Lecture 2

Intro to dimension reduction & review of linear algebra

COURSE MATH 637: Mathematical Techniques in Data Science

WEBSITE https://jacobcalvert.com/teaching/math637fall2026/

INSTRUCTOR Jacob Calvert (calvert@udel.edu)

OFFICE HOURS MW 2:45pm–3:45pm, Ewing Hall 509

Announcements & Reminders

- For further review of linear algebra, see notes on Canvas
- Monday: principal component analysis.

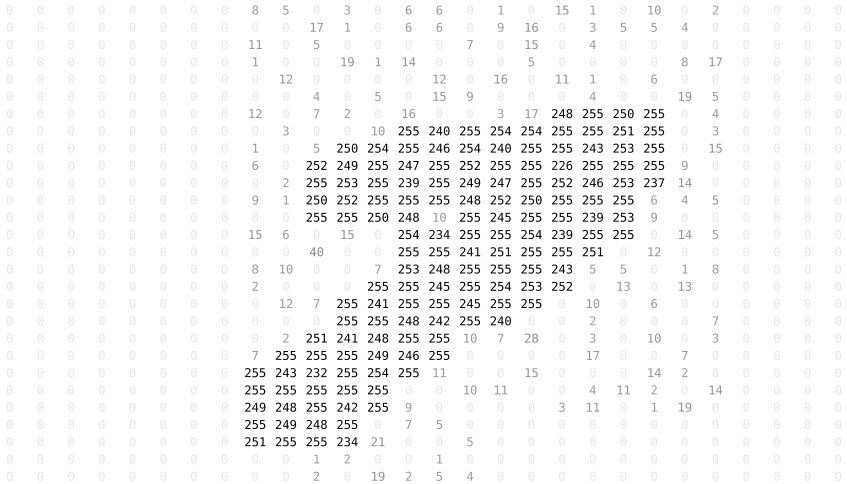
Plan for today

• Intro to dimension reduction	11 min
• Behold, the data matrix	10 min
• Outer products everywhere	16 min
• Spectral theorem	9 min
• Square matrices from the data matrix	4 min

Example: The MNIST data

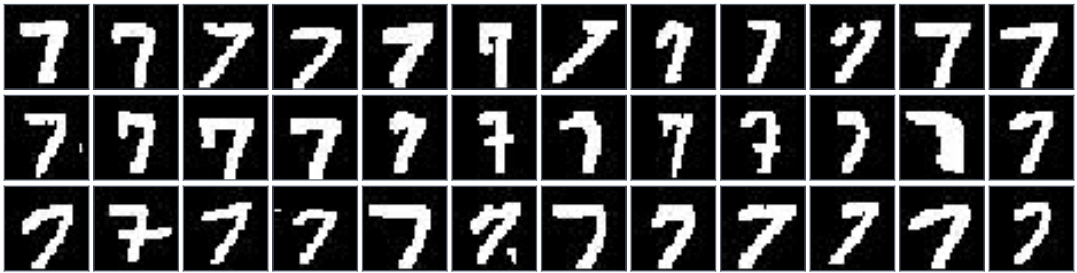
- $70{,}000$ images of handwritten digits
- Each image $= 28 \times 28$ pixels
- Each pixel $=$ integer in $\{0, \dots, 255\}$

Example: A seven, up close



Example: A seven, far away

Example: Many sevens



Example: The MNIST data

- $70{,}000$ images of handwritten digits
- Each image $= 28 \times 28$ pixels
- Each pixel $=$ integer in $\{0, \dots, 255\}$
- Pixels usually concatenated to make vector of length 784
- So each image is an element of $\{0, \dots, 255\}^{784}$

Example: The MNIST data

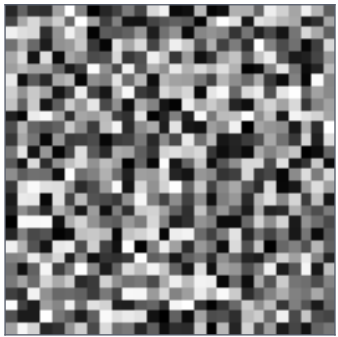
- $70{,}000$ images of handwritten digits
- Each image $= 28 \times 28$ pixels
- Each pixel $=$ integer in $\{0, \dots, 255\}$
- Pixels usually concatenated to make vector of length 784
- So each image is an element of $\{0, \dots, 255\}^{784}$

Question

What $\%$ of possible digit data does the MNIST dataset represent?

A digit is an atypical element of $\{0, \dots, 255\}^{784}$

Left: A digit.



Middle: Uniformly random.

Right: Permuted.

Do you see what this suggests?

