DSC 140B
Representation Learning

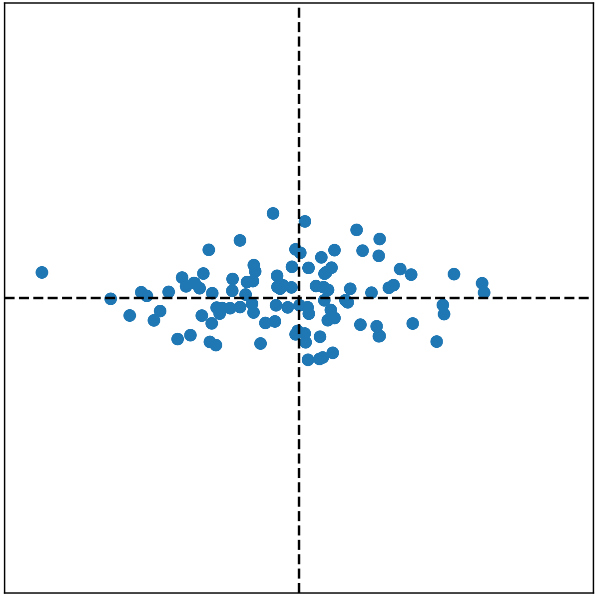
Lecture 10 | Part 1

Visualizing Covariance Matrices

Visualizing Covariance Matrices

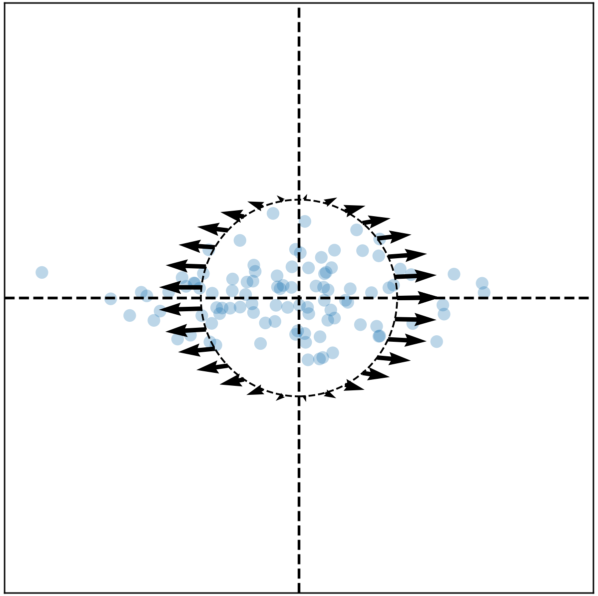
- Covariance matrices are symmetric.
- They have axes of symmetry (eigenvectors and eigenvalues).
- What are they?

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} \quad & \quad \\ \quad & \quad \end{pmatrix}$$

Visualizing Covariance Matrices

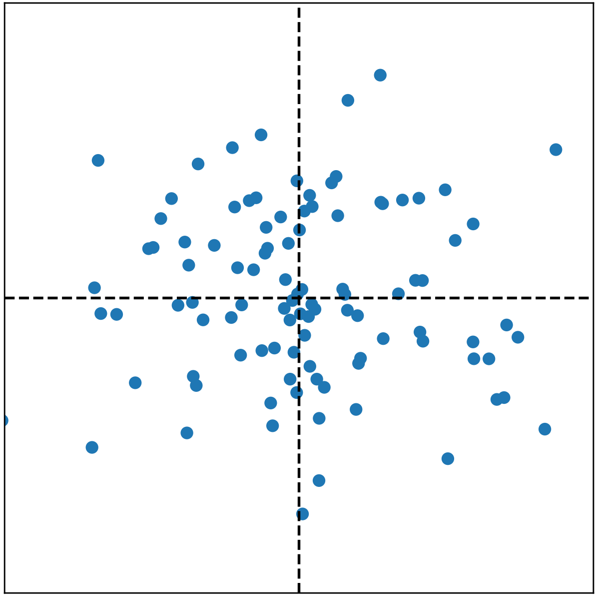


Eigenvectors:

$\vec{u}^{(1)} \approx$

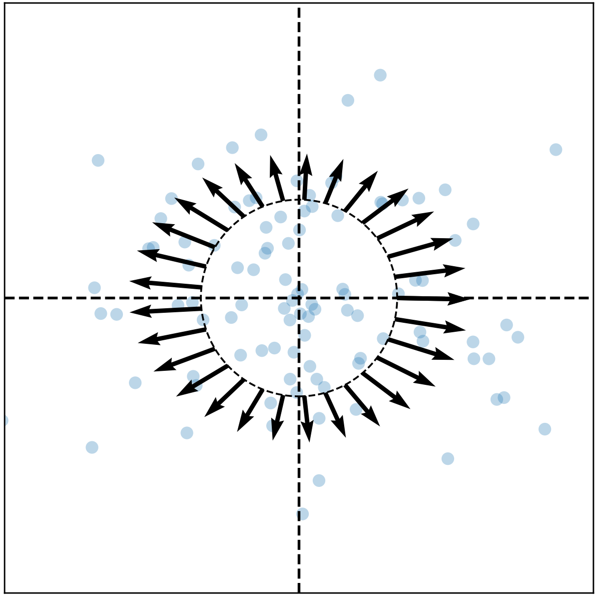
$\vec{u}^{(2)} \approx$

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} & \\ & \end{pmatrix}$$

Visualizing Covariance Matrices

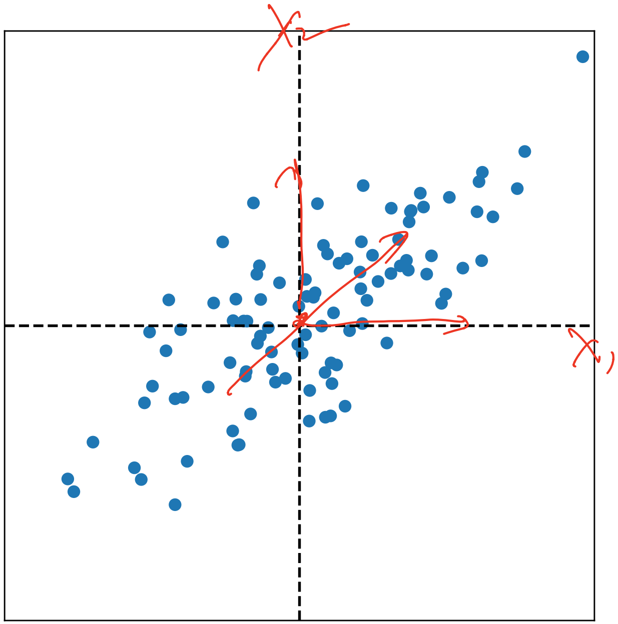


Eigenvectors:

$\vec{u}^{(1)} \approx$

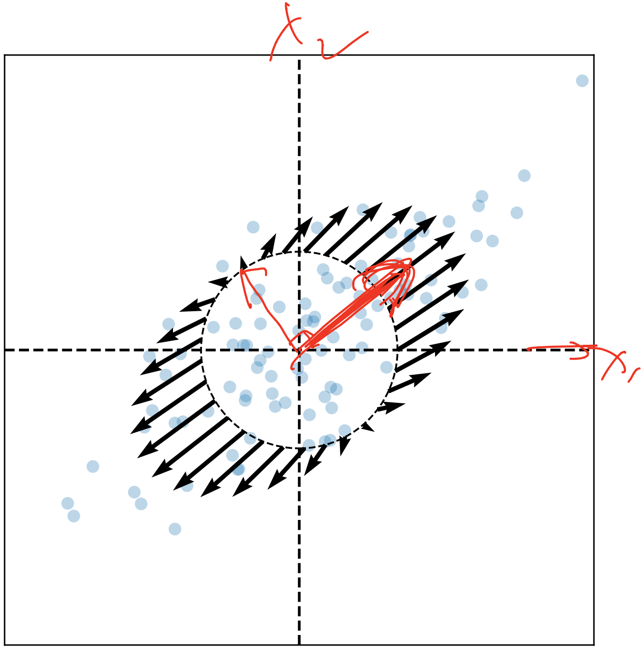
$\vec{u}^{(2)} \approx$

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 10 & 5 \\ 5 & 7 \end{pmatrix}$$

Visualizing Covariance Matrices



$$\begin{pmatrix} 60 & 5 \\ 5 & 7 \end{pmatrix}$$

Eigenvectors:

$\vec{u}^{(1)} \approx (1, 1)$

$\vec{u}^{(2)} \approx (-1, 1)$

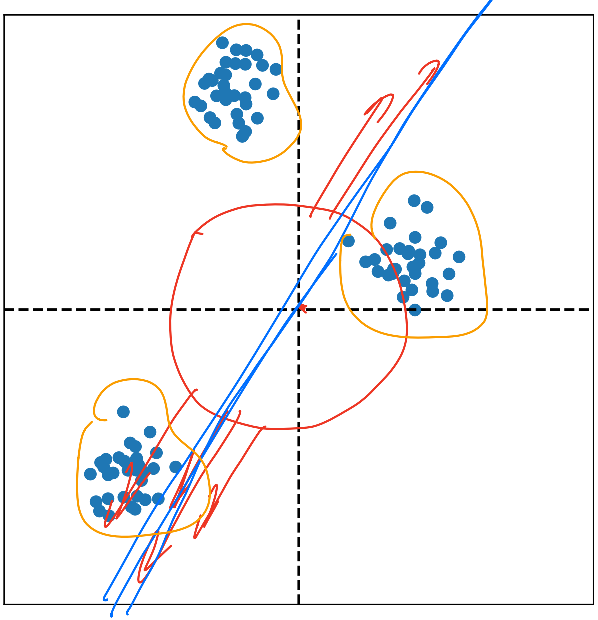
Intuitions

- The **eigenvectors** of the covariance matrix describe the data's "principal directions"
  - C tells us something about data's shape.
- The **top eigenvector** points in the direction of "maximum variance".
- The **top eigenvalue** is proportional to the variance in this direction.

Caution

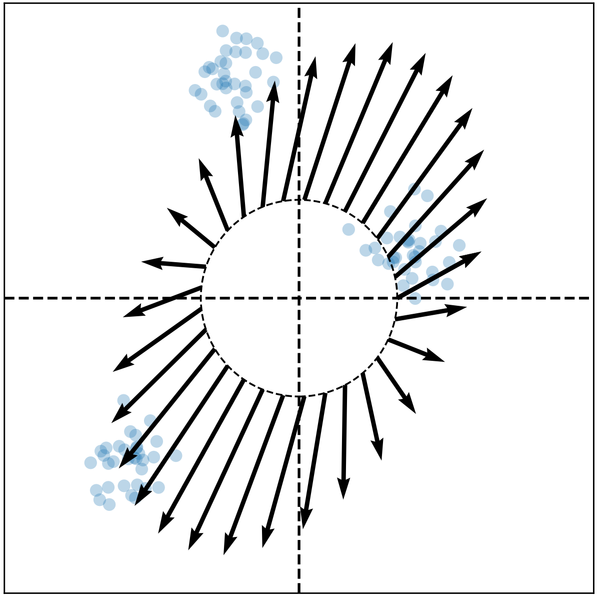
- The data doesn't always look like this.
- We can always compute covariance matrices.
- They just may not describe the data's shape very well.

$C = \frac{1}{n} X^T X$



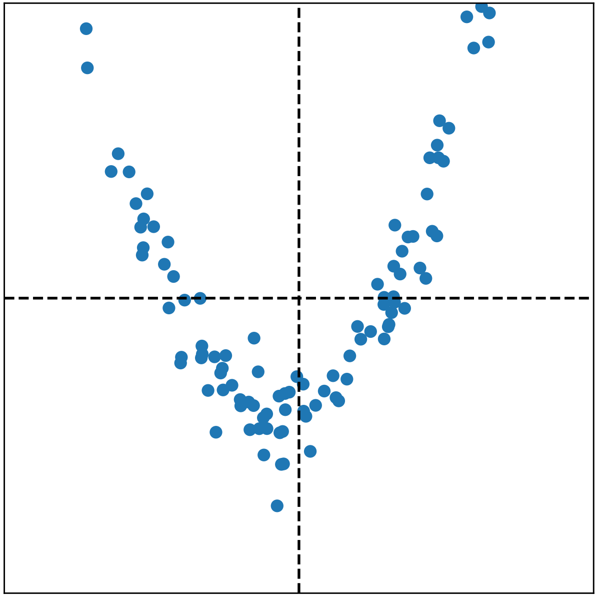
Caution

- The data doesn't always look like this.
- We can always compute covariance matrices.
- They just may not describe the data's shape very well.



