DSC 140B
Representation Learning

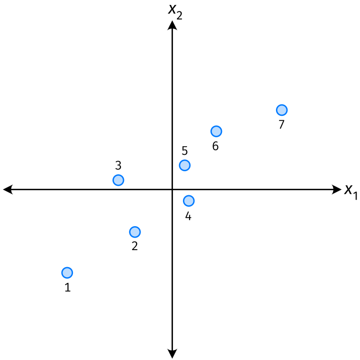
Lecture 06 | Part 1

Dimensionality Reduction

Last Time: Dimensionality Reduction

- **Given:** data points $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)} \in \mathbb{R}^d$
- **Goal:** create a new, lower-dimensional data set without losing too much useful information
- For now, focus on reducing to just one dimension.

Example



- Each point is a phone.
- $\vec{x} = (\text{width}, \text{weight})^T$.
- Can we reduce $\vec{x}$ to a single feature, z, without losing too much information?

The Idea from Last Time

- Our new feature should be a "mixture" of the old features:

$$
\begin{aligned}
z &= u_1 \times \text{width} + u_2 \times \text{weight} \\
&= u_1 x_1 + u_2 x_2 \\
&= \vec{u} \cdot \vec{x}
\end{aligned}
$$

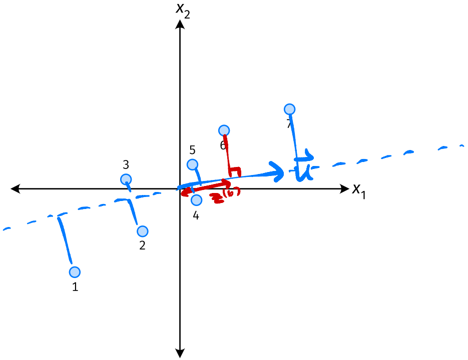
~~$|u_1| + |u_2| = 1$~~

- We get to choose $\vec{u} = (u_1, u_2)^T$.

- Constraint: $\|\vec{u}\| = 1$.

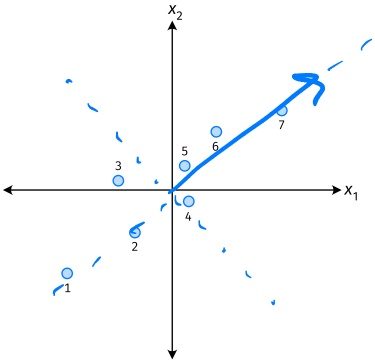
$u_1^2 + u_2^2 = 1$

Geometrically



- $\vec{u}$ defines a direction in $\mathbb{R}^2$.
- z is the projection of $\vec{x}$ onto that direction.
- Which direction should we pick?
  - Concluded: direction of max variance.

Another View



- Our data came to us in the standard basis.
- If we could pick a better basis, what would be our first basis vector?

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:

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

PCA

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

Exercise

Suppose the direction of maximum variance in a data set is

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

Let $\vec{x}^{(1)} = (3, -2)^T$ and $\vec{x}^{(2)} = (1, 4)^T$. What are $z^{(1)}$ and $z^{(2)}$?

A) $z^{(1)} = \frac{1}{\sqrt{2}}$, $z^{(2)} = \frac{3}{\sqrt{2}}$

B) $z^{(1)} = \frac{5}{\sqrt{2}}$, $z^{(2)} = \frac{3}{\sqrt{2}}$

C) $z^{(1)} = \frac{5}{\sqrt{2}}$, $z^{(2)} = \frac{-3}{\sqrt{2}}$

Live Q&A

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

Exercise

Suppose the direction of maximum variance in a data set is

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

Let $\vec{x}^{(1)} = (3, -2)^T$ and $\vec{x}^{(2)} = (1, 4)^T$.
What are $z^{(1)}$ and $z^{(2)}$?