- $256^{784} \approx 10^{1888}$ points in that space
- MNIST has $< 10^5$ of them
- Uniformly random sampling of pixels will *never* produce a digit

Do you see what this suggests?

- $256^{784} \approx 10^{1888}$ points in that space
- MNIST has $< 10^5$ of them
- Uniformly random sampling of pixels will *never* produce a digit

> **Idea:** The images inhabit $\mathbb{R}^{784}$, but they effectively lie on a much lower-dimensional surface inside it.

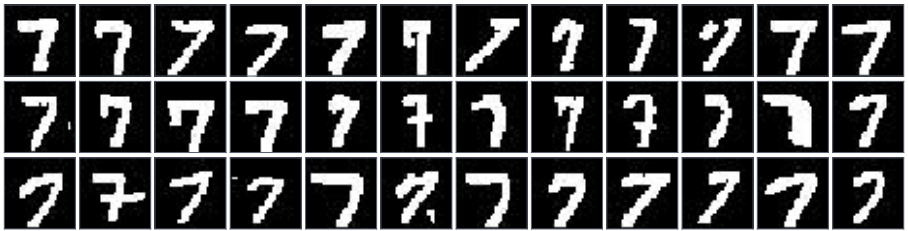
Do you see what this suggests?

- $256^{784} \approx 10^{1888}$ points in that space
- MNIST has $< 10^5$ of them
- Uniformly random sampling of pixels will *never* produce a digit

Idea: The images inhabit $\mathbb{R}^{784}$, but they effectively lie on a much lower-dimensional surface inside it.

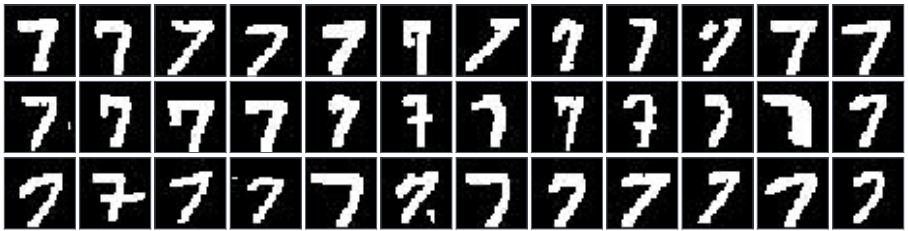
In the literature: Such a surface is called a *manifold*. The claim that real data sets lie near one is the manifold hypothesis, and finding the manifold is manifold learning.

Example: If not a dimension of 784, then what?



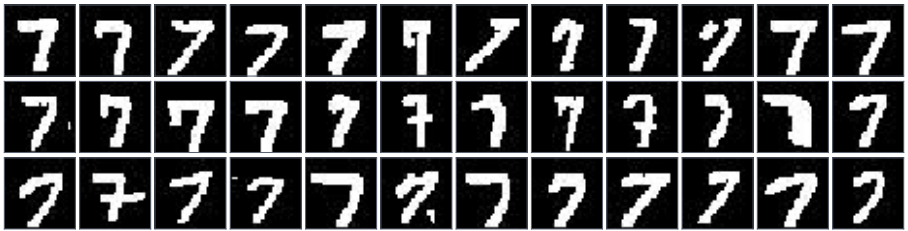
- Digits differ in relatively few "ways"

Example: If not a dimension of 784, then what?



- Digits differ in relatively few "ways"
- Intuitively, this number of ways is related to the intrinsic dimension of the data

Example: If not a dimension of 784, then what?



- Digits differ in relatively few "ways"
- Intuitively, this number of ways is related to the intrinsic dimension of the data
- The intrinsic dimension can be estimated (!) to be $\ll 784$

Guess the intrinsic dimension of MNIST

Guess the intrinsic dimension of MNIST

In the literature (Pope et al., ICLR 2021): Estimates put it between 12 and 14: roughly a dozen continuous qualities.

Example: Dimension reduction methods on MNIST

Adapted from Fig. 4 of McInnes et al. (2020)

- Each maps an image in $\mathbb{R}^{784}$ to a point in $\mathbb{R}^2$, without ever seeing a label
- Color by the true digit *afterwards*: the clusters are the digits
- Principal component analysis is left-most

Plan for today

• Intro to dimension reduction	11 min
• Behold, the data matrix	10 min
• Outer products everywhere	16 min
• Spectral theorem	9 min
• Square matrices from the data matrix	4 min

Convention: vectors are columns

- vectors $x, y \in \mathbb{R}^p$ are $p \times 1$ arrays

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \qquad x^\mathsf{T} y = \sum_i x_i y_i, \qquad \|x\|^2 = x^\mathsf{T} x = \sum_i x_i^2.$$

