

DSC 140B

Representation Learning

Lecture 09 | Part 1

Covariance Matrices

Variance

- ▶ We know how to compute the variance of a set of numbers $X = \{x^{(1)}, \dots, x^{(n)}\}$:

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu)^2$$

- ▶ The variance measures the “spread” of the data

Generalizing Variance

- ▶ If we have two features, x_1 and x_2 , we can compute the variance of each as usual:

$$\text{Var}(x_1) = \frac{1}{n} \sum_{i=1}^n (\vec{x}_1^{(i)} - \mu_1)^2$$

$$\text{Var}(x_2) = \frac{1}{n} \sum_{i=1}^n (\vec{x}_2^{(i)} - \mu_2)^2$$

- ▶ Can also measure how x_1 and x_2 vary together.

Measuring Similar Information

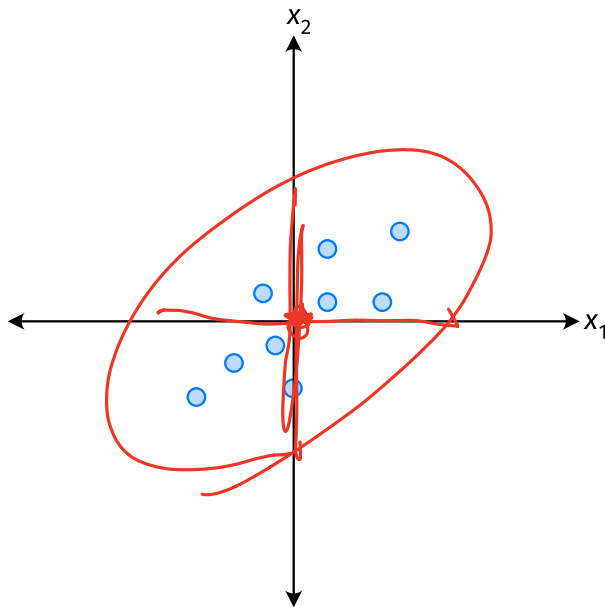
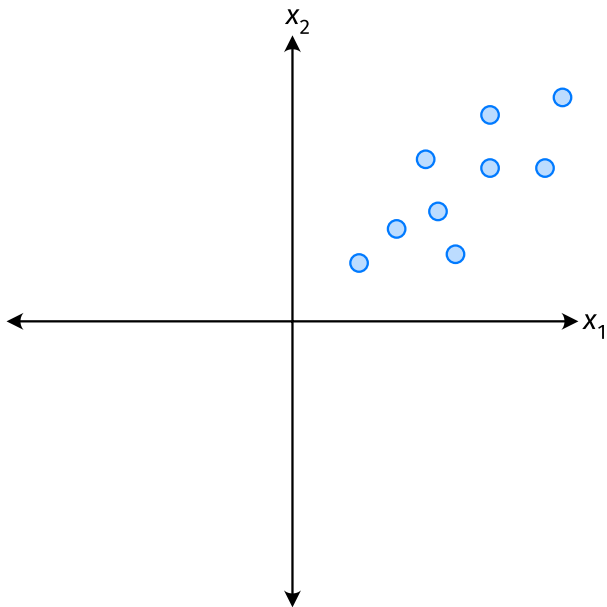
- ▶ Features which share information if they *vary together*.
 - ▶ A.k.a., they “co-vary”
- ▶ Positive association: when one is above average, so is the other
- ▶ Negative association: when one is above average, the other is below average

Examples

- ▶ Positive: temperature and ice cream cones sold.
- ▶ Positive: temperature and shark attacks.
- ▶ Negative: temperature and coats sold.

Centering

- First, it will be useful to **center** the data.