A) $z^{(1)} = \frac{1}{\sqrt{2}}$, $z^{(2)} = \frac{-3}{\sqrt{2}}$

B) $z^{(1)} = \frac{5}{\sqrt{2}}$, $z^{(2)} = \frac{3}{\sqrt{2}}$

C) $z^{(1)} = \frac{5}{\sqrt{2}}$, $z^{(2)} = \frac{-3}{\sqrt{2}}$

$$\vec{z}^{(2)} = \vec{x}^{(2)} \cdot \vec{u} = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} - \frac{4}{\sqrt{2}} = \frac{-3}{\sqrt{2}}$$

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 $X = \{x^{(1)}, \ldots, x^{(n)}\}$:

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

- 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:

feature 1 for phone i

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

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

- 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.

Quantifying Co-Variance

- One approach is as follows:

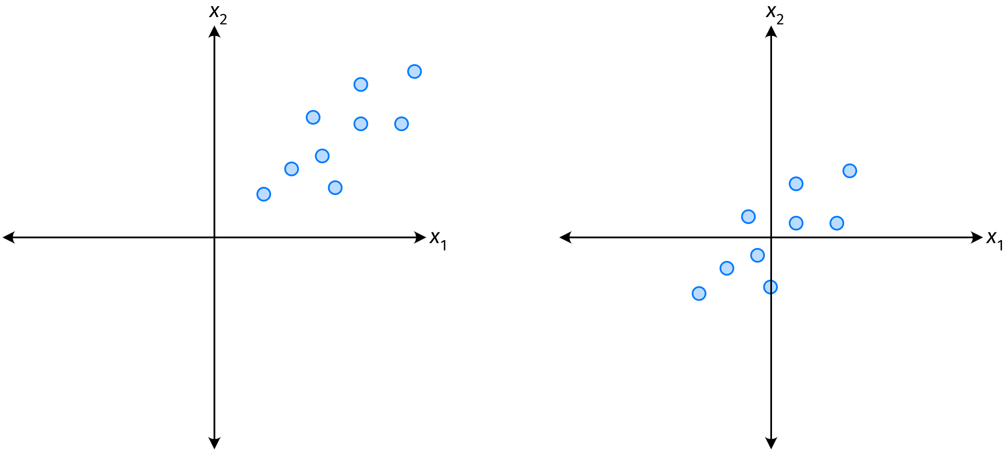
$$\text{Cov}(x_i, x_j) = \frac{1}{n} \sum_{k=1}^{n} (\vec{x}_i^{(k)} - \mu_i)(\vec{x}_j^{(k)} - \mu_j)$$

  - 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.

Centering

- We often **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:

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

Centering (Equivalently)

- Compute the mean of all data points:

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

- Define new centered data:

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

Exercise

Center the data set:

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

Covariance (Again)

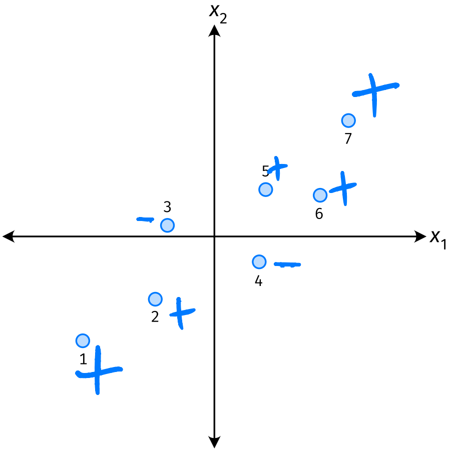
- If the data are **centered**, covariance is:

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

Quantifying Covariance

- Assume the data are **centered**.

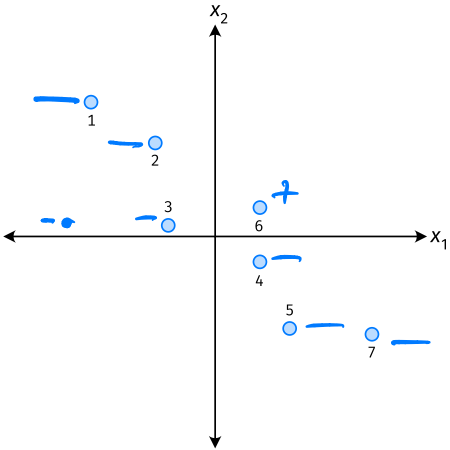
$$\text{Covariance} = \frac{1}{7}\sum_{i=1}^{7} \vec{x}_1^{(i)} \times \vec{x}_2^{(i)}$$



Quantifying Covariance

- Assume the data are **centered**.