Convention: vectors are columns

- vectors $x, y \in \mathbb{R}^p$ are $p \times 1$ arrays

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \qquad x^\mathsf{T} y = \sum_i x_i y_i, \qquad \|x\|^2 = x^\mathsf{T} x = \sum_i x_i^2.$$

- $\langle x, y \rangle$ also denotes $x^\mathsf{T} y$
- $x \perp y$ means $x^\mathsf{T} y = 0$

Convention: vectors are columns

- vectors $x, y \in \mathbb{R}^p$ are $p \times 1$ arrays

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \qquad x^\mathsf{T} y = \sum_i x_i y_i, \qquad \|x\|^2 = x^\mathsf{T} x = \sum_i x_i^2.$$

- $\langle x, y \rangle$ also denotes $x^\mathsf{T} y$
- $x \perp y$ means $x^\mathsf{T} y = 0$

> **Note:** When in doubt, check that the dimensions make sense. E.g., $x^\mathsf{T} y$ is $(1 \times p)(p \times 1) =$ a scalar.

Convention: Rows of data matrix $=$ examples

$$X = \begin{bmatrix} 5 & 5 \\ 1 & 1 \\ 4 & 2 \\ 2 & 4 \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$n = 4$ rows, $p = 2$ columns

- One row per example/observation
- One column per feature/predictor
- X is rarely square!

Convention: Rows of data matrix $=$ examples

$$X = \begin{bmatrix} 5 & 5 \\ 1 & 1 \\ 4 & 2 \\ 2 & 4 \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$n = 4$ rows, $p = 2$ columns

- One row per example/observation
- One column per feature/predictor
- X is rarely square!

Note: Vectors are columns, but stacked into a data matrix they lie along the rows: row i of X is $x_i^\intercal$. So $x_1 = (5, 5)^\intercal$, and X has $x_1^\intercal$ across its top.

Convention: Rows of data matrix $=$ examples

$$X = \begin{bmatrix} — & x_1^\mathsf{T} & — \\ — & x_2^\mathsf{T} & — \\ & \vdots & \\ — & x_n^\mathsf{T} & — \end{bmatrix} \in \mathbb{R}^{n \times p}$$

n rows, p columns

- One row per example/observation
- One column per feature/predictor
- X is rarely square!

Note: Vectors are columns, but stacked into a data matrix they lie along the rows: row i of X is x_i^T. So $x_1 = (5, 5)^\mathsf{T}$, and X has x_1^T across its top.

Definition: Dimension reduction $=$ approximate X

Fix $k < p$. Given $x_1, \dots, x_n \in \mathbb{R}^p$, and classes $\mathcal{F}$ of maps $\mathbb{R}^p \to \mathbb{R}^k$ and $\mathcal{G}$ of maps $\mathbb{R}^k \to \mathbb{R}^p$:

$$\min_{f \in \mathcal{F},\, g \in \mathcal{G}} \quad \sum_{i=1}^{n} \| x_i - g(f(x_i)) \|^2.$$

Definition: Dimension reduction $=$ approximate X

Fix $k < p$. Given $x_1, \dots, x_n \in \mathbb{R}^p$, and classes $\mathcal{F}$ of maps $\mathbb{R}^p \to \mathbb{R}^k$ and $\mathcal{G}$ of maps $\mathbb{R}^k \to \mathbb{R}^p$:

$$\min_{f \in \mathcal{F},\, g \in \mathcal{G}} \sum_{i=1}^{n} \| x_i - g(f(x_i)) \|^2.$$

- f is the encoder: the *code* $f(x_i) \in \mathbb{R}^k$ has $k < p$ numbers
- g is the decoder: $g(f(x_i)) \in \mathbb{R}^p$ is the *reconstruction*

Definition: Dimension reduction $=$ approximate X

Fix $k < p$. Given $x_1, \dots, x_n \in \mathbb{R}^p$, and classes $\mathcal{F}$ of maps $\mathbb{R}^p \to \mathbb{R}^k$ and $\mathcal{G}$ of maps $\mathbb{R}^k \to \mathbb{R}^p$:

$$\min_{f \in \mathcal{F},\, g \in \mathcal{G}} \sum_{i=1}^{n} \| x_i - g(f(x_i)) \|^2.$$

- f is the encoder: the *code* $f(x_i) \in \mathbb{R}^k$ has $k < p$ numbers
- g is the decoder: $g(f(x_i)) \in \mathbb{R}^p$ is the *reconstruction*

> **Note:** Different choices of $\mathcal{F}$ and $\mathcal{G}$ lead to different dimension reduction methods!

