DSC 140B

Representation Learning

Lecture 06 | Part 1

Dimensionality Reduction

Choosing $\vec{u}$

$\|\vec{u}\| = 1$

$\vec{u} \cdot \vec{u} = 1$

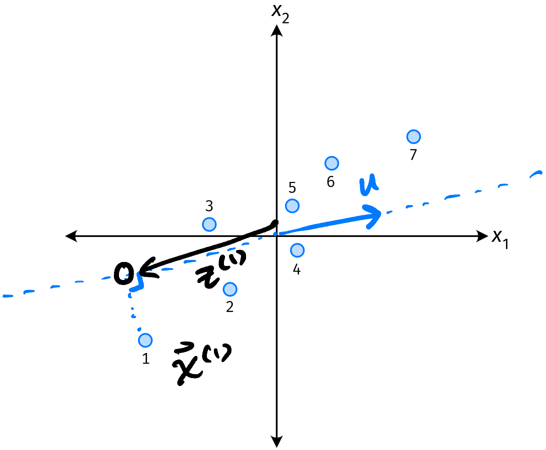
$u_1^2 + u_2^2 = 1$

- Suppose we have only two features:
  - x_1: screen size
  - x_2: phone thickness
- We'll create single new feature, z, from x_1 and x_2.
  - Assume $z = u_1 x_1 + u_2 x_2 = \vec{x} \cdot \vec{u}$
  - Interpretation: z is a measure of a phone's size
- How should we choose $\vec{u} = (u_1, u_2)^T$?

Visualization

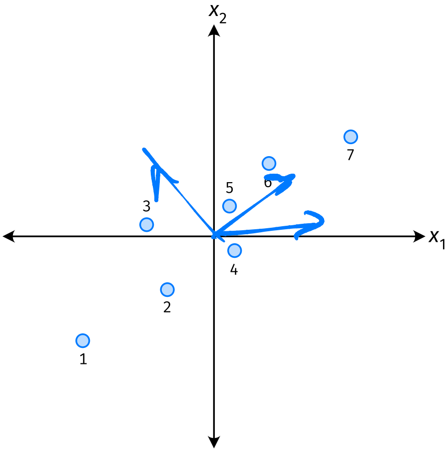
http://dsc140b.com/static/vis/pca-max_variance/

Example



- $\vec{u}$ defines a direction
- $\vec{z}^{(i)} = \vec{x}^{(i)} \cdot \vec{u}$ measures position of $\vec{x}$ along this direction

Example

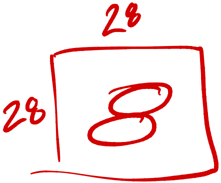


- Phone "size" varies most along a diagonal direction.
- Along direction of "max variance", phones are well-separated.
- **Idea:** $\vec{u}$ should point in direction of "max variance".

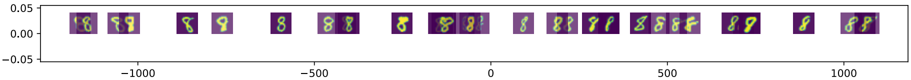
Our Algorithm (Informally)

- **Given**: data points $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)} \in \mathbb{R}^d$
- Pick $\vec{u}$ to be the direction of "max variance"
- Create a new feature, z, for each point:

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



Example



PCA

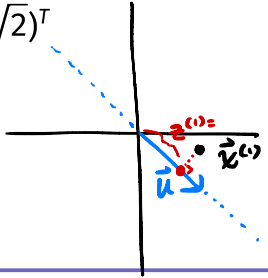
- This algorithm is called **Principal Component Analysis**, or **PCA**.
- The direction of maximum variance is called the **principal component**.

$$z^{(2)} = \frac{-3}{\sqrt{2}}$$

Exercise

Suppose the direction of maximum variance in a data set is

$$\vec{u} = (1/\sqrt{2}, -1/\sqrt{2})^T$$



Let

- $\vec{x}^{(1)} = (3, -2)^T$
- $\vec{x}^{(2)} = (1, 4)^T$

What are $z^{(1)}$ and $z^{(2)}$?

$$z^{(1)} = \vec{x}^{(1)} \cdot \vec{u} = \begin{pmatrix} 3 \\ -2 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} = \frac{3}{\sqrt{2}} + \frac{2}{\sqrt{2}} = \frac{5}{\sqrt{2}}$$

Problem

- How do we compute the “direction of maximum variance”?