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

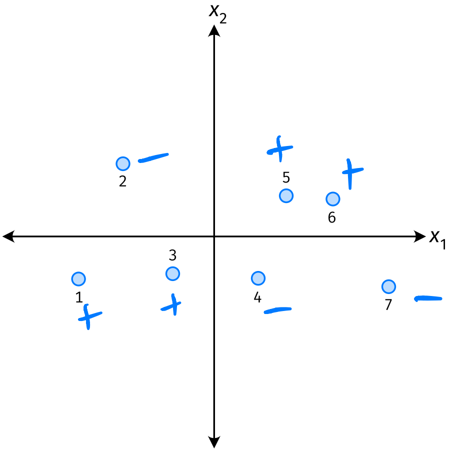


Quantifying Covariance

- 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:

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

Exercise

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

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

Covariance Matrices

$$\sigma_{11} = \frac{1}{n} \sum \vec{x}_1^{(k)} \vec{x}_1^{(k)} = \frac{1}{n} \sum \left(\vec{x}_1^{(k)}\right)^2$$

- 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}.

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

* Assuming we have centered!

Observations

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

Note

- Sometimes you'll see the sample covariance defined with $1/(n-1)$ instead of $1/n$:

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

- 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 $1/n$.

Exercise

Which of the following could be the covariance matrix for the data shown below?

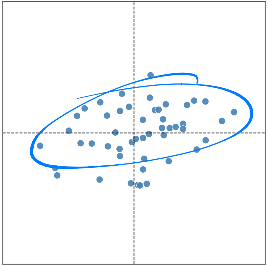
Var x_1

A) $\begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$ Var x_2

B) $\begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix}$

C) $\begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$

D) $\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$



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)[1]

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

[1] Or use `np.cov`

Meaning of the Covariance Matrix

- On the one hand, C is just a table of numbers.
- But remember: every matrix represents a linear transformation.
- What linear transformation does C represent?

Meaning of the Covariance Matrix

- Suppose $\vec{u}$ is a unit vector listing our "mixture coefficients":

$$z = u_1 x_1 + u_2 x_2 + \cdots + u_d x_d$$

- $C\vec{u}$ computes the covariances of the new feature z with each of the original features, $x_1, \ldots, x_d$:

$$C\vec{u} = \left(\text{Cov}(z, x_1),\ \text{Cov}(z, x_2),\ \ldots,\ \text{Cov}(z, x_d)\right)^T$$

- We'd like each to be large.
  - Then, the new feature would be highly correlated with the original features.

Intuition

- $\|C\vec{u}\|$ is large when the new feature $z = \vec{u} \cdot \vec{x}$ is **highly correlated** with the original features.

Intuition

- $\|C\vec{u}\|$ is large when the new feature $z = \vec{u} \cdot \vec{x}$ is **highly correlated** with the original features.
- That is, when z contains a lot of the same information.

Intuition

- $\|C\vec{u}\|$ is large when the new feature $z = \vec{u} \cdot \vec{x}$ is **highly correlated** with the original features.
- That is, when z contains a lot of the same information.
- To maximize this correlation, we want to find $\vec{u}$ which maximizes $\|C\vec{u}\|$.