Definition: Dimension reduction = approximate X

Fix $k < p$. Given $x_1, \dots, x_n \in \mathbb{R}^p$, and classes $\mathcal{F}$ of maps $\mathbb{R}^p \to \mathbb{R}^k$ and $\mathcal{G}$ of maps $\mathbb{R}^k \to \mathbb{R}^p$:

$$\min_{f \in \mathcal{F},\, g \in \mathcal{G}} \sum_{i=1}^{n} \| x_i - g(f(x_i)) \|^2.$$

- f is the encoder: the *code* $f(x_i) \in \mathbb{R}^k$ has $k < p$ numbers
- g is the decoder: $g(f(x_i)) \in \mathbb{R}^p$ is the *reconstruction*

> **Note:** Different choices of $\mathcal{F}$ and $\mathcal{G}$ lead to different dimension reduction methods!

$\mathcal{F}, \mathcal{G} =$ linear leads to principal component analysis.
$\mathcal{F}, \mathcal{G} =$ neural leads to autoencoders.

Definition: Dimension reduction $=$ approximate X

Let the reconstruction $\widehat{X}$ have rows $g(f(x_i))^\mathsf{T}$. The reconstruction error is

$$\sum_{i=1}^{n} \|x_i - g(f(x_i))\|^2 \;=\; \left\|X - \widehat{X}\right\|_F^2,$$

where the Frobenius norm adds-up every squared entry:

$$\|A\|_F^2 \;=\; \sum_{i=1}^{n} \|\text{row } i \text{ of } A\|^2 \;=\; \sum_{i=1}^{n}\sum_{j=1}^{p} A_{ij}^2.$$

Plan for today

• Intro to dimension reduction	11 min
• Behold, the data matrix	10 min
• Outer products everywhere	16 min
• Spectral theorem	9 min
• Square matrices from the data matrix	4 min

Definition: the outer product

For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, put the transpose on the *other* factor:

$$\underbrace{x^\mathsf{T} x}_{1\times 1,\ \text{a scalar}} \quad \text{versus} \quad \underbrace{xy^\mathsf{T}}_{n\times p,\ \text{a matrix}} = \begin{bmatrix} — & x_1 y^\mathsf{T} & — \\ & \vdots & \\ — & x_n y^\mathsf{T} & — \end{bmatrix}, \qquad (xy^\mathsf{T})_{ij} = x_i y_j.$$

Definition: the outer product

For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, put the transpose on the *other* factor:

$$\underbrace{x^\mathsf{T} x}_{1\times 1,\ \text{a scalar}} \quad \text{versus} \quad \underbrace{xy^\mathsf{T}}_{n\times p,\ \text{a matrix}} = \begin{bmatrix} — & x_1 y^\mathsf{T} & — \\ & \vdots & \\ — & x_n y^\mathsf{T} & — \end{bmatrix}, \qquad (xy^\mathsf{T})_{ij} = x_i y_j.$$

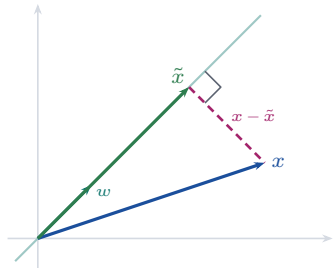
- Every column of xy^T is a multiple of x; every row is a multiple of y^T
- So its column space is the line through $x \implies xy^\mathsf{T}$ has rank one

Example: the projection matrix ww^T = outer product

For a unit vector $w \in \mathbb{R}^p$, the projection of $x \in \mathbb{R}^p$ onto w is

$$\tilde{x} = (x^\mathsf{T} w)\, w = w\,(w^\mathsf{T} x) = (ww^\mathsf{T})x.$$

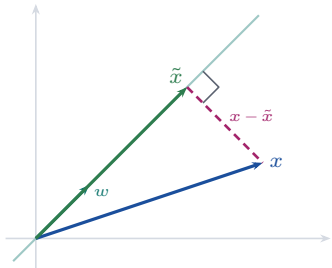
- $x^\mathsf{T} w$ is the score
- ww^T projects onto the line through w