Centering

- ▶ Compute the mean of each feature:

$$\mu_j = \frac{1}{n} \sum_1^n \vec{x}_j^{(i)}$$

- ▶ Define new centered data:

$$\vec{z}^{(i)} = \begin{pmatrix} \vec{x}_1^{(i)} - \mu_1 \\ \vec{x}_2^{(i)} - \mu_2 \\ \vdots \\ \vec{x}_d^{(i)} - \mu_d \end{pmatrix}$$

Centering (Equivalently)

- ▶ Compute the mean of all data points:

$$\mu = \frac{1}{n} \sum_1^n \vec{x}^{(i)}$$

- ▶ Define new centered data:

$$\vec{z}^{(i)} = \vec{x}^{(i)} - \mu$$

Exercise

Center the data set:

$$\vec{x}^{(1)} = (1, 2, 3)^T$$

$$\vec{x}^{(2)} = (-1, -1, 0)^T$$

$$\vec{x}^{(3)} = (0, 2, 3)^T$$

Quantifying Co-Variance

- ▶ One approach is as follows¹.

$$\text{Cov}(x_i, x_j) = \frac{1}{n} \sum_{k=1}^n \vec{x}_i^{(k)} \vec{x}_j^{(k)}$$

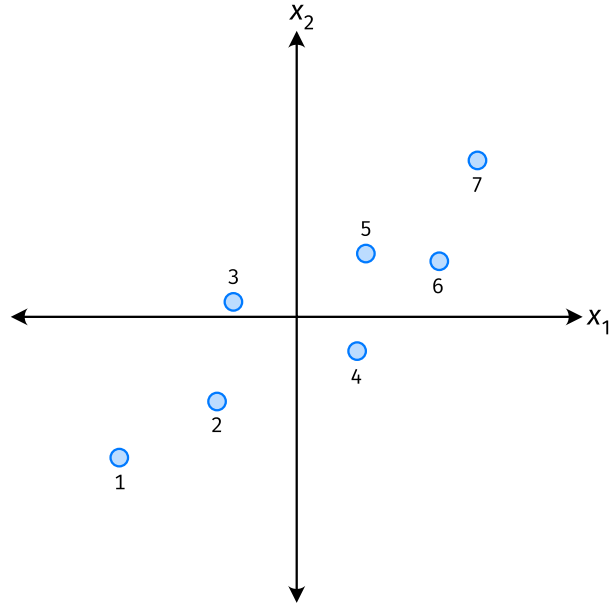
- ▶ For each data point, multiply the value of feature i and feature j , then average these products.
- ▶ This is the **covariance** of features i and j .

¹Assuming centered data

Quantifying Covariance

- ▶ Assume the data are **centered**.

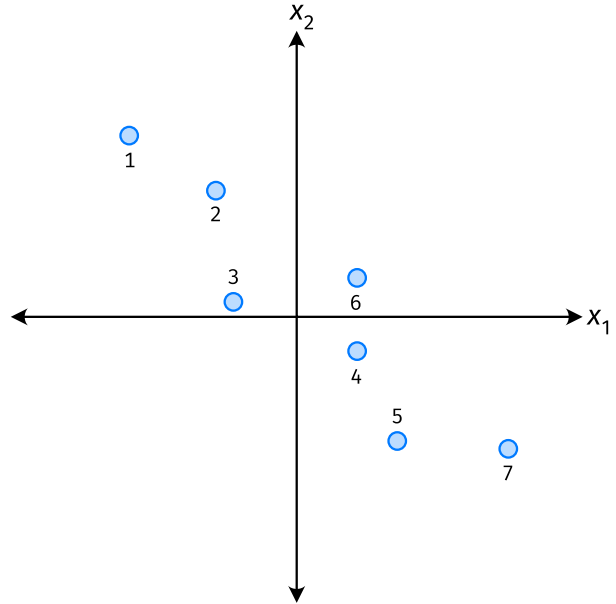
$$\text{Covariance} = \frac{1}{7} \sum_{i=1}^7 \vec{X}_1^{(i)} \times \vec{X}_2^{(i)}$$



Quantifying Covariance

- ▶ Assume the data are **centered**.

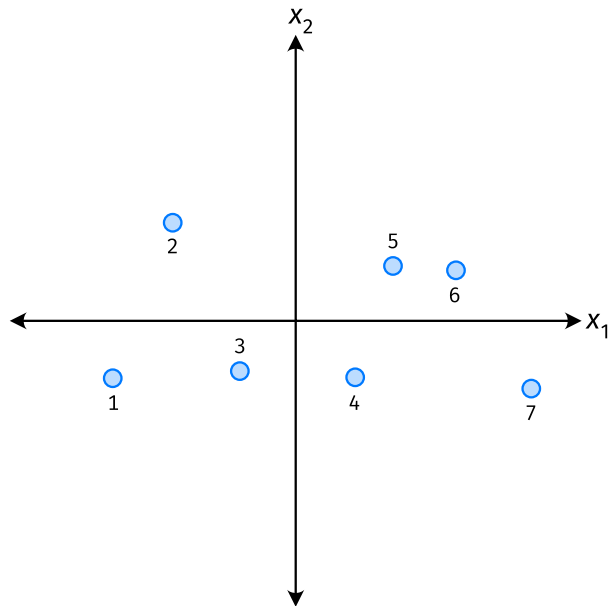
$$\text{Covariance} = \frac{1}{7} \sum_{i=1}^7 \vec{X}_1^{(i)} \times \vec{X}_2^{(i)}$$