DSC 140B
Representation Learning

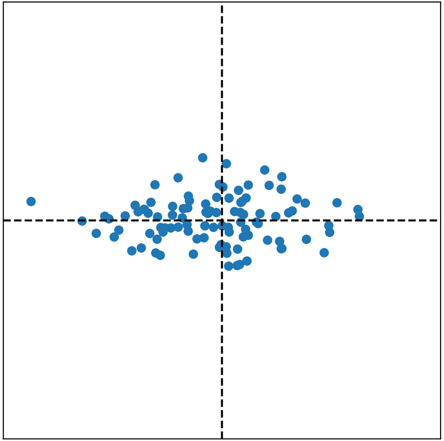
Lecture 06 | Part 3

Visualizing Covariance Matrices

Visualizing Covariance Matrices

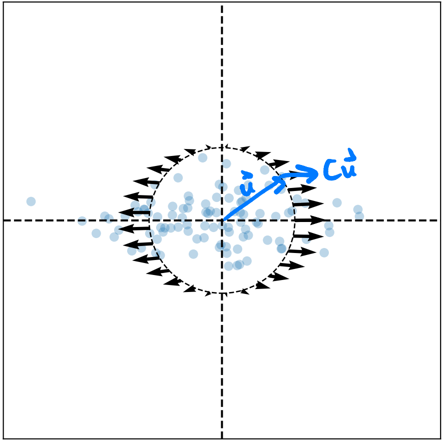
- Covariance matrices are symmetric.
- They have axes of symmetry (eigenvectors and eigenvalues).
- What are they?

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

Visualizing Covariance Matrices

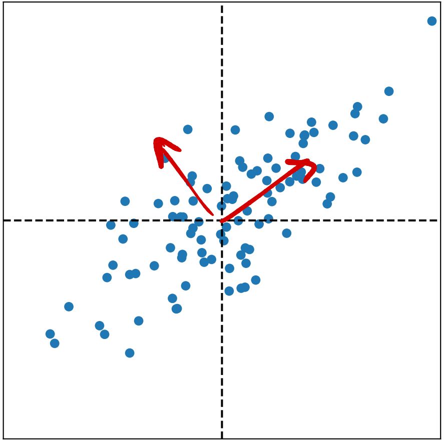


Eigenvectors:

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

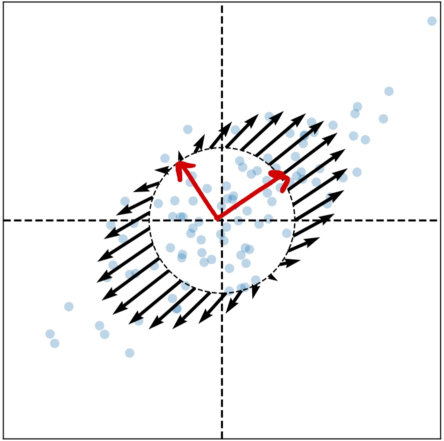
$\vec{u}^{(2)} \approx (0, 1)^T$

Visualizing Covariance Matrices



$$C \approx \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

Visualizing Covariance Matrices



Eigenvectors:

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

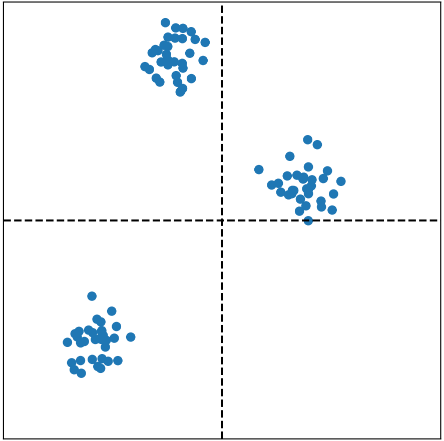
$\vec{u}^{(2)} \approx (-1, 1)^T$

Observations

- 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.

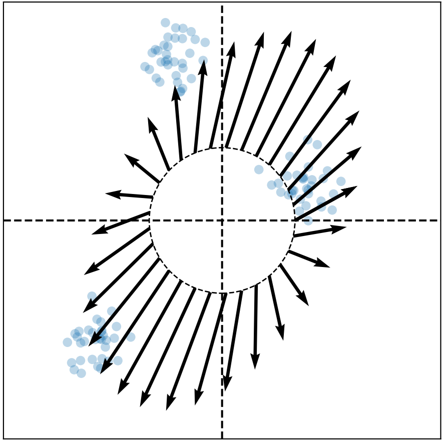
Caution

- 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.