Example: the projection matrix ww^T = outer product

For a unit vector $w \in \mathbb{R}^p$, the projection of $x \in \mathbb{R}^p$ onto w is

$$\tilde{x} = (x^\mathsf{T} w)\, w = w\,(w^\mathsf{T} x) = (ww^\mathsf{T})x.$$



- $x^\mathsf{T} w$ is the score
- ww^T projects onto the line through w

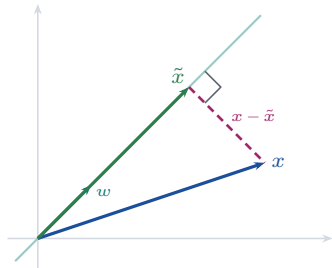
Since $x - \tilde{x} \perp w$, Pythagoras gives

$$\|x\|^2 = (x^\mathsf{T} w)^2 + \|x - \tilde{x}\|^2 .$$

Example: the projection matrix ww^T = outer product

For a unit vector $w \in \mathbb{R}^p$, the projection of $x \in \mathbb{R}^p$ onto w is

$$\tilde{x} = (x^\mathsf{T} w)\, w = w\,(w^\mathsf{T} x) = (ww^\mathsf{T})x.$$



- $x^\mathsf{T} w$ is the score
- ww^T projects onto the line through w

Since $x - \tilde{x} \perp w$, Pythagoras gives

$$\|x\|^2 = (x^\mathsf{T} w)^2 + \|x - \tilde{x}\|^2.$$

Key idea: Maximizing squared score $(x^\mathsf{T} w)^2$ $\iff$ minimizing squared projection error $\|x - \tilde{x}\|^2$.

Example: Outer products from orthonormal basis

Question

Let $v_1, \dots, v_p \in \mathbb{R}^p$ be orthonormal, meaning $v_i^\intercal v_j = \delta_{ij}$. How can we write $x \in \mathbb{R}^p$ as a sum over the v_i?

Example: Outer products from orthonormal basis

Question

Let $v_1, \dots, v_p \in \mathbb{R}^p$ be orthonormal, meaning $v_i^\mathsf{T} v_j = \delta_{ij}$. How can we write $x \in \mathbb{R}^p$ as a sum over the v_i?

Project x onto each and add:

$$\sum_{i=1}^{p} (v_i^\mathsf{T} x)\, v_i \;=\; \Big(\sum_{i=1}^{p} v_i v_i^\mathsf{T}\Big) x = x \qquad \text{because} \qquad \sum_{i=1}^{p} v_i v_i^\mathsf{T} = I_p.$$

Example: Outer products from orthonormal basis

Question

Let $v_1, \dots, v_p \in \mathbb{R}^p$ be orthonormal, meaning $v_i^\mathsf{T} v_j = \delta_{ij}$. How can we write $x \in \mathbb{R}^p$ as a sum over the v_i?

Project x onto each and add:

$$\sum_{i=1}^{p} (v_i^\mathsf{T} x)\, v_i \;=\; \Big(\sum_{i=1}^{p} v_i v_i^\mathsf{T}\Big) x = x \qquad \text{because} \qquad \sum_{i=1}^{p} v_i v_i^\mathsf{T} = I_p.$$

- Each $v_i v_i^\mathsf{T}$ is a rank-one projection matrix
- Squared lengths add as well: $\|x\|^2 = \sum_j (v_j^\mathsf{T} x)^2$

Exercise: centering, as an outer product

$$X = \begin{bmatrix} 5 & 5 \\ 1 & 1 \\ 4 & 2 \\ 2 & 4 \end{bmatrix}$$

Exercise

DS methods often assume that X is centered, meaning that the column average has been subtracted from all entries in the column.

1. Write the column vector of column sums of X, using the ones vector $\mathbf{1} \in \mathbb{R}^n$.
2. Use it to write the centered $\widetilde{X}$ as X minus one outer product.

Exercise: centering, as an outer product

$$X = \begin{bmatrix} 5 & 5 \\ 1 & 1 \\ 4 & 2 \\ 2 & 4 \end{bmatrix}$$

Exercise

DS methods often assume that X is centered, meaning that the column average has been subtracted from all entries in the column.

1. Write the column vector of column sums of X, using the ones vector $\mathbf{1} \in \mathbb{R}^n$.
2. Use it to write the centered $\widetilde{X}$ as X minus one outer product.

$$X^\mathsf{T}\mathbf{1} = \begin{bmatrix} 12 \\ 12 \end{bmatrix}, \quad \bar{x} = \tfrac{1}{n} X^\mathsf{T}\mathbf{1} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \widetilde{X} = X - \mathbf{1}\bar{x}^\mathsf{T} = \begin{bmatrix} 2 & 2 \\ -2 & -2 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

The usual rule: rows times columns

With A of size $m \times n$ and B of size $n \times p$, write A by its rows and B by its columns:

$$AB = \begin{bmatrix} — & a_1^\mathsf{T} & — \\ — & a_2^\mathsf{T} & — \\ & \vdots & \\ — & a_m^\mathsf{T} & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} a_1^\mathsf{T} b_1 & a_1^\mathsf{T} b_2 & \cdots & a_1^\mathsf{T} b_p \\ a_2^\mathsf{T} b_1 & a_2^\mathsf{T} b_2 & \cdots & a_2^\mathsf{T} b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^\mathsf{T} b_1 & a_m^\mathsf{T} b_2 & \cdots & a_m^\mathsf{T} b_p \end{bmatrix}.$$

