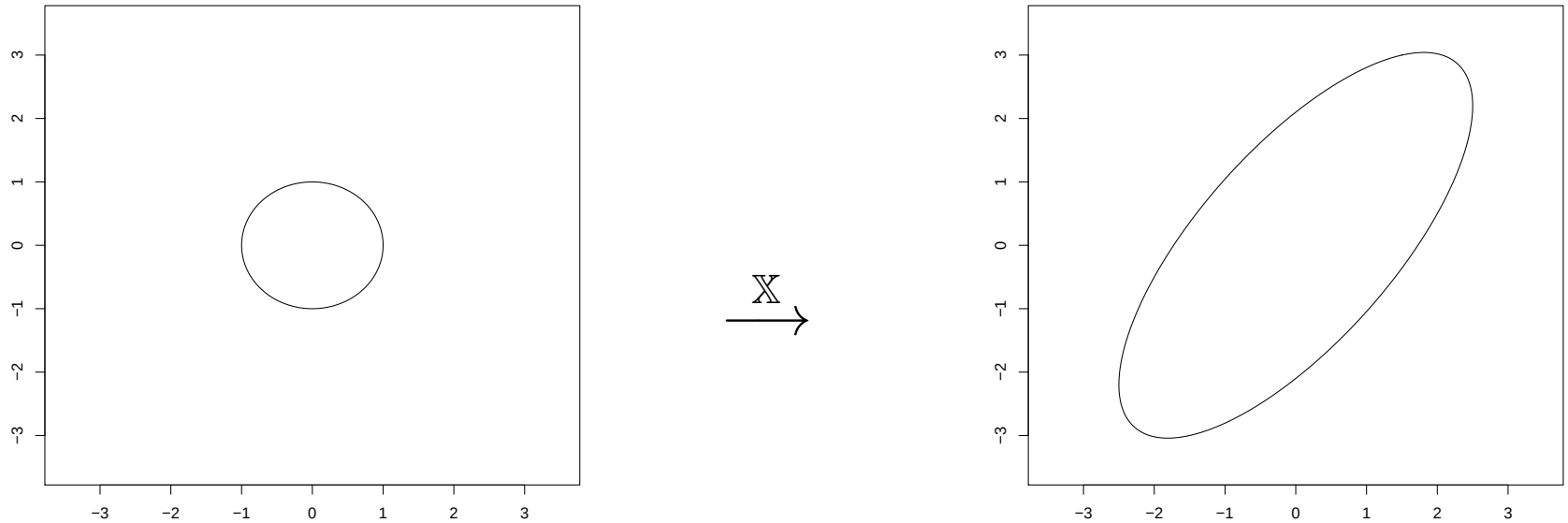
What happens when we multiply vectors in this circle by $\mathbb{X}$?

Let

$$\mathbb{X} = \begin{bmatrix} 2.0 & 0.5 \\ 1.5 & 3.0 \end{bmatrix} \text{ and } \mathbb{X}\beta = \begin{bmatrix} 2\beta_1 + 0.5\beta_2 \\ 1.5\beta_1 + 3\beta_2 \end{bmatrix}$$