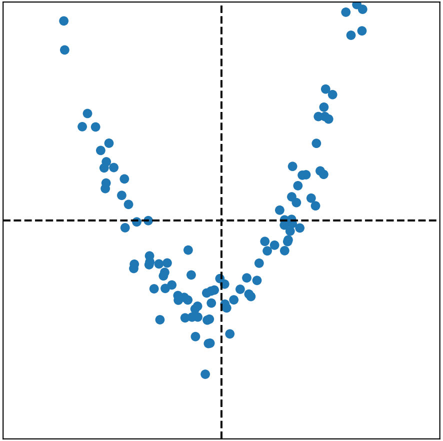
Caution

- 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.



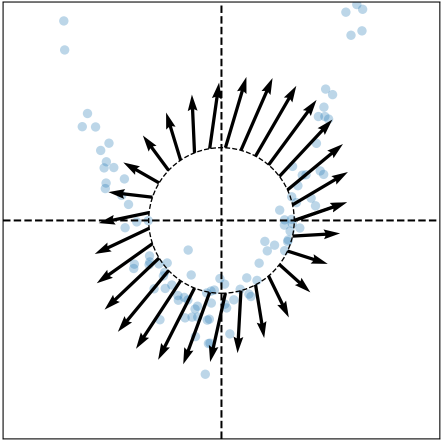
Caution

- 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"

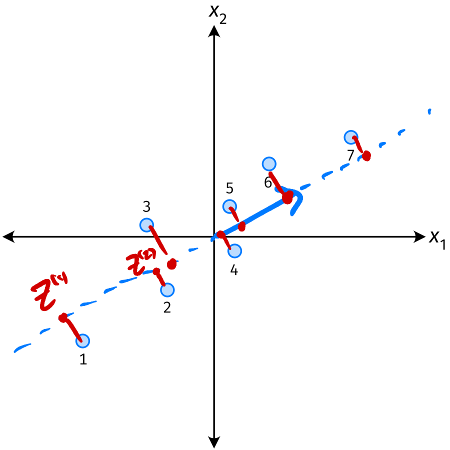
- Let's derive PCA more formally.

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 in the direction of $\vec{u}$** is defined to be the variance of the new features:

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

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

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

- Proven on this week’s homework.

Our Goal (Again)

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

Recall

- When C is symmetric, the unit vector which maximizes the quadratic form $\vec{u}^T C \vec{u}$ is the eigenvector of C with the largest eigenvalue.
- **Solution**: the direction of maximum variance is the top eigenvector of the covariance matrix.

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[2] eigenvector $\vec{u}$, of C.

3. For $i \in \{1, \ldots, n\}$, create new feature:

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

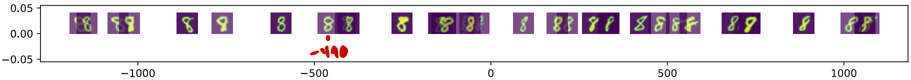
[2] All eigenvalues are positive. Why?

2 8

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 06 | Part 5

Dimensionality Reduction with d ≥ 2

So far: PCA

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

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 $\mathbb{R}^d$

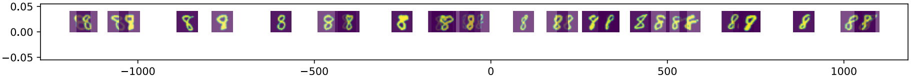
- Make a new feature $z \in \mathbb{R}$ for point $\vec{x} \in \mathbb{R}^d$ by computing $z = \vec{x} \cdot \vec{u}$

- 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?

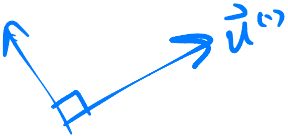
- Clearly, mapping from $\mathbb{R}^{784} \to \mathbb{R}^1$ loses a lot of information
- What about mapping from $\mathbb{R}^{784} \to \mathbb{R}^2$? $\mathbb{R}^k$?