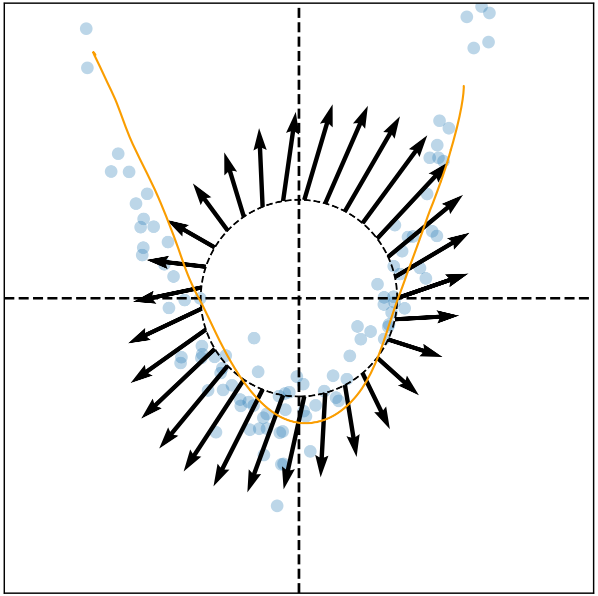
Caution

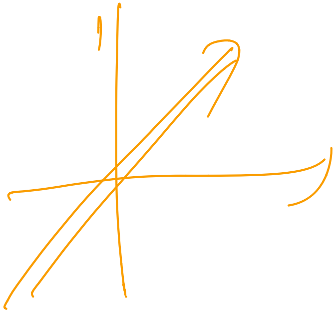
- The data doesn't always look like this.
- We can always compute covariance matrices.
- They just may not describe the data's shape very well.



Caution

- The data doesn't always look like this.
- We can always compute covariance matrices.
- They just may not describe the data's shape very well.





DSC 140B
Representation Learning

Lecture 10 | Part 2

PCA, More Formally

The Story (So Far)

- We want to create a single new feature, z.
- Our idea: $z = \vec{x} \cdot \vec{u}$; choose $\vec{u}$ to point in the "direction of maximum variance".
- Intuition: the top eigenvector of the covariance matrix points in direction of maximum variance.

More Formally...

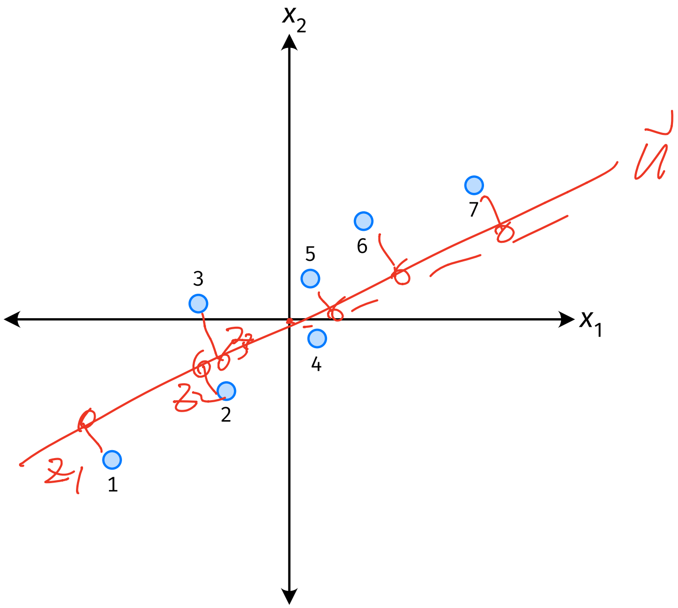
- We haven’t actually defined “direction of maximum variance”
- Let’s derive PCA more formally.

Variance in a Direction

- Let $\vec{u}$ be a unit vector.
- $z^{(i)} = \vec{x}^{(i)} \cdot \vec{u}$ is the new feature for $\vec{x}^{(i)}$.
- The variance of the new features is:

$$\begin{aligned}\text{Var}(z) &= \frac{1}{n}\sum_{i=1}^{n}(z^{(i)} - \mu_z)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\vec{x}^{(i)} \cdot \vec{u} - \mu_z\right)^2\end{aligned}$$

Example



Note

- If the data are centered, then $\mu_z = 0$ and the variance of the new features is:

$$\text{Var}(z) = \frac{1}{n}\sum_{i=1}^{n}(z^{(i)})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\vec{x}^{(i)} \cdot \vec{u}\right)^2$$

Goal

- The variance of a data set in the direction of $\vec{u}$ is:

$$g(\vec{u}) = \frac{1}{n} \sum_{i=1}^{n} \left(\vec{x}^{(i)} \cdot \vec{u}\right)^2$$

$$\max_{\vec{u}} g(\vec{u})$$

- Our goal: Find a unit vector $\vec{u}$ which maximizes g.

Claim

$$\frac{1}{n}\sum_{i=1}^{n}\left(\vec{x}^{(i)} \cdot \vec{u}\right)^2 = \vec{u}^T C \vec{u}$$

$$C = \frac{1}{n} X^T X$$

$$\frac{1}{n}\vec{u}^T X^T X \vec{u} = \frac{1}{n}(X\cdot\vec{u})^T (X\cdot\vec{u}) = \frac{1}{n}\vec{v}\cdot\vec{v}$$

$$X \in \mathbb{R}^{n \times d}, \quad X\vec{u} = \begin{pmatrix} \vec{x}^{(1)}\cdot\vec{u} \\ \vec{x}^{(2)}\cdot\vec{u} \\ \vdots \\ \vec{x}^{(n)}\cdot\vec{u} \end{pmatrix} = X\cdot\vec{u} = \vec{v}$$

Our Goal (Again)

- Find a unit vector $\vec{u}$ which maximizes $\vec{u}^T C \vec{u}$.

$$\max_{\vec{u}} \vec{u}^T C \vec{u} \quad \text{s.t. } \|\vec{u}\| = 1$$

Claim

- To maximize $\vec{u}^T C \vec{u}$ over unit vectors, choose $\vec{u}$ to be the top eigenvector of C.