SVD

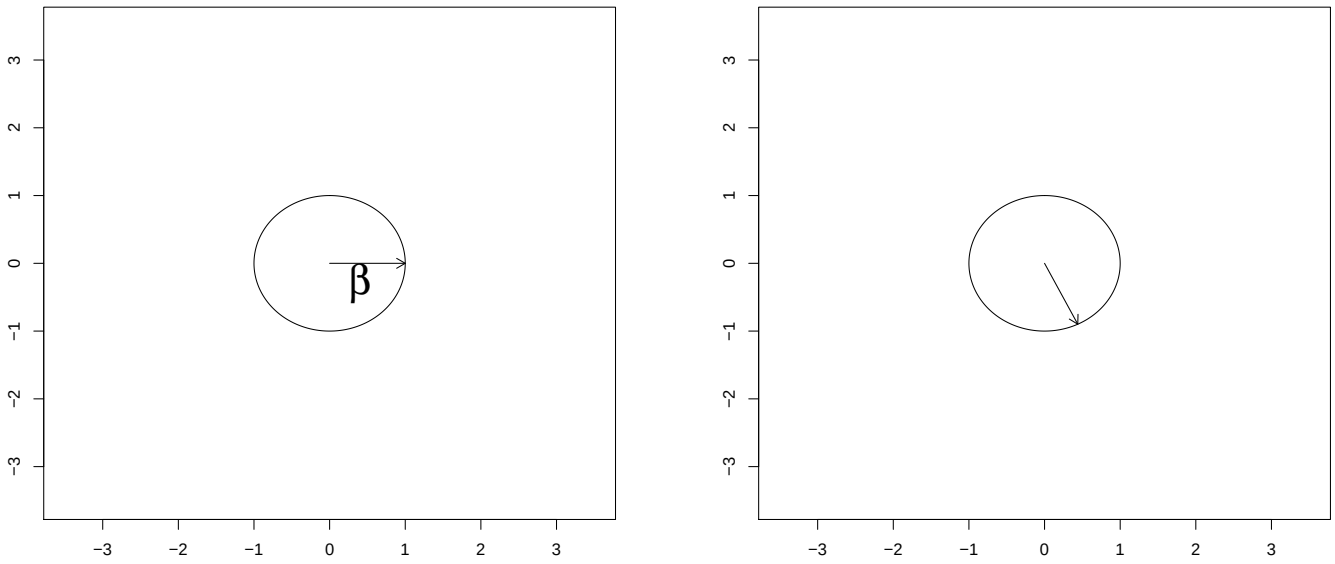


What happened?

1. The coodinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

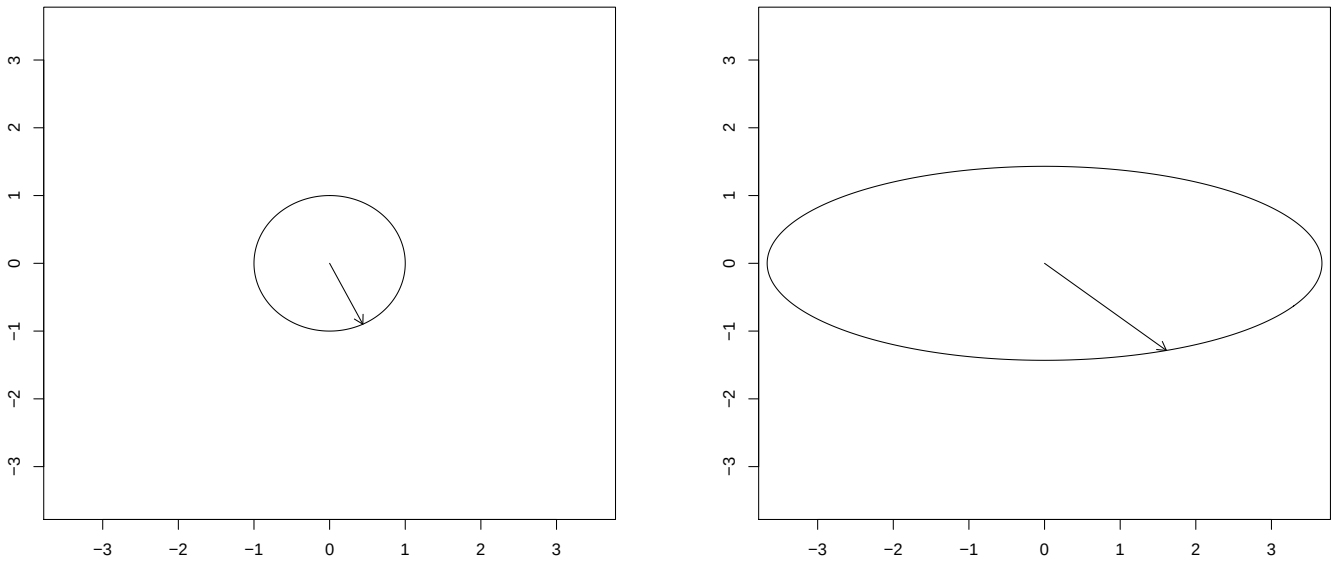
Let's break this down into parts...