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

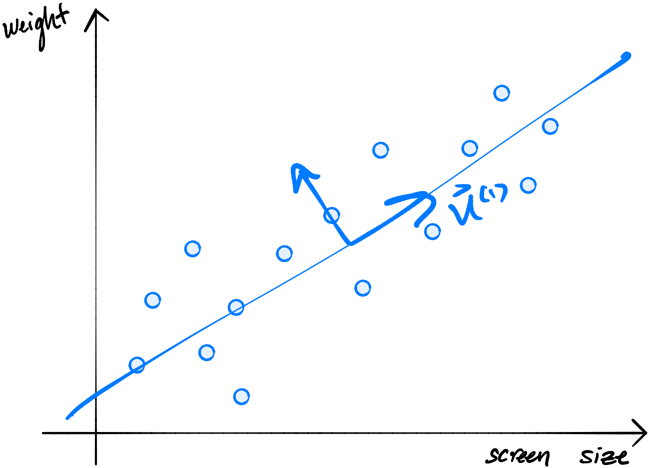
- Make same assumption for second feature:

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

- 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

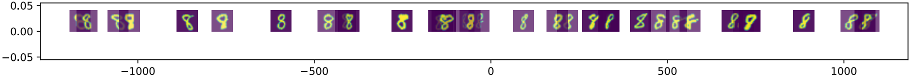
- Given a covariance matrix C.
- The principal component $\vec{u}^{(1)}$ is the top eigenvector of C.
  - Points in the direction of maximum variance.
- The *second* principal component $\vec{u}^{(2)}$ is the *second* eigenvector of C.
  - Out of all vectors orthogonal to the principal component, points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

PCA: Two Components

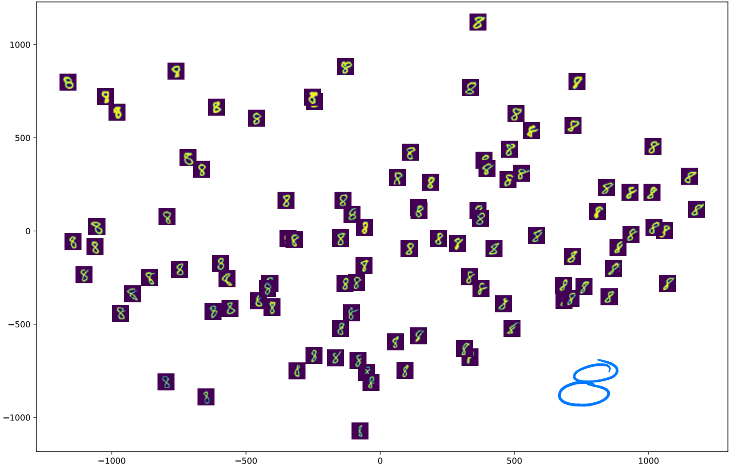
- Given data $\{\vec{x}^{(1)}, \dots, \vec{x}^{(n)}\} \in \mathbb{R}^d$.
- Compute covariance matrix C, top two eigenvectors $\vec{u}^{(1)}$ and $\vec{u}^{(2)}$.
- For any vector $\vec{x} \in \mathbb{R}$, its new representation in $\mathbb{R}^2$ is $\vec{z} = (z_1, z_2)^T$, where:

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

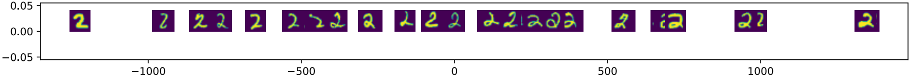
Example



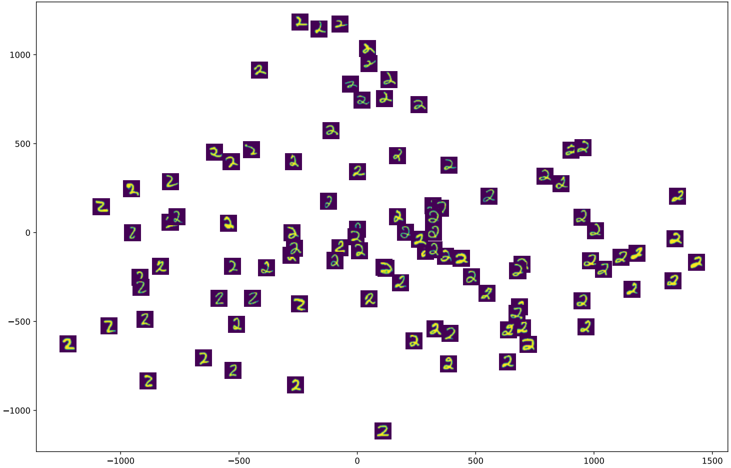
Example



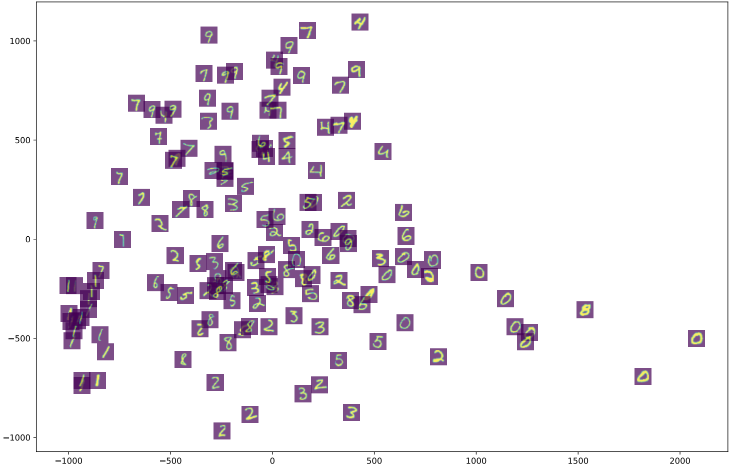
Example



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: $Z = XU$

DSC 140B

Representation Learning

Lecture 06 | Part 6

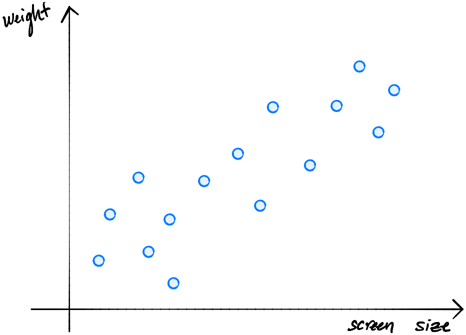
Reconstructions

Reconstructing Points

- PCA helps us reduce dimensionality from $\mathbb{R}^d \to R^k$
- Suppose we have the “new” representation in $\mathbb{R}^k$.
- Can we “go back” to $\mathbb{R}^d$?
- And why would we want to?

Back to $\mathbb{R}^d$

- Suppose new representation of $\vec{x}$ is z.
- $z = \vec{x} \cdot \vec{u}^{(1)}$
- Idea: $\vec{x} \approx z\vec{u}^{(1)}$



Reconstructions

- Given a “new” representation of $\vec{x}$, $\vec{z} = (z_1, \ldots, z_k) \in \mathbb{R}^k$
- And top k eigenvectors, $\vec{u}^{(1)}, \ldots, \vec{u}^{(k)}$
- The **reconstruction** of $\vec{x}$ is

$$z_1 \vec{u}^{(1)} + z_2 \vec{u}^{(2)} + \ldots + z_k \vec{u}^{(k)} = U\vec{z}$$

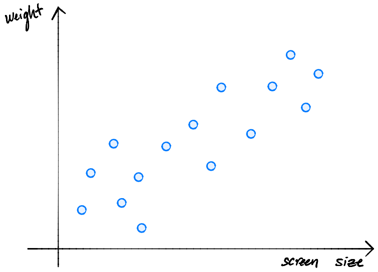
Reconstruction Error

- The reconstruction *approximates* the original point, $\vec{x}$.
- The **reconstruction error** for a single point, $\vec{x}$:

$$\|\vec{x} - U\vec{z}\|^2$$

- Total reconstruction error:

$$\sum_{i=1}^{n} \|\vec{x}^{(i)} - U\vec{z}^{(i)}\|^2$$



DSC 140B

Representation Learning

Lecture 06 | Part 7

Interpreting PCA

Three Interpretations

- What is PCA doing?

