Lecture 3

Principal component analysis (PCA)

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

- This Friday: first quiz, then problem solving and programming

Plan for today

• Recap of Lecture 2	10 min
• Principal component analysis	5 min
• Demo	15 min
• Why it solves the (linear) dim. reduction problem	20 min

Recap: Data matrix

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

n rows, p columns

- One row per example
- One column per feature
- We generally center the columns

Recap: We usually center the data matrix

$$X - \mathbf{1}\overline{x}^\mathsf{T} = \begin{bmatrix} — & (x_1 - \overline{x})^\mathsf{T} & — \\ — & (x_2 - \overline{x})^\mathsf{T} & — \\ & \vdots & \\ — & (x_n - \overline{x})^\mathsf{T} & — \end{bmatrix}$$

- $\mathbf{1} \in \mathbb{R}^n$ is all-ones vector
- $\overline{x} = \frac{1}{n} X^\mathsf{T} \mathbf{1}$ is vector of col. averages
- $\mathbf{1}\overline{x}^\mathsf{T}$ stacks n copies of $\overline{x}^\mathsf{T}$

Recap: We usually center the data matrix

$$X - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}X = \begin{bmatrix} — & (x_1 - \overline{x})^\mathsf{T} & — \\ — & (x_2 - \overline{x})^\mathsf{T} & — \\ & \vdots & \\ — & (x_n - \overline{x})^\mathsf{T} & — \end{bmatrix}$$

- $\mathbf{1} \in \mathbb{R}^n$ is all-ones vector
- $\overline{x} = \frac{1}{n}X^\mathsf{T}\mathbf{1}$ is vector of col. averages
- $\mathbf{1}\overline{x}^\mathsf{T}$ stacks n copies of $\overline{x}^\mathsf{T}$

Recap: A 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 $\lambda_1 \geq \cdots \geq \lambda_p$.

Recap: A 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 $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$.

Recap: A 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 $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$.

Key idea: Although the data matrix $X \in \mathbb{R}^{n \times p}$ is rarely symmetric, the covariance matrix $X^\mathsf{T} X \in \mathbb{R}^{p \times p}$ always is!

Recap: Application of spectral theorem to $X^\mathsf{T}X$

Key idea: Although the data matrix $X \in \mathbb{R}^{n \times p}$ is rarely symmetric, the covariance matrix $X^\mathsf{T}X \in \mathbb{R}^{p \times p}$ always is!

- The spectral theorem applies to $X^\mathsf{T}X$

Recap: Application of spectral theorem to $X^\mathsf{T}X$

Key idea: Although the data matrix $X \in \mathbb{R}^{n \times p}$ is rarely symmetric, the covariance matrix $X^\mathsf{T}X \in \mathbb{R}^{p \times p}$ always is!

- The spectral theorem applies to $X^\mathsf{T}X$
- So there is an orthonormal basis $v_1, \dots, v_p$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ such that

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

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

Recap: Application of spectral theorem to $X^\mathsf{T}X$

Key idea: Although the data matrix $X \in \mathbb{R}^{n\times p}$ is rarely symmetric, the covariance matrix $X^\mathsf{T}X \in \mathbb{R}^{p\times p}$ always is!

- The spectral theorem applies to $X^\mathsf{T}X$
- So there is an orthonormal basis $v_1, \dots, v_p$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ such that

$$X^\mathsf{T}X = V_p \Lambda V_p^\mathsf{T} = \sum_i \lambda_i v_i v_i^\mathsf{T},$$

where $V_k = [v_1, \dots, v_k]$ for $1 \leq k \leq p$, and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$.

Plan for today

- Recap of Lecture 2 — 10 min
- Principal component analysis — 5 min
- Demo — 15 min
- Why it solves the (linear) dim. reduction problem — 20 min

Algorithm: Principal component analysis

Input. Data matrix $X \in \mathbb{R}^{n \times p}$, a dimension $k \in \{1, \dots, p-1\}$.

Output. Principal components $Z \in \mathbb{R}^{n \times k}$, directions $V_k \in \mathbb{R}^{p \times k}$.