Quantifying Covariance

- ▶ Assume the data are **centered**.

$$\text{Covariance} = \frac{1}{7} \sum_{i=1}^7 \vec{X}_1^{(i)} \times \vec{X}_2^{(i)}$$



Quantifying Covariance

- ▶ The **covariance** quantifies extent to which two variables vary together.
- ▶ Assume we have centered the data.
- ▶ The **sample covariance** of feature i and j is:

$$\sigma_{ij} = \frac{1}{n} \sum_{k=1}^n \vec{x}_i^{(k)} \vec{x}_j^{(k)}$$

Exercise

True or False: $\sigma_{ij} = \sigma_{ji}$?

$$\sigma_{ij} = \frac{1}{n} \sum_{k=1}^n \vec{X}_i^{(k)} \vec{X}_j^{(k)}$$

$$\begin{aligned} \sigma_{ij} &= \frac{1}{n} \sum_{k=1}^n X_j^{(k)} X_i^{(k)} \\ &= \sigma_{ji} \end{aligned}$$

$i=j$ $\sigma_{ii} = \frac{1}{n} \sum_{k=1}^n \frac{\vec{x}_i^{(k)} \vec{x}_i^{(k)}}{\|\vec{x}_i^{(k)}\|^2}$ centered. $\vec{\mu} = \vec{0}$

Covariance Matrices

$\Rightarrow \frac{1}{n} \sum_{k=1}^n \vec{x}_i^{(k)} \vec{x}_i^{(k)T}$

- ▶ Given data $\vec{x}^{(1)}, \dots, \vec{x}^{(n)} \in \mathbb{R}^d$.
- ▶ The **sample covariance matrix** C is the $d \times d$ matrix whose ij entry is defined to be σ_{ij} .

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \\ & \dots \\ & & \sigma_{dd} \end{bmatrix}$$

$$\sigma_{ij} = \frac{1}{n} \sum_{k=1}^n \vec{x}_i^{(k)} \vec{x}_j^{(k)}$$

Observations

- ▶ Diagonal entries of C are the variances.
- ▶ The matrix is **symmetric!**

$$\sigma_{ij} = \sigma_{ji}$$

Note

- ▶ Sometimes you'll see the sample covariance defined as:

$$\sigma_{ij} = \frac{1}{n-1} \sum_{k=1}^n \vec{X}_i^{(k)} \vec{X}_j^{(k)}$$

Note the $1/(n-1)$

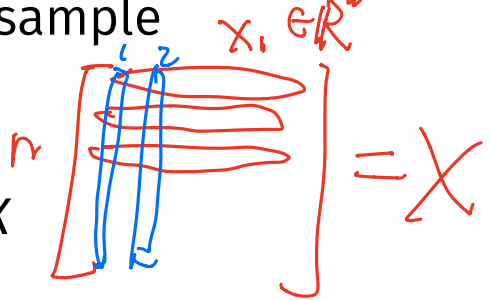
- ▶ This is an **unbiased** estimator of the population covariance.
- ▶ Our definition is the **maximum likelihood** estimator.
- ▶ In practice, it doesn't matter: $1/(n-1) \approx 1/n$.
- ▶ For consistency, in this class use $1/n$.

Computing Covariance

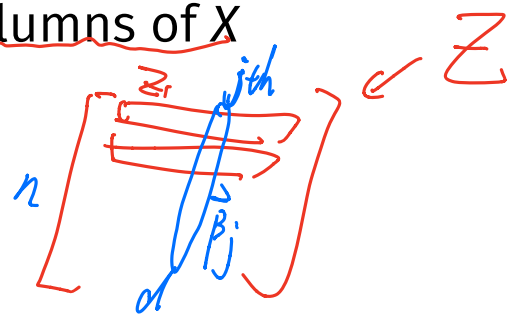
$$C_{ij} = \frac{1}{n} \sum_{k=1}^n z_i^{(k)} \cdot z_j^{(k)}$$

- ▶ There is a “trick” for computing sample covariance matrices.