- Proof:

Proof

Show that the maximizer of $\|A\vec{x}\|$ s.t., $\|\vec{x}\| = 1$ is the top eigenvector of A.

eigen decomposition

assume A: $\vec{u}^{(1)}$ $\vec{u}^{(2)}$

$\|b_1\vec{u}^{(1)} + b_2\vec{u}^{(2)}\|^2 = 1 \Rightarrow b_1^2 + b_2^2 = 1$

A symmetric

$$A\vec{x} = A(b_1\vec{u}^{(1)} + b_2\vec{u}^{(2)})$$
$$= b_1 A\vec{u}^{(1)} + b_2 A\vec{u}^{(2)}$$
$$= b_1\lambda_1\vec{u}^{(1)} + b_2\lambda_2\vec{u}^{(2)}$$

$$\|A\vec{x}\|^2 = (b_1\lambda_1\vec{u}^{(1)} + b_2\lambda_2\vec{u}^{(2)})\cdot(b_1\lambda_1\vec{u}^{(1)} + b_2\lambda_2\vec{u}^{(2)})$$
$$= b_1^2\lambda_1^2\vec{u}^{(1)T}\vec{u}^{(1)} + 2b_1b_2\lambda_1\lambda_2\vec{u}^{(1)T}\vec{u}^{(2)} + b_2^2\lambda_2^2\vec{u}^{(2)T}\vec{u}^{(2)}$$

max $b_1^2\lambda_1^2 + b_2^2\lambda_2^2$

assume $\lambda_1^2 \geq \lambda_2^2$

$\Rightarrow b_1 = 1 \quad b_2 = 0$

Claim

- To maximize $\vec{u}^T C \vec{u}$ over unit vectors, choose $\vec{u}$ to be the top eigenvector of C.
- Proof:

Claim

- To maximize $\vec{u}^T C \vec{u}$ over unit vectors, choose $\vec{u}$ to be the top eigenvector of C.
- Proof:

PCA (for a single new feature)

- **Given**: data points $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)} \in \mathbb{R}^d$

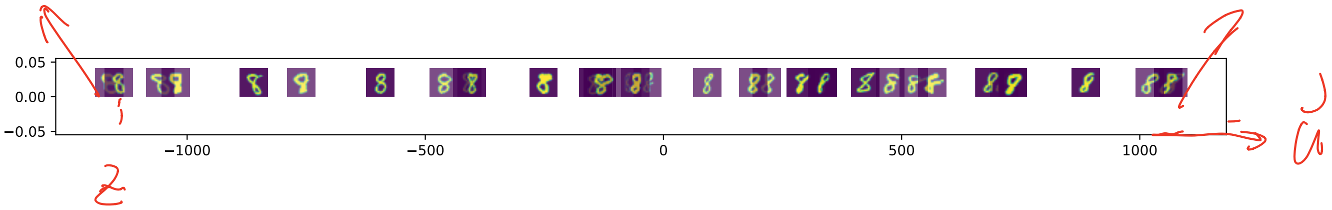
1. Compute the covariance matrix, C.
2. Compute the top eigenvector $\vec{u}$, of C.
3. For $i \in \{1, \ldots, n\}$, create new feature:

$$z^{(i)} = \vec{u} \cdot \vec{x}^{(i)}$$

A Parting Example

- MNIST: 60,000 images in 784 dimensions
- Principal component: $\vec{u} \in \mathbb{R}^{784}$
- We can project an image in $\mathbb{R}^{784}$ onto $\vec{u}$ to get a single number representing the image

Example



DSC 140B
Representation Learning

Lecture 10 | Part 3

Dimensionality Reduction with d ≥ 2

So far: PCA

- **Given**: data $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)} \in \mathbb{R}^d$
- **Map**: each data point $\vec{x}^{(i)}$ to a single feature, z_i.
  - Idea: maximize the variance of the new feature
- **PCA**: Let $z_i = \vec{x}^{(i)} \cdot \vec{u}$, where $\vec{u}$ is top eigenvector of covariance matrix, C.