- Three interpretations:
    1. Mazimizing variance
    2. Finding the best reconstruction
    3. Decorrelation

Recall: Matrix Formulation

- Given data matrix X.
- Compute new data matrix $Z = XU$.
- PCA: choose U to be matrix of eigenvectors of C.
- For now: suppose U can be anything – but columns should be orthonormal
  - Orthonormal = “not redundant”

View #1: Maximizing Variance

- This was the view we used to derive PCA
- Define the **total variance** to be the sum of the variances of each column of Z.
- Claim: Choosing U to be top eigenvectors of C maximizes the total variance among all choices of orthonormal U.

Main Idea

PCA maximizes the total variance of the new data. I.e., chooses the most “interesting” new features which are not redundant.

View #2: Minimizing Reconstruction Error

- Recall: total reconstruction error

$$\sum_{i=1}^{n} \| \vec{x}^{(i)} - U\vec{z}^{(i)} \|^2$$

- Goal: minimize total reconstruction error.

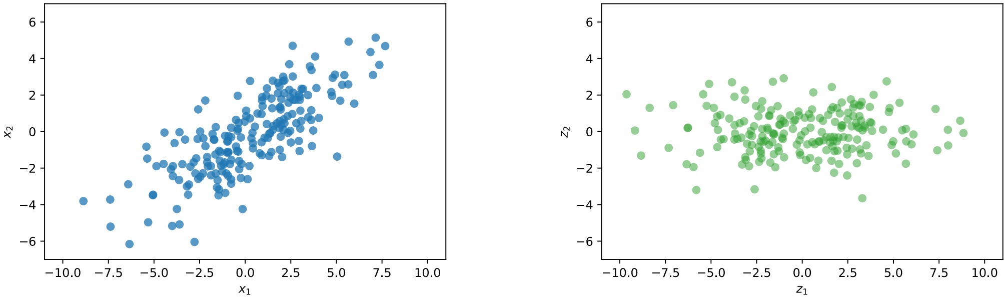
- Claim: Choosing U to be top eigenvectors of C minimizes reconstruction error among all choices of orthonormal U

Main Idea

PCA minimizes the reconstruction error. It is the "best" projection of points onto a linear subspace of dimensionality k. When $k = d$, the reconstruction error is zero.

View #3: Decorrelation

- PCA has the effect of "decorrelating" the features.



Main Idea

PCA learns a new representation by rotating the data into a basis where the features are uncorrelated (not redundant). That is: the natural basis vectors are the principal directions (eigenvectors of the covariance matrix). PCA changes the basis to this natural basis.

DSC 140B
Representation Learning

Lecture 06 | Part 8

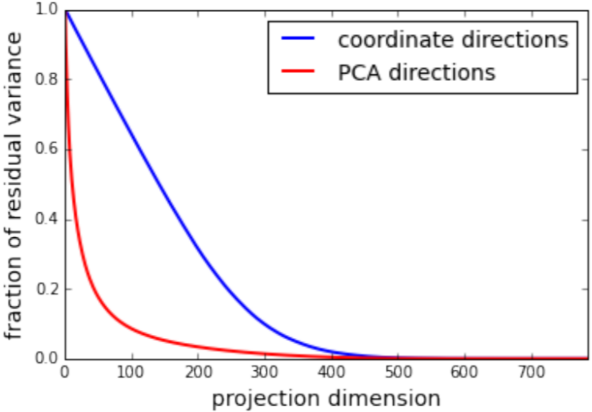
PCA in Practice

PCA in Practice

- PCA is often used in **preprocessing** before classifier is trained, etc.
- Must choose number of dimensions, k.
- One way: cross-validation.
- Another way: the elbow method.

Total Variance

- The **total variance** is the sum of the eigenvalues of the covariance matrix.
- Or, alternatively, sum of variances in each orthogonal basis direction.


coordinate directions
PCA directions
fraction of residual variance
projection dimension
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700

Caution

- PCA's assumption: variance is interesting
- PCA is totally unsupervised
- The direction most meaningful for classification may not have large variance!