The usual rule: rows times columns

With A of size $m \times n$ and B of size $n \times p$, write A by its rows and B by its columns:

$$AB = \begin{bmatrix} — & a_1^\mathsf{T} & — \\ — & a_2^\mathsf{T} & — \\ & \vdots & \\ — & a_m^\mathsf{T} & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} a_1^\mathsf{T} b_1 & a_1^\mathsf{T} b_2 & \cdots & a_1^\mathsf{T} b_p \\ a_2^\mathsf{T} b_1 & a_2^\mathsf{T} b_2 & \cdots & a_2^\mathsf{T} b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^\mathsf{T} b_1 & a_m^\mathsf{T} b_2 & \cdots & a_m^\mathsf{T} b_p \end{bmatrix}.$$

- $(AB)_{ij} = a_i^\mathsf{T} b_j$
- Every entry is an inner product, so a_i and b_j both live in $\mathbb{R}^n$

Key identity: a sum of outer products

Now slice the *other* way: $A \in \mathbb{R}^{m \times n}$ by its columns, $B \in \mathbb{R}^{n \times p}$ by its rows.

$$AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^\intercal & — \\ — & b_2^\intercal & — \\ & \vdots & \\ — & b_n^\intercal & — \end{bmatrix} = \sum_{\ell=1}^{n} a_\ell \, b_\ell^\intercal.$$

Key identity: a sum of outer products

Now slice the *other* way: $A \in \mathbb{R}^{m \times n}$ by its columns, $B \in \mathbb{R}^{n \times p}$ by its rows.

$$AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^\mathsf{T} & — \\ — & b_2^\mathsf{T} & — \\ & \vdots & \\ — & b_n^\mathsf{T} & — \end{bmatrix} = \sum_{\ell=1}^{n} a_\ell \, b_\ell^\mathsf{T}.$$

- Same matrix, now built one rank-one matrix at a time
- n terms, one for each column of A and the matching row of B

Key identity: a sum of outer products

Now slice the *other* way: $A \in \mathbb{R}^{m \times n}$ by its columns, $B \in \mathbb{R}^{n \times p}$ by its rows.

$$AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^\intercal & - \\ - & b_2^\intercal & - \\ & \vdots & \\ - & b_n^\intercal & - \end{bmatrix} = \sum_{\ell=1}^{n} a_\ell \, b_\ell^\intercal .$$

- Same matrix, now built one rank-one matrix at a time
- n terms, one for each column of A and the matching row of B

> **Note:** a_ℓ is $m \times 1$ and $b_\ell^\intercal$ is $1 \times p$, so every term is $m \times p$, the size of AB itself.

Example: when $A = X^\mathsf{T}$ and $B = X$

Coming up: Entries of $X^\mathsf{T}X \in \mathbb{R}^{p \times p}$ are features/predictor covariances.

The columns of X^T are $x_1, \ldots, x_n$ and the rows of X are $x_1^\mathsf{T}, \ldots, x_n^\mathsf{T}$, so

$$X^\mathsf{T}X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & x_1^\mathsf{T} & - \\ - & x_2^\mathsf{T} & - \\ & \vdots & \\ - & x_n^\mathsf{T} & - \end{bmatrix} = \sum_{i=1}^n x_i x_i^\mathsf{T}.$$

Example: when $A = X^\mathsf{T}$ and $B = X$

Coming up: Entries of $X^\mathsf{T}X \in \mathbb{R}^{p \times p}$ are features/predictor covariances.

The columns of X^T are $x_1, \ldots, x_n$ and the rows of X are $x_1^\mathsf{T}, \ldots, x_n^\mathsf{T}$, so

$$X^\mathsf{T}X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & x_1^\mathsf{T} & - \\ - & x_2^\mathsf{T} & - \\ & \vdots & \\ - & x_n^\mathsf{T} & - \end{bmatrix} = \sum_{i=1}^n x_i x_i^\mathsf{T}.$$

E.g., for the centered matrix from before:

$$\widetilde{X} = \begin{bmatrix} 2 & 2 \\ -2 & -2 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \widetilde{X}^\mathsf{T}\widetilde{X} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} + \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

The pattern to watch for

Coming up: Many important ideas in linear algebra (and data science) involve writing a matrix (e.g., $\widetilde{X}^T\widetilde{X}$) as a sum of outer products.