SVD



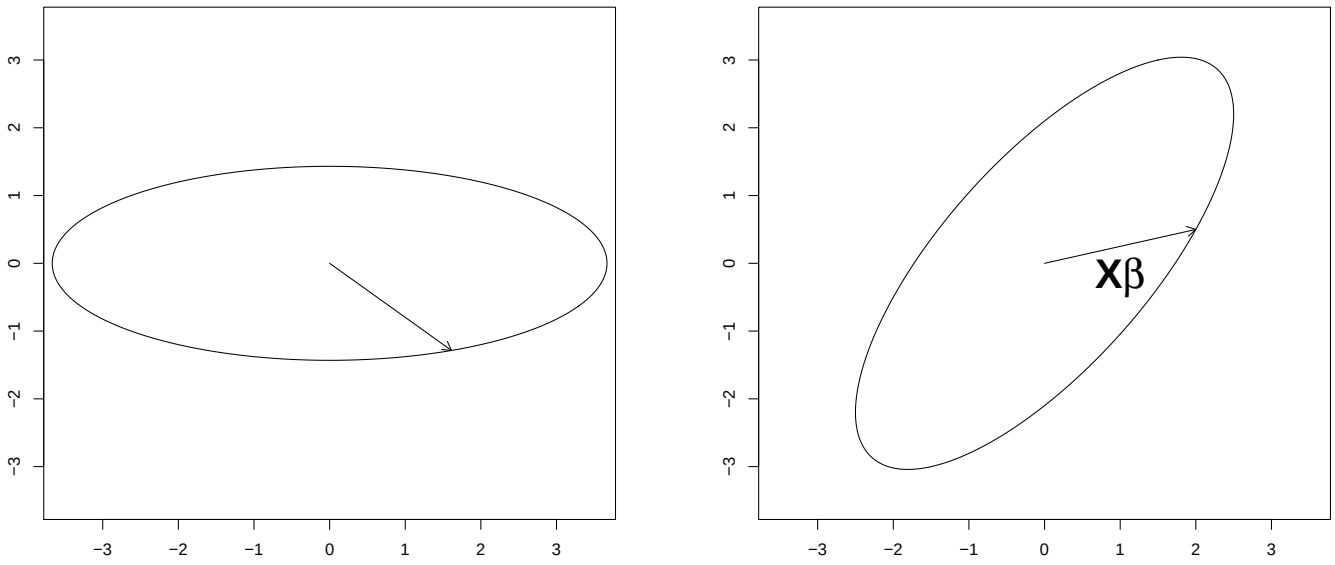
1. The coordinate axis gets rotated

SVD



1. The coordinate axis gets rotated
2. The new axis gets elongated (making an ellipse)

SVD



1. The coordinate axis gets rotated
2. The new axis gets elongated (making an ellipse)
3. This ellipse gets rotated

Necessary Background: Rotation

Rotations: These can be thought of as just reparameterizing the coordinate axis. This means that they don't change the geometry.

As the original axis was orthogonal (that is; perpendicular), the new axis must be as well.

Necessary Background: Rotation

Let v_1, v_2 be two normalized, orthogonal vectors. This means that:

$$v_1^\top v_2 = 0 \quad \text{and} \quad v_1^\top v_1 = v_2^\top v_2 = 1$$

In matrix notation, if we create V as a matrix with normalized, orthogonal vectors as columns, then:

$$V^\top V = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = I$$

Here, I is the identity matrix.

MMA section 2.10

Necessary Background: Elongation

Elongation: These can be thought of as stretching vectors along the current coordinate axis. This means that they **do** change the geometry by distorting distances.

Elongations are the result of multiplication by a diagonal matrix (note: we just saw a very special case of such a matrix: the identity matrix I)

All diagonal matrices have the form:

$$D = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & d_p \end{bmatrix}$$

MMA section 2.2.4

SVD

Using this intuition, for any matrix $\mathbb{X}$ it is possible to write its SVD:

$$\mathbb{X} = UDV^\top$$

where

- U and V are orthogonal (think: rotations)
- D is diagonal (think: elongation)
- The diagonal elements of D are ordered as

$$d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$$

Many properties of matrices can be 'read off' from the SVD.