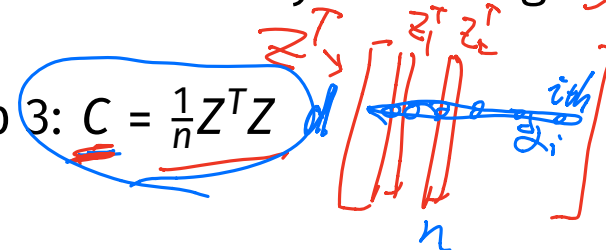
- ▶ Step 1: make $n \times d$ data matrix, X



- ▶ Step 2: make Z by centering columns of X



- ▶ Step 3: $C = \frac{1}{n} Z^T Z$



Computing Covariance (in code)²

```
»» mu = X.mean(axis=0)
»» Z = X - mu
»» C = 1 / len(X) * Z.T @ Z
```

$\frac{1}{n}$ $Z^T \cdot Z$

²Or use np.cov ✓

DSC 140B

Representation Learning

Lecture 09 | Part 2

Visualizing Covariance Matrices

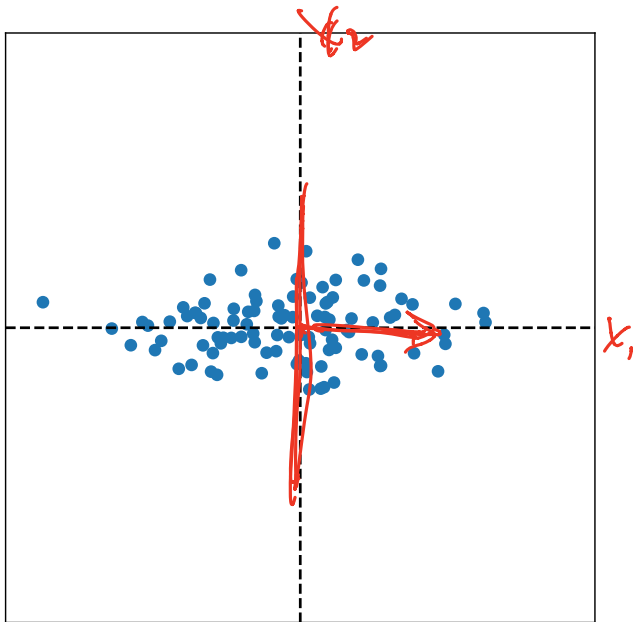
Visualizing Covariance Matrices

- ▶ Covariance matrices are symmetric.

$$\sigma_{ij} = \sigma_{ji}$$

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

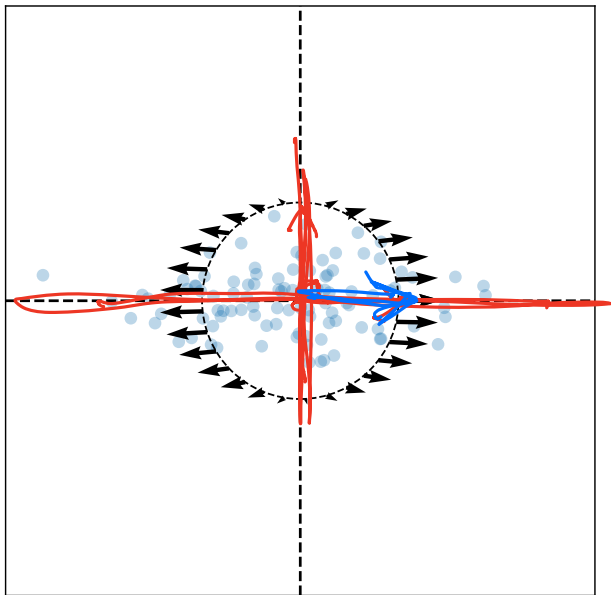
Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} C_{11} & C_{12} \\ \textcircled{10} & 0 \\ 0 & \textcircled{5} \\ & C_{22} \end{pmatrix}$$

vector field

Visualizing Covariance Matrices



$$C = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$$

$$C \vec{u}^{(1)} = 1 \vec{u}^{(1)}$$
$$C \vec{u}^{(2)} = 5 \vec{u}^{(2)}$$

Eigenvectors:

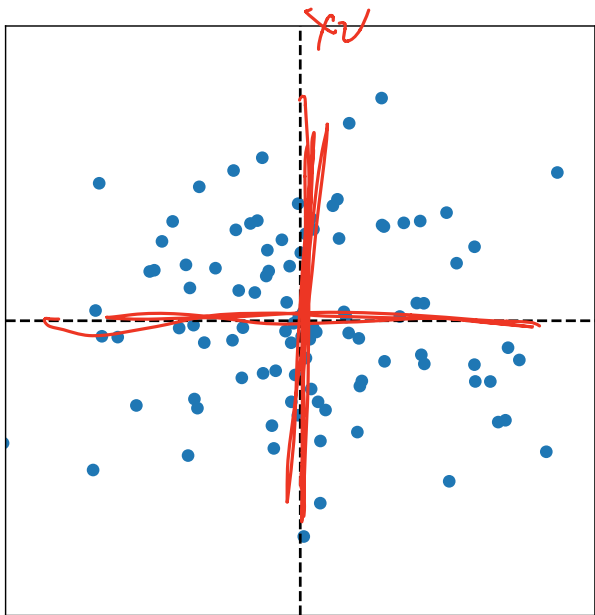
$$\vec{u}^{(1)} \approx (1, 0)$$

$$\vec{u}^{(2)} \approx (0, 1)$$

$$\lambda_1 \approx 1$$

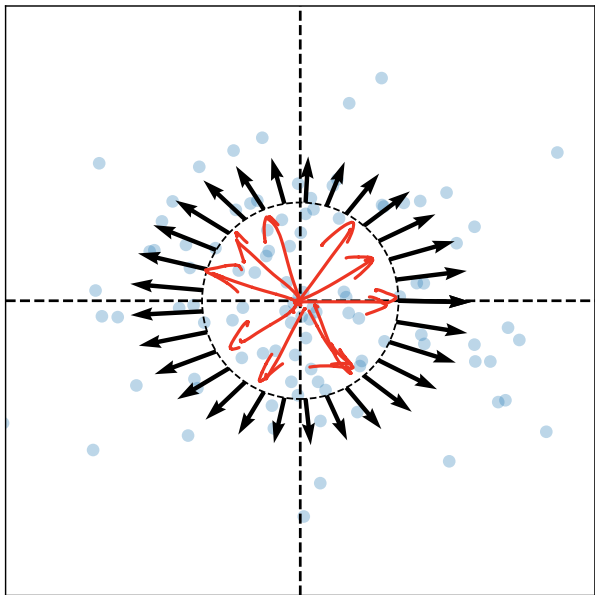
$$\lambda_2 \approx 5$$

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$$

Visualizing Covariance Matrices

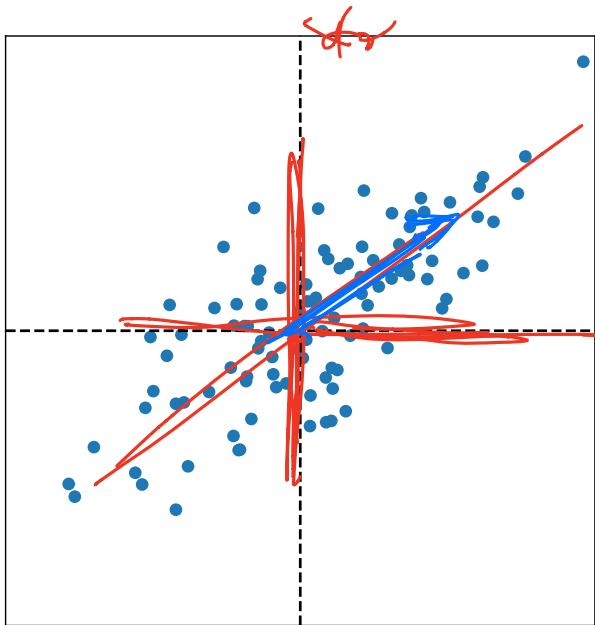


$$C = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

Eigenvectors:

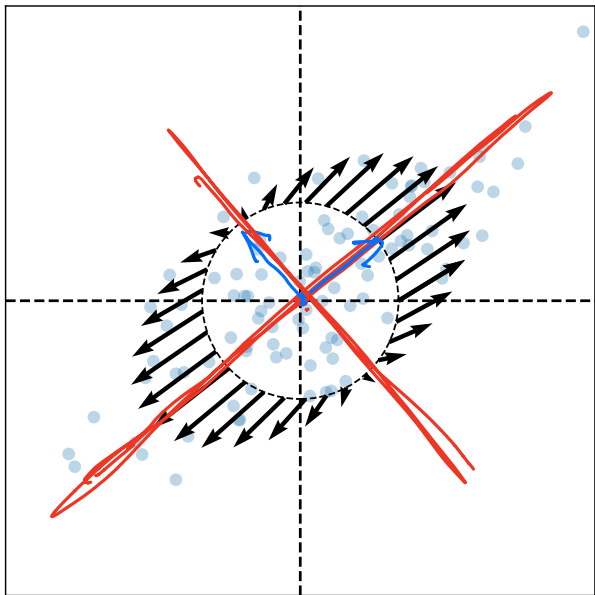
$$\vec{u}^{(1)} \approx \frac{(\sigma^2, 1)}{\sigma^2 + 1} \circlearrowleft$$
$$\vec{u}^{(2)} \approx \frac{(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})}{\sqrt{\frac{1}{\sigma^2} + \frac{1}{\sigma^2}}}$$

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$$

Visualizing Covariance Matrices



$$C = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$$

Eigenvectors:

$$\vec{u}^{(1)} \approx \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$

$$\vec{u}^{(2)} \approx \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$

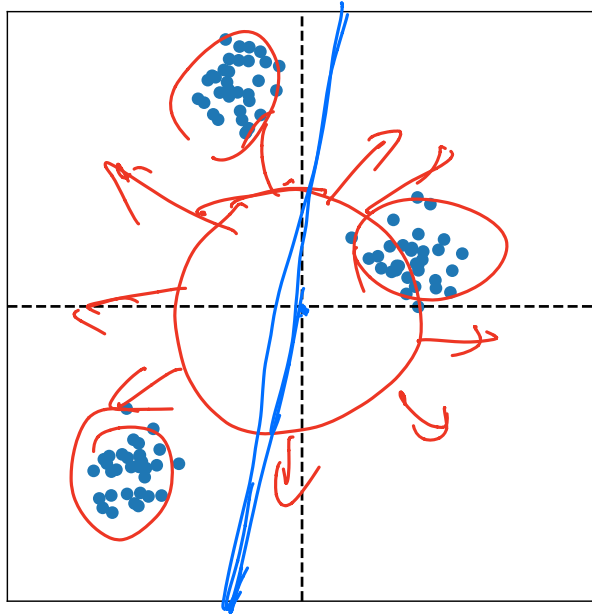
Intuitions

- ▶ 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

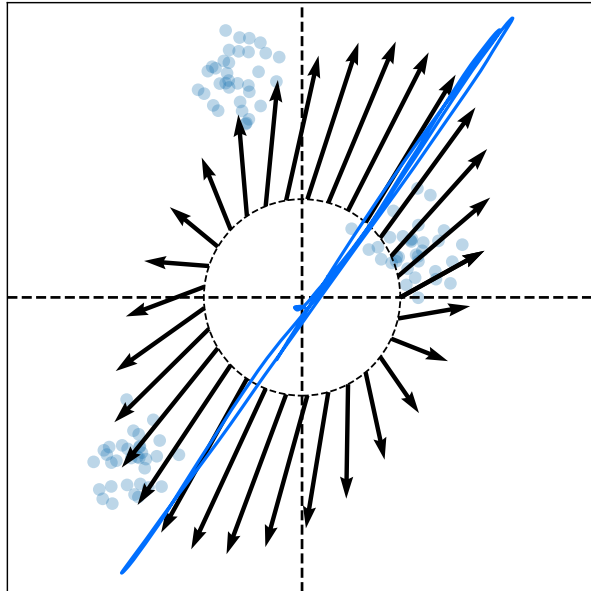
- ▶ The data doesn't always look like this.
- ▶ We can always compute covariance matrices.
- ▶ They just may not describe the data's shape very well.