1. Center: $X \leftarrow X - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}X$
2. Form: the matrix $X^\mathsf{T}X \in \mathbb{R}^{p \times p}$
3. Diagonalize: get its orthonormal eigenvectors $v_1, \dots, v_p$, ordered so that $\lambda_1 \geq \cdots \geq \lambda_p$
4. Truncate: $V_k = [\, v_1 \ \cdots \ v_k \,]$
5. Project: $Z = XV_k$

Algorithm: Principal component analysis

Input. Data matrix $X \in \mathbb{R}^{n \times p}$, a dimension $k \in \{1, \dots, p-1\}$.

Output. Principal components $Z \in \mathbb{R}^{n \times k}$, directions $V_k \in \mathbb{R}^{p \times k}$.

1. Center: $X \leftarrow X - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}X$
2. Form: the sample covariance matrix $\frac{1}{n-1}X^\mathsf{T}X \in \mathbb{R}^{p \times p}$
3. Diagonalize: get its orthonormal eigenvectors $v_1, \dots, v_p$, ordered so that $\lambda_1 \geq \cdots \geq \lambda_p$
4. Truncate: $V_k = [\, v_1 \ \cdots \ v_k \,]$
5. Project: $Z = XV_k$

> **Note:** For reasons we'll later discuss, the "sample covariance matrix" $\frac{1}{n-1}X^\mathsf{T}X$ is often used in the place of $X^\mathsf{T}X$, even though its eigenvectors are the same.

Plan for today

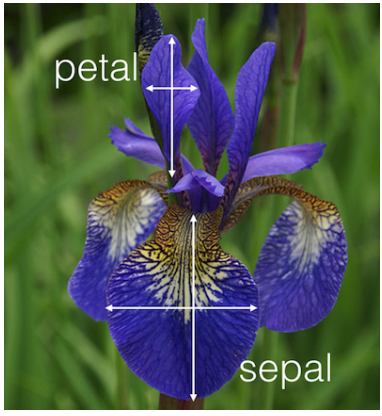
- Recap of Lecture 2 — 10 min
- Principal component analysis — 5 min
- Demo — 15 min
- Why it solves the (linear) dim. reduction problem — 20 min

PCA demo

1. Download `lec03_pca_demo.ipynb` from course webpage or Canvas
2. Go to Google Colab and use your UD email address to make an account
3. Open the .ipynb file on Google Colab
4. Let me know when you have it open

Example: Iris dataset

- Contains measurements of 150 irises
- For each, we know: petal and sepal length and width (in cm)
- So each flower is a point in $\mathbb{R}^4$
- How should we explore these data?



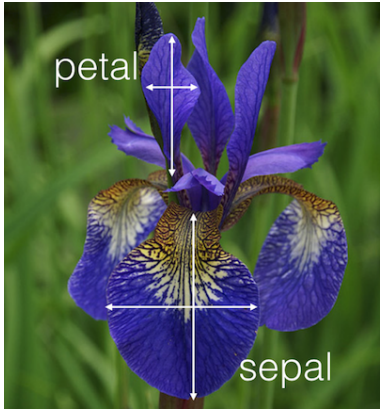
From Jadhav et al., ICCP (2020).

Example: Iris dataset

- Contains measurements of 150 irises
- For each, we know: petal and sepal length and width (in cm)
- So each flower is a point in $\mathbb{R}^4$
- How should we explore these data?

Problem

We cannot visualize four-dimensional data.



From Jadhav et al., ICCP (2020).

Example: Iris dataset

After subtracting-off column averages, the data matrix $X \in \mathbb{R}^{150 \times 4}$ looks like

$X =$

Out[1]:

	sepal_length	sepal_width	petal_length	petal_width
0	-0.7	0.4	-2.4	-1.0
1	-0.9	-0.1	-2.4	-1.0
2	-1.1	0.1	-2.5	-1.0
3	-1.2	0.0	-2.3	-1.0
4	-0.8	0.5	-2.4	-1.0

Example: Iris dataset

After subtracting-off column averages, the data matrix $X \in \mathbb{R}^{150 \times 4}$ looks like