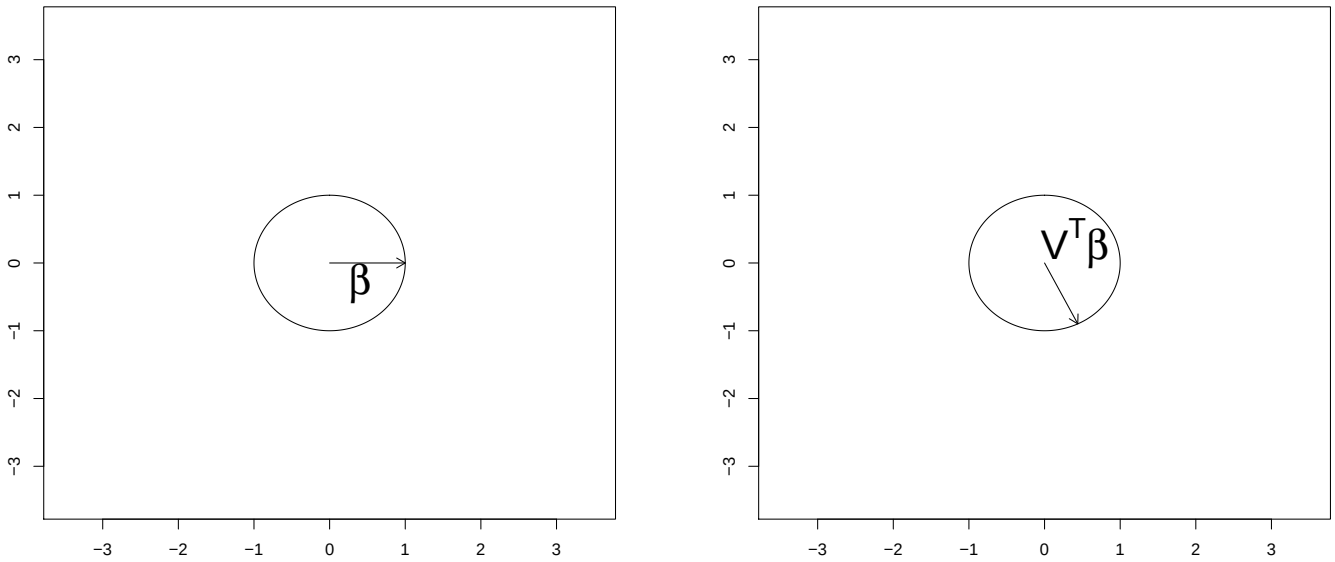
SVD

EXAMPLE: The rank of a matrix answers the question: how many dimensions does the ellipse live in? In other words, it is the number of columns of the matrix $\mathbb{X}$, not counting the columns that are 'redundant'

It turns out the rank is exactly the quantity q such that $d_q > 0$ and $d_{q+1} = 0$