- The spectral theorem, in a few minutes, is one of them
- The singular value decomposition, on Wednesday, is another
- The useful expansions are the ones whose vectors are orthonormal

The pattern to watch for

Coming up: Many important ideas in linear algebra (and data science) involve writing a matrix (e.g., $\widetilde{X}^T\widetilde{X}$) as a sum of outer products.

- The spectral theorem, in a few minutes, is one of them
- The singular value decomposition, on Wednesday, is another
- The useful expansions are the ones whose vectors are orthonormal

Key idea: Once a matrix is a sum of rank-one pieces, you can *keep some and drop the rest*. This is a key idea in dimension reduction.

Plan for today

• Intro to dimension reduction	11 min
• Behold, the data matrix	10 min
• Outer products everywhere	16 min
• Spectral theorem	9 min
• Square matrices from the data matrix	4 min

Recall: eigenvectors

Definition

$v \neq 0$ is an eigenvector of A with eigenvalue λ if $Av = \lambda v$

Recall: eigenvectors

Definition

$v \neq 0$ is an eigenvector of A with eigenvalue λ if $Av = \lambda v$

For $A = \widetilde{X}^\intercal \widetilde{X} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$:

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 16 \\ 16 \end{bmatrix} = 16 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad A \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Recall: eigenvectors

Definition

$v \neq 0$ is an eigenvector of A with eigenvalue λ if $Av = \lambda v$

For $A = \widetilde{X}^\intercal \widetilde{X} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$:

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 16 \\ 16 \end{bmatrix} = 16 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad A \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Note: Normalize them: $v_1 = \frac{1}{\sqrt{2}}(1, 1)^\intercal$ and $v_2 = \frac{1}{\sqrt{2}}(1, -1)^\intercal$. Note $v_1 \perp v_2$, which is guaranteed by the spectral theorem.

An important fact about symmetric matrices

Theorem (Spectral theorem)

If $A \in \mathbb{R}^{p \times p}$ is symmetric, then there is an orthonormal basis of $\mathbb{R}^p$ consisting of eigenvectors $v_1, \dots, v_p$ of A with corresponding real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.

An important fact about symmetric matrices

Theorem (Spectral theorem)

If $A \in \mathbb{R}^{p \times p}$ is symmetric, then there is an orthonormal basis of $\mathbb{R}^p$ consisting of eigenvectors $v_1, \dots, v_p$ of A with corresponding real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, and

$$A = V \Lambda V^\mathsf{T} = \sum_i \lambda_i v_i v_i^\mathsf{T},$$

where $V = [v_1, \dots, v_p]$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$.

Spectral theorem: matrix form $A = V\Lambda V^\mathsf{T}$

Collect eigenvectors $v_1, \dots, v_p$ as columns of V, and let $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$.

Express definition of eigenvectors in matrix form. Then use fact that, since V is square with orthonormal columns, $VV^\mathsf{T} = I_p$:

$$Av_i = \lambda_i v_i,\ i = 1, \dots, p \quad \iff \quad AV = V\Lambda \quad \implies \quad A = V\Lambda V^\mathsf{T}.$$

Spectral theorem: matrix form $A = V\Lambda V^\mathsf{T}$

Collect eigenvectors $v_1, \dots, v_p$ as columns of V, and let $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$.

Express definition of eigenvectors in matrix form. Then use fact that, since V is square with orthonormal columns, $VV^\mathsf{T} = I_p$:

$$Av_i = \lambda_i v_i,\ i = 1, \dots, p \quad \iff \quad AV = V\Lambda \quad \implies \quad A = V\Lambda V^\mathsf{T}.$$

- V^T rotates into the eigenvector coordinates
- Λ scales each axis by its λ_j
- V rotates back

Spectral theorem: matrix form $A = V\Lambda V^\mathsf{T}$

Collect eigenvectors $v_1, \dots, v_p$ as columns of V, and let $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$.

Express definition of eigenvectors in matrix form. Then use fact that, since V is square with orthonormal columns, $VV^\mathsf{T} = I_p$:

$$Av_i = \lambda_i v_i,\ i = 1, \dots, p \quad \iff \quad AV = V\Lambda \quad \implies \quad A = V\Lambda V^\mathsf{T}.$$

- V^T rotates into the eigenvector coordinates
- Λ scales each axis by its λ_j
- V rotates back

> **Note:** Symmetric matrices merely stretch along p perpendicular axes.