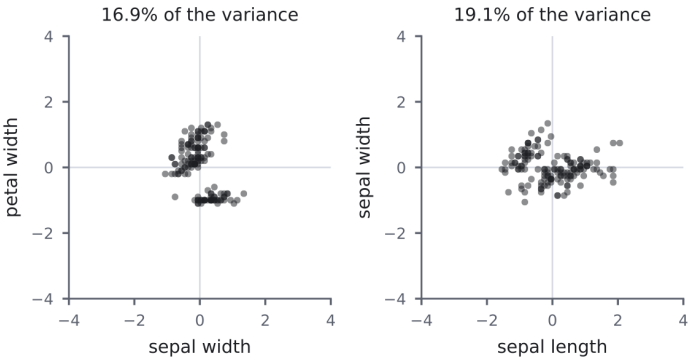
$X =$

Out[1]:

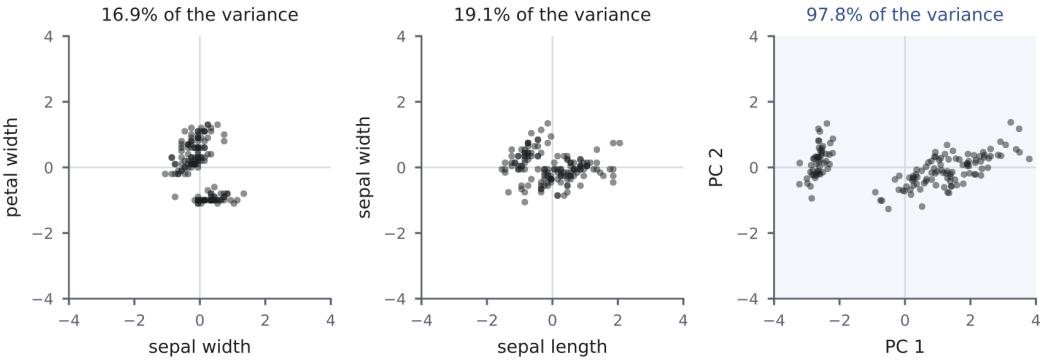
	sepal_length	sepal_width	petal_length	petal_width
0	-0.7	0.4	-2.4	-1.0
1	-0.9	-0.1	-2.4	-1.0
2	-1.1	0.1	-2.5	-1.0
3	-1.2	0.0	-2.3	-1.0
4	-0.8	0.5	-2.4	-1.0

Idea: To visualize, why not scatter pairs of columns?

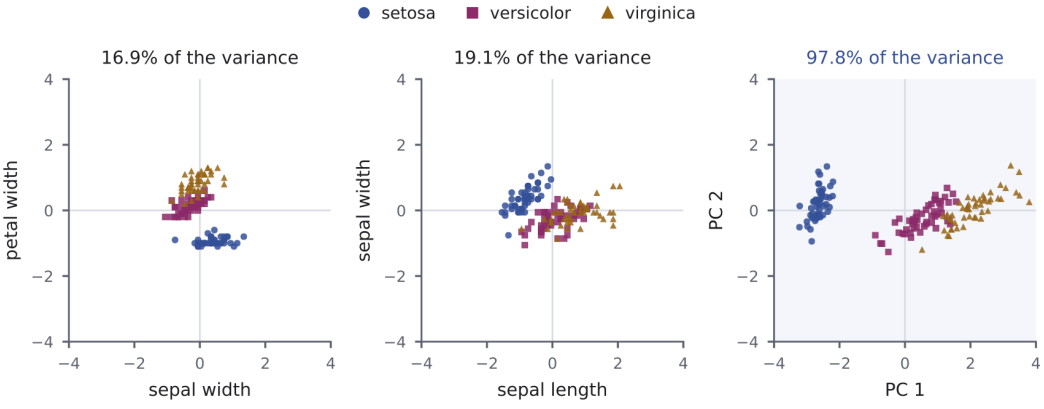
Example: What about pairs of columns?



Example: Two linear combinations of all columns?



Example: Two linear combinations of all columns?