Now: More PCA

- **Given**: data $\vec{x}^{(1)}, \dots, \vec{x}^{(n)} \in \mathbb{R}^d$
- **Map**: each data point $\vec{x}^{(i)}$ to k new features, $\vec{z}^{(i)} = (z_1^{(i)}, \dots, z_k^{(i)})$.

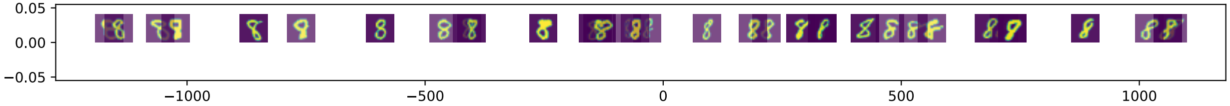
A Single Principal Component

- Recall: the **principal component** is the top eigenvector $\vec{u}$ of the covariance matrix, C
- It is a unit vector in $\mathbb{R}^d$
- Make a new feature $z \in \mathbb{R}$ for point $\vec{x} \in \mathbb{R}^d$ by computing $z = \vec{x} \cdot \vec{u}$
- This is dimensionality reduction from $\mathbb{R}^d \to \mathbb{R}^1$

Example

- MNIST: 60,000 images in 784 dimensions
- Principal component: $\vec{u} \in \mathbb{R}^{784}$
- We can project an image in $\mathbb{R}^{784}$ onto $\vec{u}$ to get a single number representing the image

Example



Another Feature?

- Clearly, mapping from $\mathbb{R}^{784} \to \mathbb{R}^1$ loses a lot of information

- What about mapping from $\mathbb{R}^{784} \to \mathbb{R}^2$? $\mathbb{R}^k$?

A Second Feature

- Our first feature is a mixture of features, with weights given by unit vector $\vec{u}^{(1)} = (u_1^{(1)}, u_2^{(1)}, \ldots, u_d^{(1)})^T$.

$$z_1 = \vec{u}^{(1)} \cdot \vec{x} = u_1^{(1)} x_1 + \ldots + u_d^{(1)} x_d$$

- To maximize variance, choose $\vec{u}^{(1)}$ to be top eigenvector of C.

A Second Feature

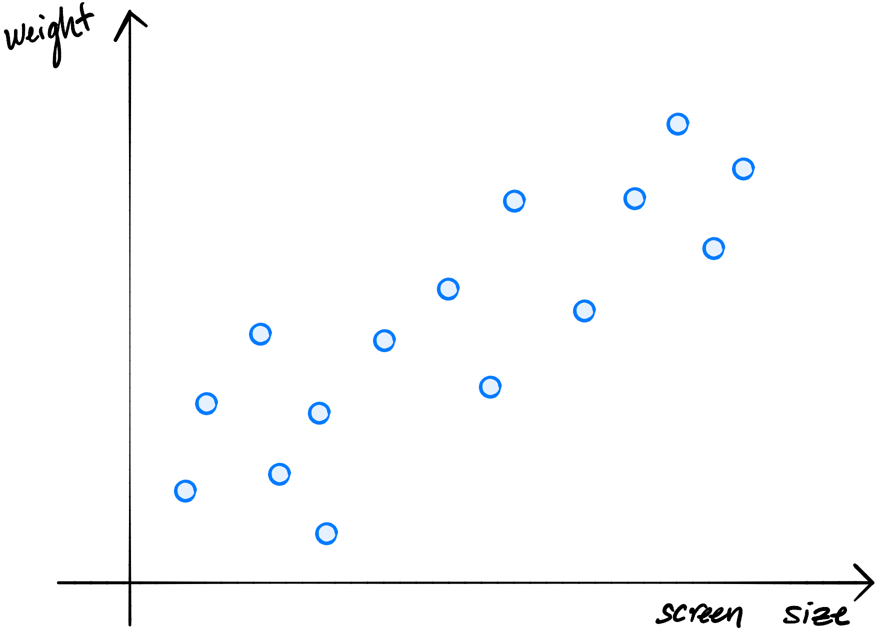
- Make same assumption for second feature:

$$z_2 = \vec{u}^{(2)} \cdot \vec{x} = u_1^{(2)} x_1 + \ldots + u_d^{(2)} x_d$$

- How do we choose $\vec{u}^{(2)}$?

- We should choose $\vec{u}^{(2)}$ to be **orthogonal** to $\vec{u}^{(1)}$.
  - No “redundancy”.

A Second Feature



Intuition

- Claim: if $\vec{u}$ and $\vec{v}$ are eigenvectors of a symmetric matrix with distinct eigenvalues, they are orthogonal.
- We should choose $\vec{u}^{(2)}$ to be an **eigenvector** of the covariance matrix, C.
- The second eigenvector of C is called the **second principal component**.