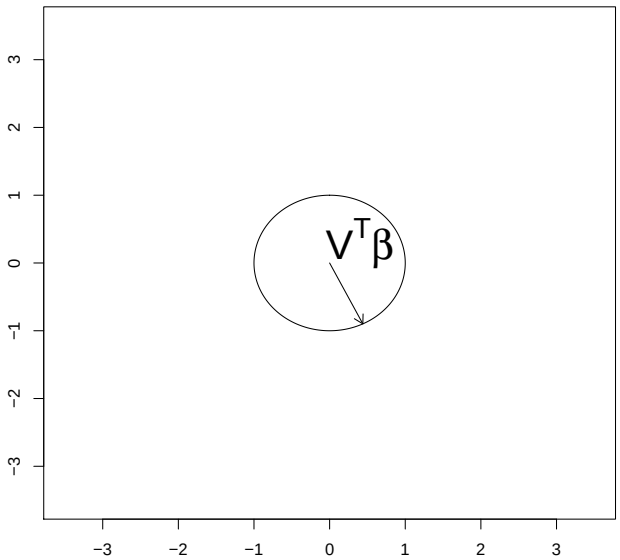
MMA section 2.2.5

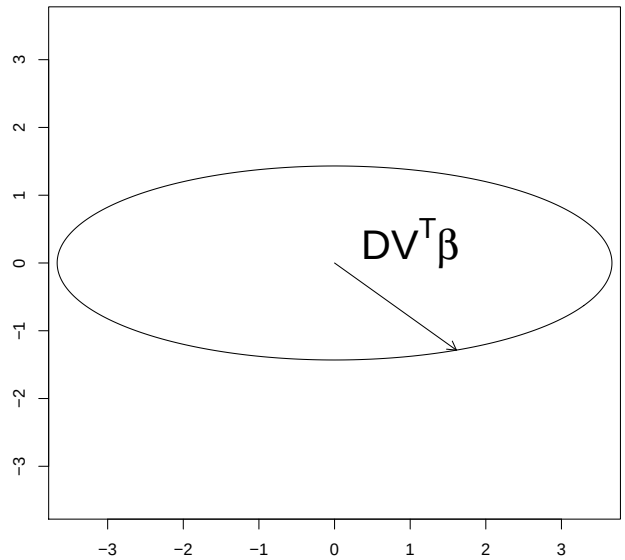
SVD: Recap



1. The coordinate axis gets rotated (Multiplication by $V^\top$)

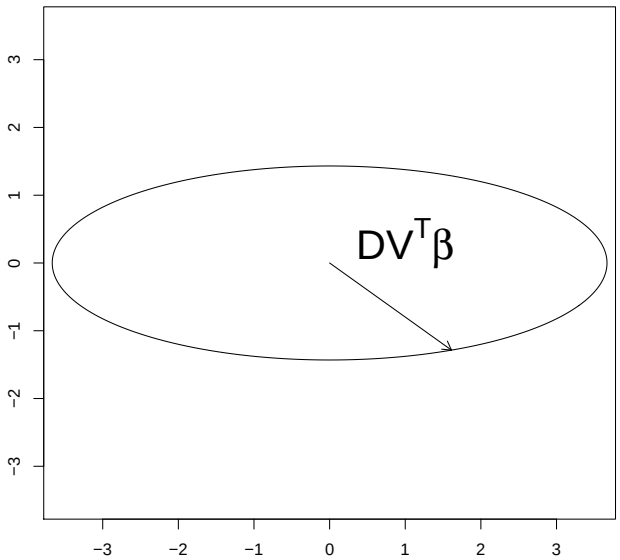
SVD: Recap

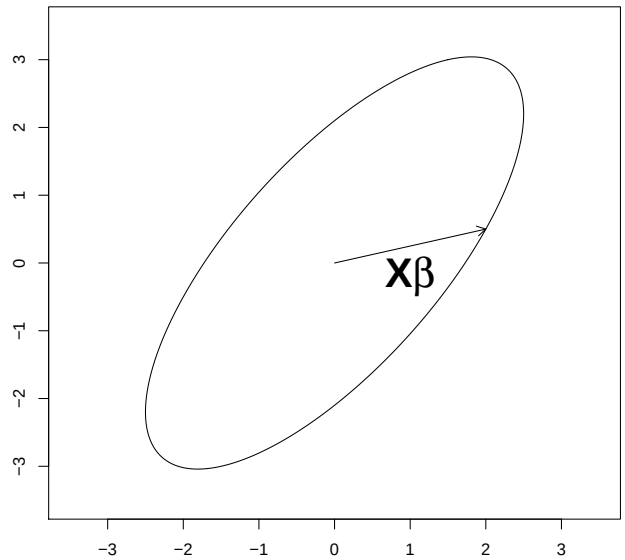




1. The coordinate axis gets rotated (Multiplication by $V^\top$)
1. The new axis gets elongated (Multiplication by D)

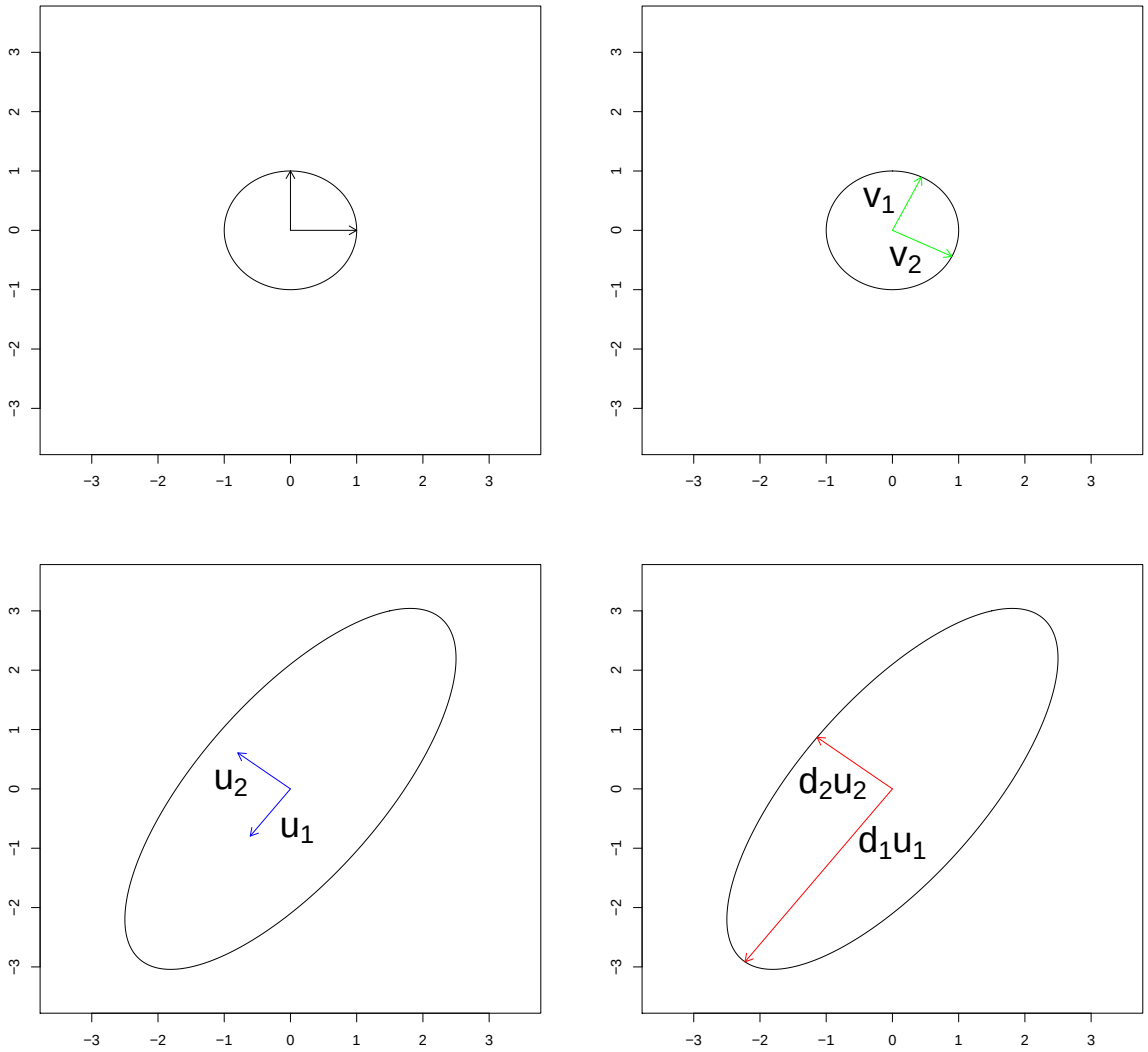
SVD: Recap





1. The coordinate axis gets rotated (Multiplication by $V^\top$)
1. The new axis gets elongated (Multiplication by D)
2. This ellipse gets rotated (Multiplication by U)

SVD [one last time]



Summary:

Of all the possible axes of the original circle, the one given by v_1, v_2 has the unique property:

$$\mathbb{X} v_j = d_j u_j$$

for all j.

Lastly:

$$\mathbb{X} = \sum_j d_j u_j v_j^\top$$

Computing the SVD: svd

The SVD can be computed in R in two ways:

The svd function is a base R function that directly uses the underlying LAPACK code

Note that in help(svd) you'll see

Usage

```
svd(x, nu = min(n, p), nv = min(n, p), LINPACK = FALSE)
```

LINPACK is a older version of LAPACK that isn't optimized to use modern 'cache-based' architectures

(Note that the only two choices for nu or nv are the default or 0)

Computing the SVD: irlba

Alternatively, there is irlba

(Needs to be installed/loaded into memory to use)

It uses an iterative method, but usually converges to within machine precision to the svd results

We will discuss irlba more when we get to Principal Components Analysis

(An interesting bit of history: one of the key parts of the BellKor team's $1,000,000 netflix prize was using irlba to compute the SVD for matrix completion)

Probability

What's a random variable?

Let X be a random variable. That is, X...

- Has a probability density function p_X such that the probability (denote this by $\mathbb{P}$) that X takes on a set of values A is given by[2]

$$\mathbb{P}(A) = \int_A p_X(x)dx$$

- And p_X has certain properties such as $p_X \geq 0$ and $\int p_X = 1$.

[2] Anyone who has studied probability would take issue with this statement. If this is you, don't quibble; we're trying to avoid unnecessary complications.

What are the properties of a random variable?

In this class, we really only care about X's

- mean (alternatively known as its expectation)

  (This is all about finding its center)

- and variance/covariance

  (This is all about finding its spread and orientation)

What's expectation?

Imagine taking a metal rod of a certain mass.

However, its mass isn't necessarily even along its length.

Attempt to balance the rod on your finger. The balancing point is the center of mass of the rod.

Figure: A family calculates expectations

What's expectation?