DSC 140B

Representation Learning

Lecture 06 | Part 2

Covariance Matrices

Variance

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

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

- 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

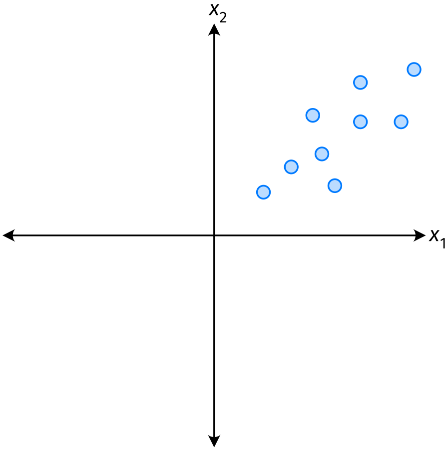
- Negative association: when one is above average, the other is below average

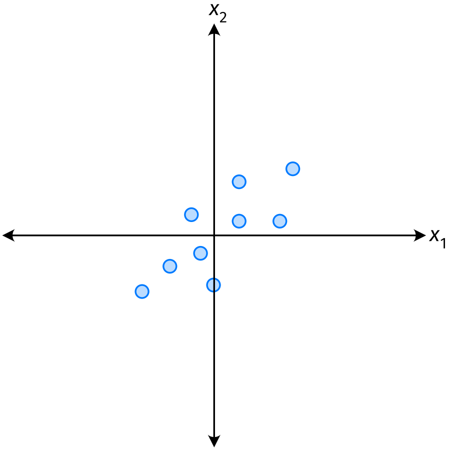
Examples

- 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:

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

- Define new centered data:

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

$$\vec{\mu} = \frac{1}{3}\begin{pmatrix}0\\3\\6\end{pmatrix} = \begin{pmatrix}0\\1\\2\end{pmatrix} \qquad \vec{z}^{(2)} = \begin{pmatrix}-1\\-1\\0\end{pmatrix} - \begin{pmatrix}0\\1\\2\end{pmatrix} = \begin{pmatrix}-1\\-2\\-2\end{pmatrix}$$

Exercise

Center the data set:

$$\vec{x}^{(1)} = (1, 2, 3)^T$$
$$\vec{x}^{(2)} = (-1, -1, 0)^T$$
$$\vec{x}^{(3)} = (0, 2, 3)^T$$

$$\vec{z}^{(1)} = \vec{x}^{(1)} - \mu = \begin{pmatrix}1\\2\\3\end{pmatrix} - \begin{pmatrix}0\\1\\2\end{pmatrix} = \begin{pmatrix}1\\1\\1\end{pmatrix}$$

Quantifying Co-Variance

- One approach is as follows[1].

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

  - For each data point, multiply the value of feature i and feature j, then average these products.

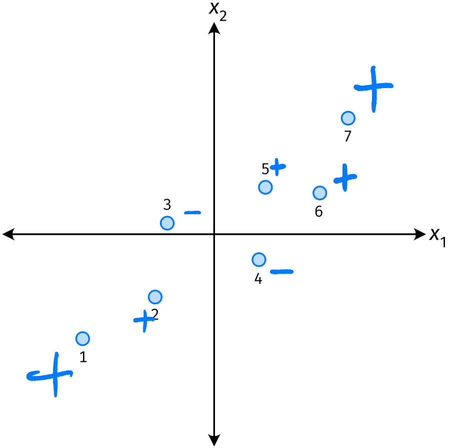
  - This is the **covariance** of features i and j.

[1] Assuming centered data

Quantifying Covariance

- Assume the data are **centered**.

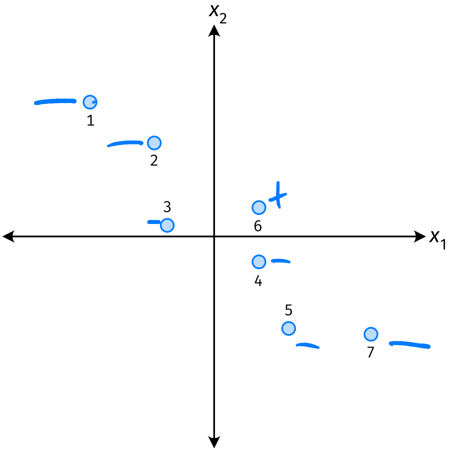
$$\text{Covariance} = \frac{1}{7}\sum_{i=1}^{7} \vec{x}_1^{(i)} \times \vec{x}_2^{(i)}$$



Quantifying Covariance

- Assume the data are **centered**.

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

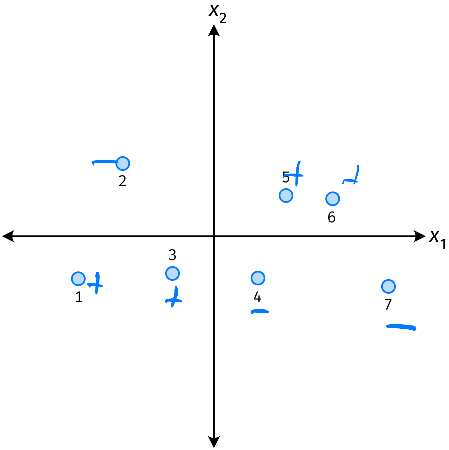


Quantifying Covariance

- Assume the data are **centered**.