Takeaways

- PCA can help to visualize high-dimensional data
- It identifies directions along which the data vary most
- We project the data onto these directions to get the principal components
- PCA can serve as an input to predictive analyses

Plan for today

- Recap of Lecture 2 — 10 min
- Principal component analysis — 5 min
- Demo — 15 min
- Why it solves the (linear) dim. reduction problem — 20 min

Recall: General dimension reduction problem

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*

$\mathcal{F}, \mathcal{G} =$ linear leads to PCA
$\mathcal{F}, \mathcal{G} =$ neural leads to "autoencoders"

Linear dimension reduction to 1D

Given examples $x_1, \dots, x_n \in \mathbb{R}^p$, solve

$$\min_{w \in \mathbb{R}^p : \|w\| = 1} \quad \sum_{i=1}^{n} \| x_i - (x_i^\intercal w) w \|^2.$$

Linear dimension reduction to 1D

Given examples $x_1, \dots, x_n \in \mathbb{R}^p$, solve

$$\min_{w \in \mathbb{R}^p : \|w\|=1} \sum_{i=1}^{n} \| x_i - (x_i^\mathsf{T} w) w \|^2.$$

- linearly "encode" x_i as the scalar $z = x_i^\mathsf{T} w$
- linearly "decode" z into $\mathbb{R}^p$ as zw
- $\hat{x}_i = zw = (x_i^\mathsf{T} w) w$ is "reconstruction" of x_i

Linear dimension reduction to 1D

Given examples $x_1, \dots, x_n \in \mathbb{R}^p$, solve

$$\min_{w \in \mathbb{R}^p \colon \|w\|=1} \quad \sum_{i=1}^{n} \| x_i - (x_i^\mathsf{T} w) w \|^2.$$

Recall that (Pythagoras):

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

Linear dimension reduction to 1D

Given examples $x_1, \dots, x_n \in \mathbb{R}^p$, solve

$$\min_{w \in \mathbb{R}^p : \|w\|=1} \sum_{i=1}^{n} \| x_i - (x_i^\mathsf{T} w) w \|^2.$$

Recall that (Pythagoras):

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

Hence, minimizing reconstruction error is the same as maximizing

$$\sum_{i=1}^{n} (x_i^\mathsf{T} w)^2 = \sum_{i=1}^{n} (Xw)_i^2 = (Xw)^\mathsf{T}(Xw) = w^\mathsf{T} X^\mathsf{T} X w.$$

Along which direction do the data vary most?

Problem (Linear dimension reduction to 1D)

Find unit vector $w^{(1)} \in \mathbb{R}^p$ along which data vary most:

$$w^{(1)} = \underset{\|w\|=1}{\arg\max}\, w^\mathsf{T} X^\mathsf{T} X w$$

Along which direction do the data vary most?

Problem (Linear dimension reduction to 1D)

Find unit vector $w^{(1)} \in \mathbb{R}^p$ along which data vary most:

$$w^{(1)} = \underset{\|w\|=1}{\arg\max}\, w^\mathsf{T} X^\mathsf{T} X w$$

Note that:

- no solution without constraint of $\|w\| = 1$
- we can use $w^{(1)}$ as an axis of our scatter plots,
- then find $w^{(2)} \perp w^{(1)}$ in a similar way, and so on

Solution: Linear dimension reduction to 1D

Theorem

The maximizer $w^{(1)}$ of $w^\mathsf{T} X^\mathsf{T} X w$ over all unit vectors $w \in \mathbb{R}^p$ is $w^{(1)} = v_1$, the eigenvector of $X^\mathsf{T} X$ with greatest eigenvalue.

Solution: Linear dimension reduction to 1D

Theorem

The maximizer $w^{(1)}$ of $w^\mathsf{T} X^\mathsf{T} X w$ over all unit vectors $w \in \mathbb{R}^p$ is $w^{(1)} = v_1$, the eigenvector of $X^\mathsf{T} X$ with greatest eigenvalue.