Spectral theorem: outer product form

Apply outer product identity to $V\Lambda V^\mathsf{T}$: columns of $V\Lambda$ are $\lambda_j v_j$ and the rows of V^T are v_j^T, so

$$A = (V\Lambda)V^\mathsf{T} = \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_p v_p \\ | & & | \end{bmatrix} \begin{bmatrix} - & v_1^\mathsf{T} & - \\ & \vdots & \\ - & v_p^\mathsf{T} & - \end{bmatrix} = \sum_{j=1}^{p} \lambda_j v_j v_j^\mathsf{T}.$$

Spectral theorem: outer product form

Apply outer product identity to $V\Lambda V^\mathsf{T}$: columns of $V\Lambda$ are $\lambda_j v_j$ and the rows of V^T are v_j^T, so

$$A = (V\Lambda)V^\mathsf{T} = \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_p v_p \\ | & & | \end{bmatrix} \begin{bmatrix} - & v_1^\mathsf{T} & - \\ & \vdots & \\ - & v_p^\mathsf{T} & - \end{bmatrix} = \sum_{j=1}^{p} \lambda_j v_j v_j^\mathsf{T}.$$

With $\lambda_1 = 16$, $\lambda_2 = 4$ and our two v_j:

$$16 \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + 4 \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix} + \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}.$$

Spectral theorem: outer product form

Apply outer product identity to $V\Lambda V^\mathsf{T}$: columns of $V\Lambda$ are $\lambda_j v_j$ and the rows of V^T are v_j^T, so

$$A = (V\Lambda)V^\mathsf{T} = \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_p v_p \\ | & & | \end{bmatrix} \begin{bmatrix} - & v_1^\mathsf{T} & - \\ & \vdots & \\ - & v_p^\mathsf{T} & - \end{bmatrix} = \sum_{j=1}^{p} \lambda_j v_j v_j^\mathsf{T}.$$

With $\lambda_1 = 16$, $\lambda_2 = 4$ and our two v_j:

$$16 \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + 4 \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix} + \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}.$$

Key idea: Compare $\sum_j v_j v_j^\mathsf{T} = I$: the *same* expansion, now weighted by the λ_j. A symmetric matrix is the identity with the axes rescaled.

Plan for today

• Intro to dimension reduction	11 min
• Behold, the data matrix	10 min
• Outer products everywhere	16 min
• Spectral theorem	9 min
• Square matrices from the data matrix	4 min

Square from rectangular: $X^\mathsf{T}X$ and XX^T

Eigenvalues only defined for square matrices, but $n \times p$ data matrices are rarely square. However,

$\underbrace{X^\mathsf{T}X}_{p\times p}$ and $\underbrace{XX^\mathsf{T}}_{n\times n}$ always are, and both are symmetric!

$$(X^\mathsf{T}X)^\mathsf{T} = X^\mathsf{T}(X^\mathsf{T})^\mathsf{T} = X^\mathsf{T}X.$$

Square from rectangular: $X^\mathsf{T}X$ and XX^T

Eigenvalues only defined for square matrices, but $n \times p$ data matrices are rarely square. However,

$$\underbrace{X^\mathsf{T}X}_{p\times p} \quad \text{and} \quad \underbrace{XX^\mathsf{T}}_{n\times n} \quad \text{always are, and both are symmetric!}$$

$$(X^\mathsf{T}X)^\mathsf{T} = X^\mathsf{T}(X^\mathsf{T})^\mathsf{T} = X^\mathsf{T}X.$$

Key idea: This is how we can apply the spectral theorem to data matrices. We'll get a lot of use out of this trick!

$$\sum_{i=1}^{n} x_i x_i^\intercal \;=\; X^\intercal X \;=\; \sum_{j=1}^{p} \lambda_j v_j v_j^\intercal$$

Two expansions of one matrix

$$\sum_{i=1}^{n} x_i x_i^\intercal \;=\; X^\intercal X \;=\; \sum_{j=1}^{p} \lambda_j v_j v_j^\intercal$$

Two expansions of one matrix

$$\sum_{i=1}^{n} x_i x_i^\mathsf{T} \;=\; X^\mathsf{T} X \;=\; \sum_{j=1}^{p} \lambda_j v_j v_j^\mathsf{T}$$

- Left: n terms, one per example/observation, in no particular direction and not orthogonal
- Right: p terms, one per direction, orthonormal, each labeled with its "size" λ_j

Two expansions of one matrix

$$\sum_{i=1}^{n} x_i x_i^\mathsf{T} \;=\; X^\mathsf{T} X \;=\; \sum_{j=1}^{p} \lambda_j v_j v_j^\mathsf{T}$$

- Left: n terms, one per example/observation, in no particular direction and not orthogonal
- Right: p terms, one per direction, orthonormal, each labeled with its "size" λ_j

Coming up: On Monday, keep the biggest terms on the right-hand sum and discard the rest.