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

≈ 0



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:

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

$3 \times 5 = 5 \times 3$

Exercise

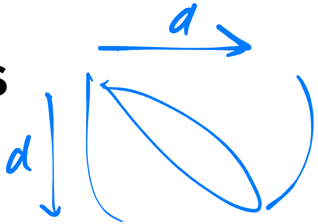
True or False: $\sigma_{ij} = \sigma_{ji}$?

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

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

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

Covariance Matrices



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

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

$$\sigma_{ii} = \frac{1}{n} \sum (x_i^{(k)} - \mu)^2 \qquad \mu \to 0$$

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 $1/n$.

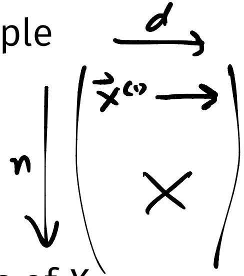
Computing Covariance

- 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
```

[2] Or use `np.cov`

DSC 140B
Representation Learning

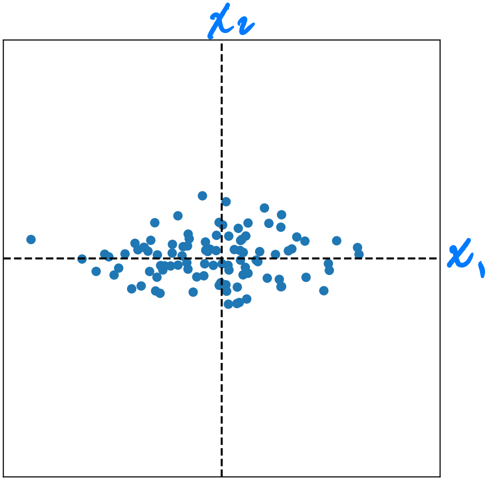
Lecture 06 | Part 3

Visualizing Covariance Matrices

Visualizing Covariance Matrices

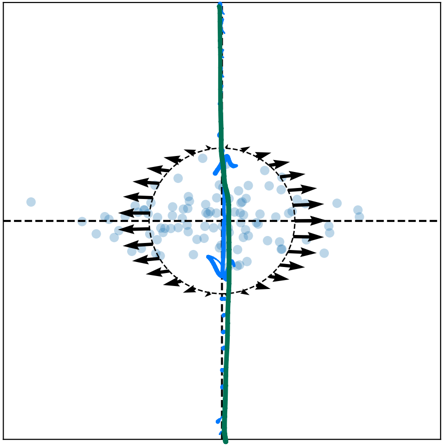
- Covariance matrices are symmetric.
- They have axes of symmetry (eigenvectors and eigenvalues).
- What are they?

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 10 & \approx 0 \\ \approx 0 & 5 \end{pmatrix}$$

Visualizing Covariance Matrices



$\lambda_1 > \lambda_2$

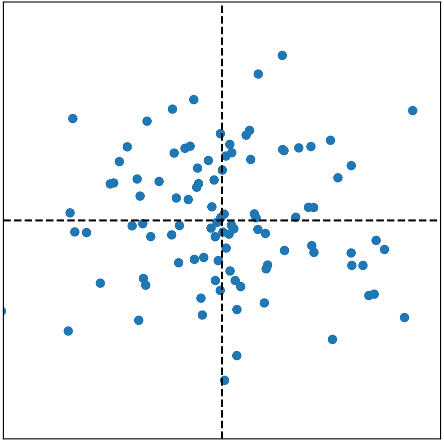
Eigenvectors:

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

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

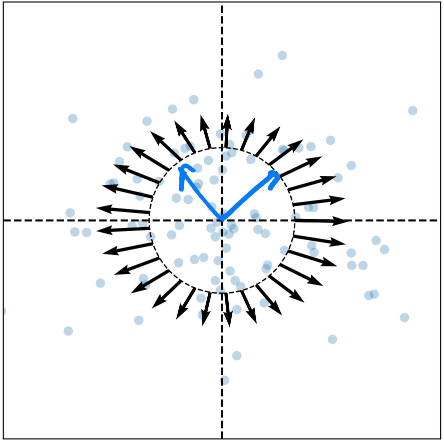
$(0, -1)$

Visualizing Covariance Matrices



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