Proof. First, $w^\mathsf{T} X^\mathsf{T} X w$ is at least λ_1 because the choice $w = v_1$ gives

$$v_1^\mathsf{T}(X^\mathsf{T} X v_1) = v_1^\mathsf{T}(\lambda_1 v_1) = \lambda_1 v_1^\mathsf{T} v_1 = \lambda_1.$$

□

Solution: Linear dimension reduction to 1D

Theorem

The maximizer $w^{(1)}$ of $w^\mathsf{T} X^\mathsf{T} X w$ over all unit vectors $w \in \mathbb{R}^p$ is $w^{(1)} = v_1$, the eigenvector of $X^\mathsf{T} X$ with greatest eigenvalue.

Proof. First, $w^\mathsf{T} X^\mathsf{T} X w$ is at least λ_1 because the choice $w = v_1$ gives

$$v_1^\mathsf{T}(X^\mathsf{T} X v_1) = v_1^\mathsf{T}(\lambda_1 v_1) = \lambda_1 v_1^\mathsf{T} v_1 = \lambda_1.$$

Second, write w in the eigenbasis of $X^\mathsf{T} X$ as $w = \sum_{i=1}^p (w^\mathsf{T} v_i) v_i$. Then,

$$w^\mathsf{T} X^\mathsf{T} X w = \sum_{i=1}^p (w^\mathsf{T} v_i) w^\mathsf{T} X^\mathsf{T} X v_i = \sum_{i=1}^p \lambda_i (w^\mathsf{T} v_i)^2 \leq \lambda_1 \underbrace{\sum_{i=1}^p (w^\mathsf{T} v_i)^2}_{= \|w\|^2 = 1} = \lambda_1.$$

□

What it means: The first principal component

Theorem

The maximizer $w^{(1)}$ of $w^\mathsf{T} X^\mathsf{T} X w$ over all unit vectors $w \in \mathbb{R}^p$ is $w^{(1)} = v_1$, the eigenvector of $X^\mathsf{T} X$ with greatest eigenvalue.

Loadings $w^{(1)} \in \mathbb{R}^p$
how much of each feature goes into the first PC

Scores $Xw^{(1)} \in \mathbb{R}^n$
the coordinates of each example in the first PC's direction

Watch out: "Principal component" is sometimes used to refer to one of these things or the other.

Linear dimension reduction to $k+1$ dimensions

Having found $w^{(1)}, \ldots, w^{(k)}$, ask the same question of the directions we have not used:

$$w^{(k+1)} = \underset{\substack{\|w\|=1 \\ w \perp w^{(1)},\ldots,w^{(k)}}}{\arg\max} \sum_{i=1}^{n} (x_i^\mathsf{T} w)^2 = \underset{\substack{\|w\|=1 \\ w \perp w^{(1)},\ldots,w^{(k)}}}{\arg\max} w^\mathsf{T} X^\mathsf{T} X w.$$

- An entirely analogous argument works!
- So $w^{(k+1)} = v_{k+1}$, with variation λ_{k+1}

Key idea: One eigendecomposition does PCA for all k at once, and in order!

Summary of PCA

Theorem

Let $X \in \mathbb{R}^{n \times p}$ be centered and let $X^\mathsf{T} X = V_p \Lambda V_p^\mathsf{T}$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p$, and $V_p = [\, v_1 \cdots v_p \,]$ orthogonal. Then the k-th column of the "score" matrix $Z = XV_p$ is the k-th principal component of X, and its variation is

$$\|Xv_k\|^2 \;=\; v_k^\mathsf{T} X^\mathsf{T} X v_k \;=\; \lambda_k.$$

Summary of PCA

Theorem

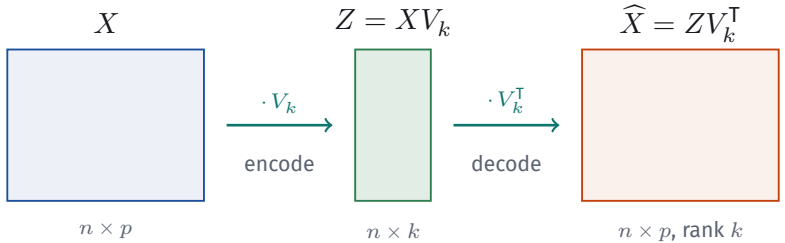
Let $X \in \mathbb{R}^{n \times p}$ be centered and let $X^\mathsf{T} X = V_p \Lambda V_p^\mathsf{T}$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p$, and $V_p = [\, v_1 \cdots v_p \,]$ orthogonal. Then the k-th column of the "score" matrix $Z = XV_p$ is the k-th principal component of X, and its variation is