C



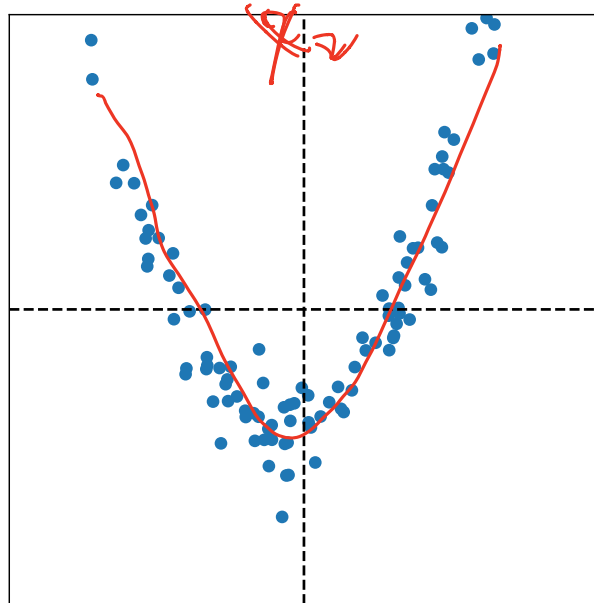
Caution

- ▶ 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.



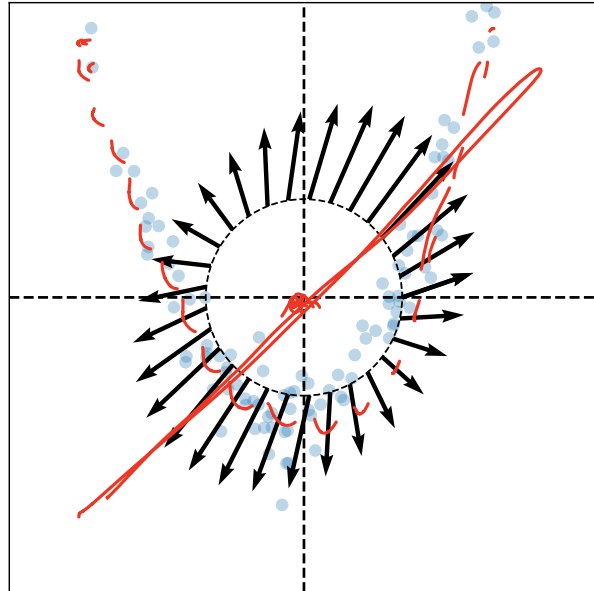
Caution

- ▶ 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 09 | Part 3

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...

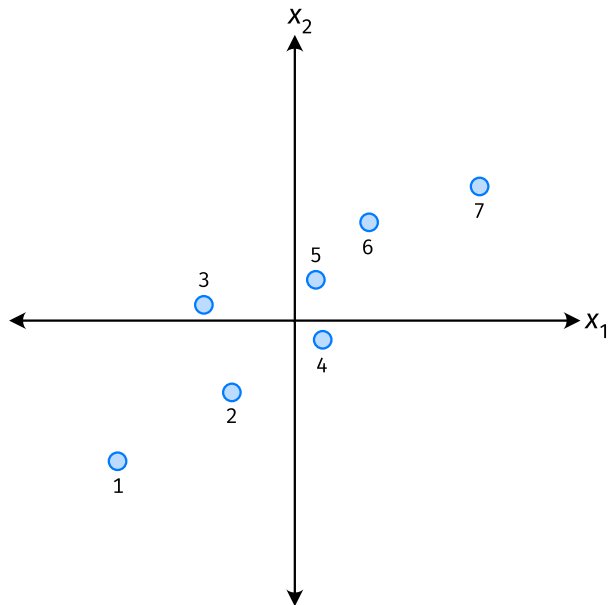
- ▶ We haven't actually defined "direction of maximum variance"
- ▶ Let's derive PCA more formally.

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:

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

Example



Note

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

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

Goal

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

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

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

Claim

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

Our Goal (Again)

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

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 $\vec{x}^{(1)}, \dots, \vec{x}^{(n)} \in \mathbb{R}^d$
- 1. Compute the covariance matrix, C .
- 2. Compute the top eigenvector \vec{u} , of C .
- 3. For $i \in \{1, \dots, n\}$, 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