A Second Principal Component

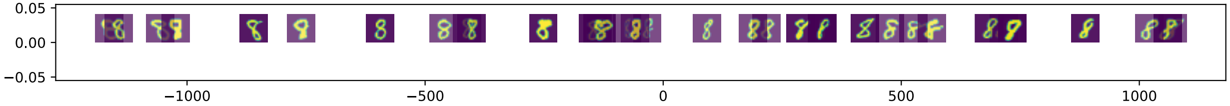
- Given a covariance matrix C.
- The principal component $\vec{u}^{(1)}$ is the top eigenvector of C.
  - Points in the direction of maximum variance.
- The *second* principal component $\vec{u}^{(2)}$ is the *second* eigenvector of C.
  - Out of all vectors orthogonal to the principal component, points in the direction of max variance.

PCA: Two Components

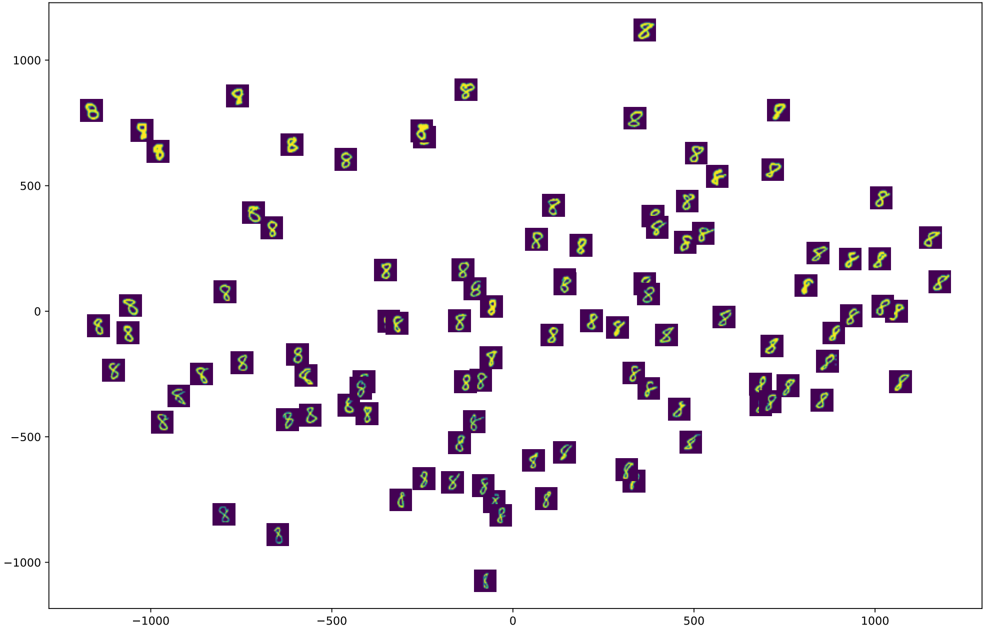
- Given data $\{\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}\} \in \mathbb{R}^d$.
- Compute covariance matrix C, top two eigenvectors $\vec{u}^{(1)}$ and $\vec{u}^{(2)}$.
- For any vector $\vec{x} \in \mathbb{R}$, its new representation in $\mathbb{R}^2$ is $\vec{z} = (z_1, z_2)^T$, where:

$$z_1 = \vec{x} \cdot \vec{u}^{(1)}$$
$$z_2 = \vec{x} \cdot \vec{u}^{(2)}$$

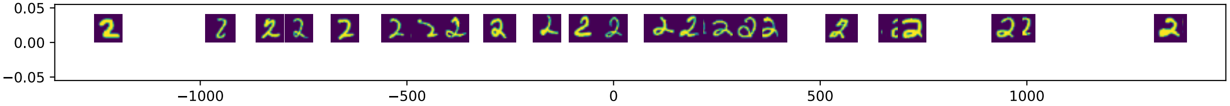
Example



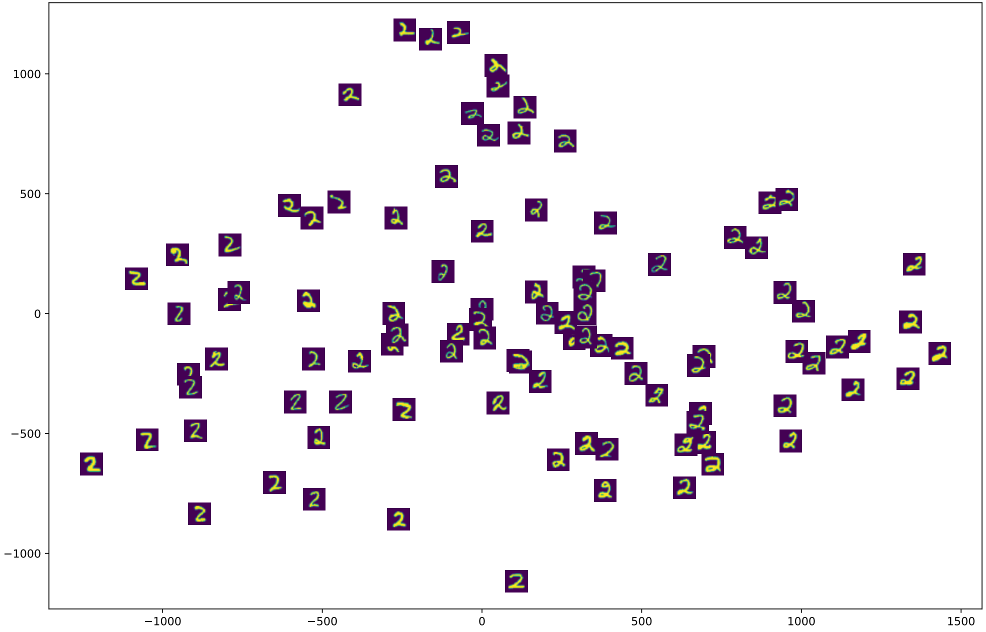
Example



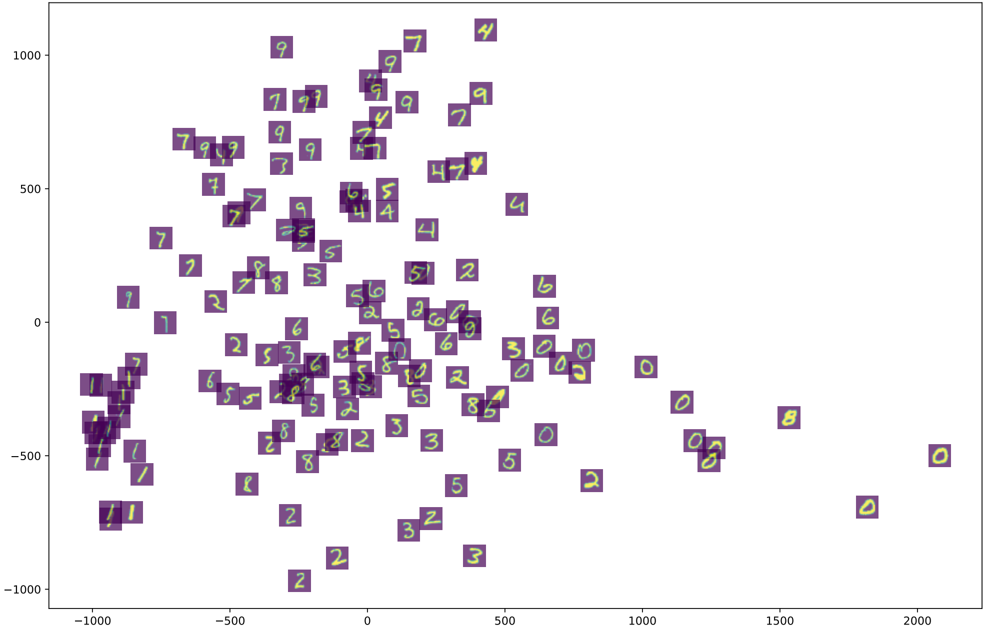
Example



Example



Example



PCA: k Components

- Given data $\{\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}\} \in \mathbb{R}^d$, number of components k.
- Compute covariance matrix C, top $k \leq d$ eigenvectors $\vec{u}^{(1)}$, $\vec{u}^{(2)}$, …, $\vec{u}^{(k)}$.
- For any vector $\vec{x} \in \mathbb{R}$, its new representation in $\mathbb{R}^k$ is $\vec{z} = (z_1, z_2, \ldots z_k)^T$, where:

$$
\begin{aligned}
z_1 &= \vec{x} \cdot \vec{u}^{(1)} \\
z_2 &= \vec{x} \cdot \vec{u}^{(2)} \\
&\vdots \\
z_k &= \vec{x} \cdot \vec{u}^{(k)}
\end{aligned}
$$

Matrix Formulation

- Let X be the **data matrix** (n rows, d columns)
- Let U be matrix of the k eigenvectors as columns (d rows, k columns)
- The new representation: $Z = XU$