Visualizing Covariance Matrices

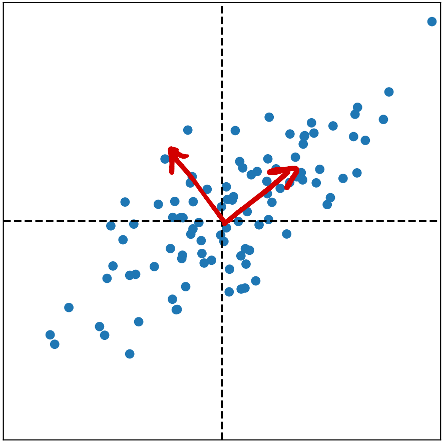


Eigenvectors:

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

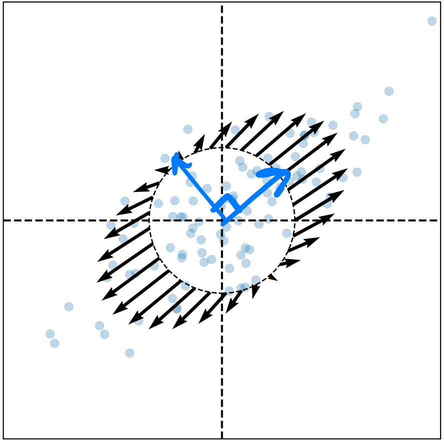
$\vec{u}^{(2)} \approx$

Visualizing Covariance Matrices



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

Visualizing Covariance Matrices



$\lambda_1 > \lambda_2$

Eigenvectors:

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

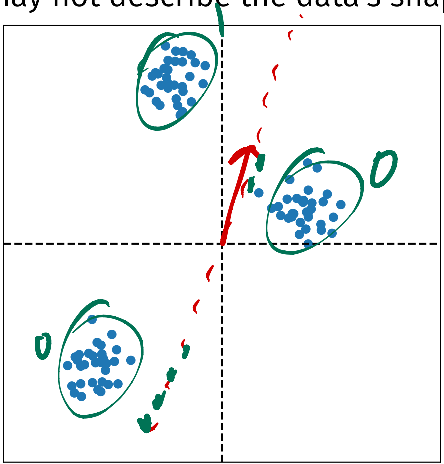
$\vec{u}^{(2)} \approx (-1, 1)$

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.

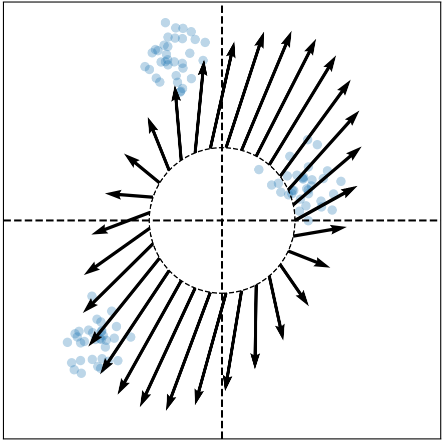
Caution

- 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.



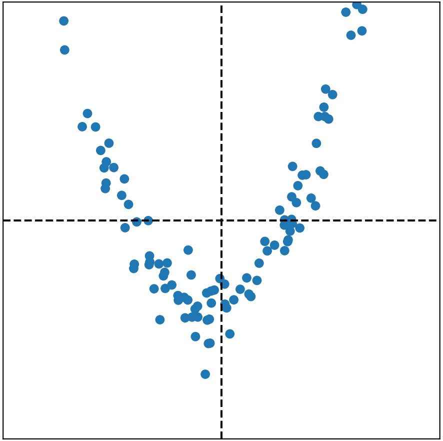
Caution

- 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.



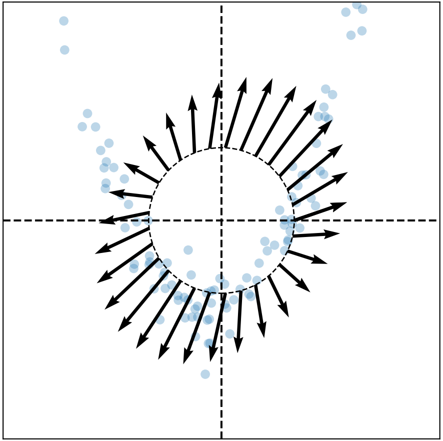
Caution

- 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 06 | Part 4

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...

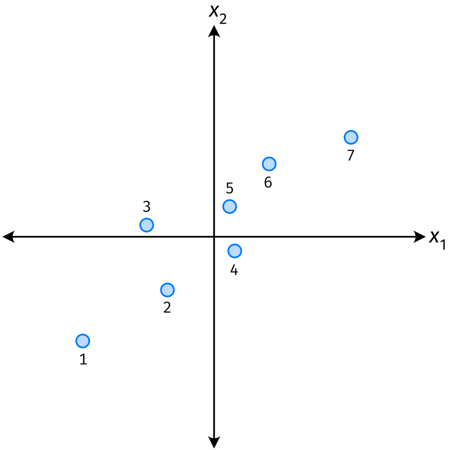
- We haven’t actually defined “direction of maximum variance”
- Let’s derive PCA more formally.

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:

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

Goal

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

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

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

Claim

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

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 $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)} \in \mathbb{R}^d$

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