Crucial connection: If we think about the density of the random variable determining where the rod's mass is distributed, then the "center of mass" is the expectation.

$$\mathbb{E}[X] = \int x\, p_X(x) dx$$

What's variance?

Variance is defined as

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sigma^2$$

and covariance is defined as

$$\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

In words:

"variance is the average squared deviation from the average"

"covariance is the average amount each r.v. spends above/below its mean relative to another r.v."

Note:

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Can you take me higher?

Implicitly, we were assuming that $X \in \mathbb{R}$.

What happens if $X \in \mathbb{R}^p$?

The expectation is going to look the same, but be a vector

$$\mathbb{E}[X] \in \mathbb{R}^p$$

For variance, we need to use some matrix notation:

$$\mathbb{V}[X] = \mathrm{Cov}[X] = \Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \in \mathbb{R}^{p \times p}$$

If you write this out, you'll see that this a matrix with

- The variances of the components on the diagonal
- The covariances of any two components in the off-diagonal entries.

MMA chapter 3.6 & 3.7

Covariance

Covariance between, e.g., X and Y, takes on units of the product of the two measurements

Example: If C is temperature in Celsius and F is temperature in Fahrenheit. Then, taken as random variables, σ_{CF} has units C$*$F

```
C = c(-10,0,10,20)
F = C*9/5. + 32
sigmaHat_CF = cov(cbind(C,F))
> sigmaHat_CF
          C   F
C 166.6667 300
F 300.0000 540
> sum((C - mean(C))*(F- mean(F)))/3.
[1] 300
```

This makes the comparison difficult as what does a covariance of 300 really mean?

Correlation

We can standardize covariance into a scale-free version

$$\rho_{CF} = \frac{\sigma_{CF}}{\sigma_C \sigma_F}$$

(This the (Pearson's) correlation, there are others (e.g. Spearmans)

In the preceding example, ρ_{CF} can be estimated by

```
r_CF = cor(cbind(C,F))
> r_CF
  C F
C 1 1
F 1 1
```

(Hmm.. what does this indicate? What is the nature of the SVD and rank?)

If Σ is the covariance matrix, then

$$\rho = \text{diag}(\Sigma)^{-1/2} \, \Sigma \, \text{diag}(\Sigma)^{-1/2}$$

Combine matrices and probability

We will combine matrix multiplication with probability statements

Suppose that $Y \in \mathbb{R}^n$ is a random variable such that $\mathbb{E}[Y] = \mu$ and $\mathbb{V}[Y] = \Sigma$.

What is the distribution of $\mathbb{A}Y$?

It turns out expectation is linear and hence we can rearrange '$\mathbb{E}$' and '$\mathbb{A}$'

$$\mathbb{E}[\mathbb{A}Y] = \mathbb{A}\mathbb{E}[Y] = \mathbb{A}\mu$$

Variance is a little more complicated, but not much

$$\mathbb{V}[\mathbb{A}Y] = \mathbb{A}\mathbb{V}[Y]\mathbb{A}^\top = \mathbb{A}\Sigma\mathbb{A}^\top$$

(This is the multivariate analogue of saying $\mathbb{V}[aY] = a^2\mathbb{V}[Y]$ when $Y \in \mathbb{R}$)

MMA chapter 3.10.2

Combine matrices and probability

Suppose we want to simulate a random vector with a given mean/covariance structure:

$$Y \sim (\mu, \Sigma)$$

but we can only generate "standardized" versions:

$$Y \sim (0, I_n)$$

(I_n is the $n \times n$ identity)

The univariate version is

$$Y \leftarrow \sigma Y + \mu$$

Hence, we need the square root of Σ

Combine matrices and probability

The square root of a matrix can be found via the Cholesky decomposition

$$Y \leftarrow \Sigma^{1/2} Y + \mu$$

We can compute this in R via the function chol

Note that we can also use our friend the SVD to compute $\Sigma^{1/2}$...

(See HW1)

MMA chapter 2.7 & 2.11.8

Postamble:

- Give an overview of the course

  (What is statistical learning?)
- Go over terminology and introduce notation

  (Define feature, supervisor, training data and give some examples)
- Discuss some main themes for the class

  (Assumptions; convexity; define "classical", "big data", and "high dimensional" types of problems; and define unsupervised and supervised learning)
- Cover the relevant topics in linear algebra & probability up to singular value decomposition (SVD)

  (SVD is the natural coordinate system for multiplication by a matrix and is incredibly useful for understanding/computing many methods)