$$\|Xv_k\|^2 \;=\; v_k^\mathsf{T} X^\mathsf{T} X v_k \;=\; \lambda_k.$$

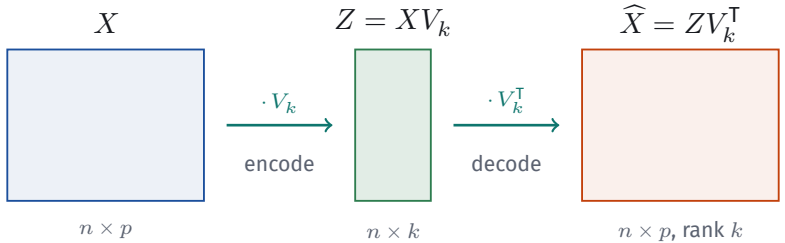
Check (Encode data matrix as XV_k, but examples as $V_k^\mathsf{T} x_i$!)

The i-th row of $Z = XV_k$ is the (transposed) result of applying the encoder V_k^T to example x_i. The reconstruction of x_i is $\widehat{x}_i = V_k V_k^\mathsf{T} x_i$.

Where the dimensions go



Where the dimensions go



Key idea: Although $\widehat{X}$ and X have the same number of columns (p), the rank of $\widehat{X}$ is only k.

How much variation do k dimensions capture?

The total variation $\mathrm{tr}(X^\mathsf{T} X) = \lambda_1 + \cdots + \lambda_p$ is fixed, and the first k components explain

$$\mathsf{PVE}(k) \;=\; \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\%.$$

- PCA works well when the first few λ_i dominate the sum
- This may not be the case! E.g., isotropic noise has $\lambda_1 = \cdots = \lambda_p$.

Example: what are PC_1 and PC_2 made of?

	sepal len.	sepal wid.	petal len.	petal wid.
$w^{(1)}$	0.361	−0.085	0.857	0.358
$w^{(2)}$	0.657	−0.730	0.173	0.076

- PC_1 is mostly petal length
- PC_2 is balanced between sepal length and width

Watch out: The signs are irrelevant because, e.g., v_i and $-v_i$ are both eigenvectors.

Using PCA to explain a certain amount of variation

Definition (The recipe)

Given a target percentage s of explained variation:

1. Center each column of X.
2. Eigendecompose $X^\mathsf{T}X = V_p \Lambda V_p^\mathsf{T}$ and choose the smallest k with $\text{PVE}(k) \geq s$.
3. Replace $X \in \mathbb{R}^{n \times p}$ by the first k columns of XV_k.

Same n observations, now described by $k \ll p$ variables!

Three mistakes with PCA

- Forgetting to center. $X^\mathsf{T}X$ on uncentered data has a large first eigenvector pointing at the mean.

Three mistakes with PCA

- Forgetting to center. $X^\mathsf{T}X$ on uncentered data has a large first eigenvector pointing at the mean.
- Ignoring units. While the iris data are all measured in cm, if one column was instead measured in mm, it would dominate PC_1. In the case of columns with different units, it is generally advisable to standardize the columns.

Three mistakes with PCA

- Forgetting to center. $X^\mathsf{T}X$ on uncentered data has a large first eigenvector pointing at the mean.
- Ignoring units. While the iris data are all measured in cm, if one column was instead measured in mm, it would dominate PC_1. In the case of columns with different units, it is generally advisable to standardize the columns.
- Reading variance as separation. PCA never sees the class labels (when they exist). Just because there is high variation in a direction does not mean it will nicely separate the classes.

A final note on implementation

We derived everything from $X^\mathsf{T}X$. It is generally unwise to *compute* with it.

- Forming $X^\mathsf{T}X$ squares the condition number $\kappa(X^\mathsf{T}X) = \kappa(X)^2$, which can easily lead to a loss of numerical precision.
- Implementations of PCA instead use the singular value decomposition (SVD), which we will discuss on Wednesday.