DATASCI 347 Machine Learning

Lecture 15: PCA

Ruoxuan Xiong

Suggested reading: ISL Chapter 6

EMORY

Lecture plan

- Principal component analysis

Principal component analysis (PCA)

X_1	X_2	X_3	⋯	⋯	⋯	X_p
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*
*	*	*	…	…	…	*

Use M features to summarize most of the information in the original p features

Reduce dimensionality (Principal component analysis)

Small table ($M = 2$)

Z_1	Z_2
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*

Reduce dimensionality (stock return data)

	AAPL	MSFT	AMZN	GOOG	NVDA	…
20220103	*	*	*	*	*	…
20220104	*	*	*	*	*	…
20220105	*	*	*	*	*	…
20220106	*	*	*	*	*	…
20220107	*	*	*	*	*	…
20220110	*	*	*	*	*	…
20220111	*	*	*	*	*	…
20220112	*	*	*	*	*	…
…	*	*	*	*	*	…
…	*	*	*	*	*	…
…	*	*	*	*	*	…
…	*	*	*	*	*	…

Reduce dimensionality →

$M = 1$

SP500
*
*
*
*
*
*
*
*
*
*
*
*

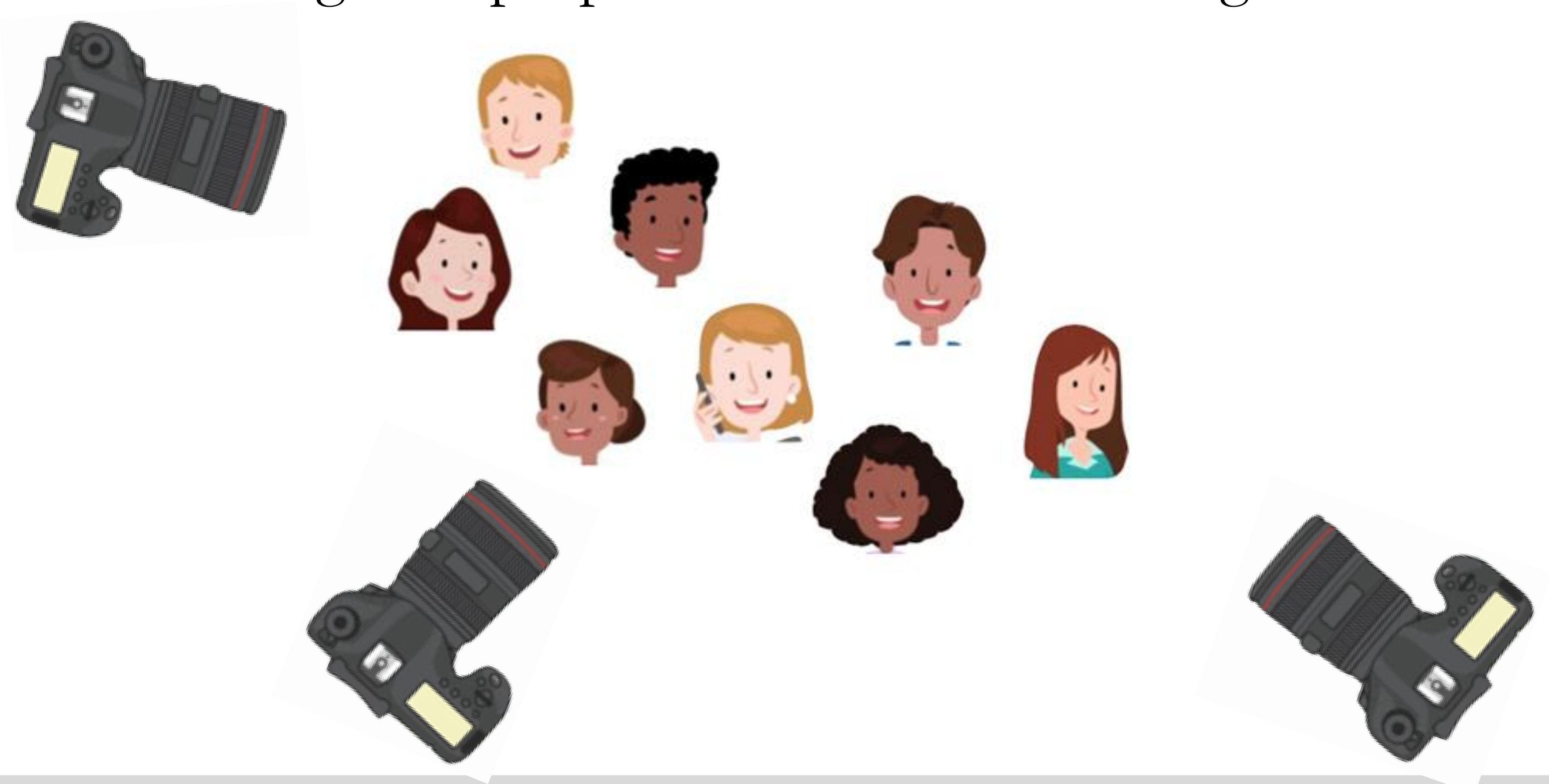
EMORY

Principal component analysis (PCA)

- Find M features, $Z_1, Z_2, \cdots, Z_M$, that can "best represent" the original p features $X_1, X_2, \cdots, X_p$
  - $M \ll p$
  - Reduce the dimensionality of $X_1, X_2, \cdots, X_p$
  - **Unsupervised learning** method
- Question: How should we select the M features?

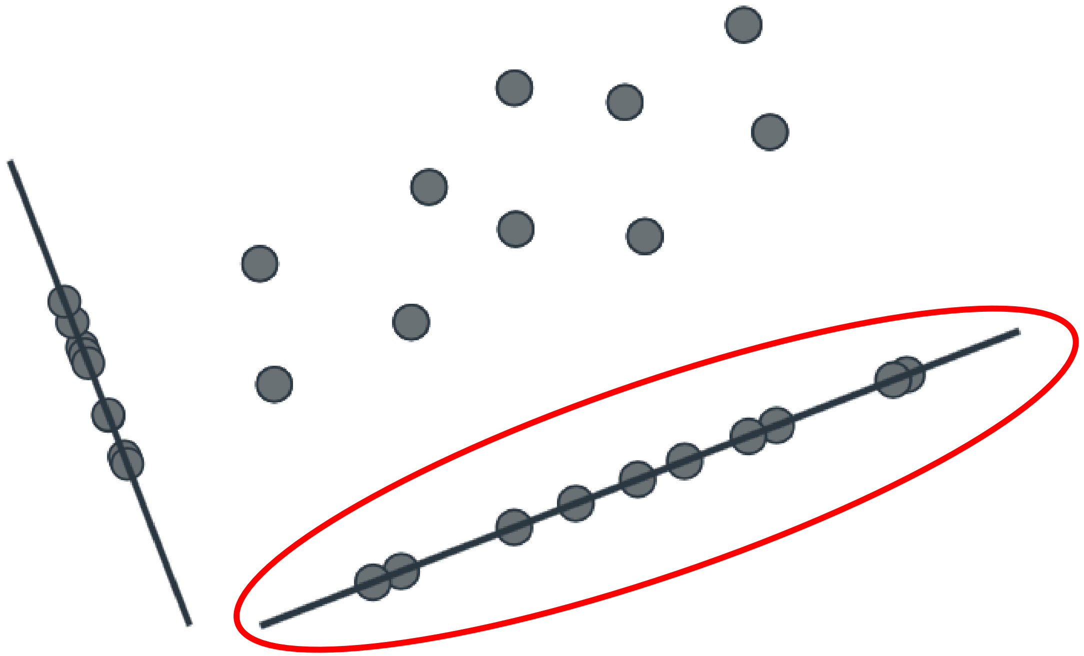
Intuition

- When your "big data" is too big
- Suppose we are taking multiple pictures from different angles



Intuition

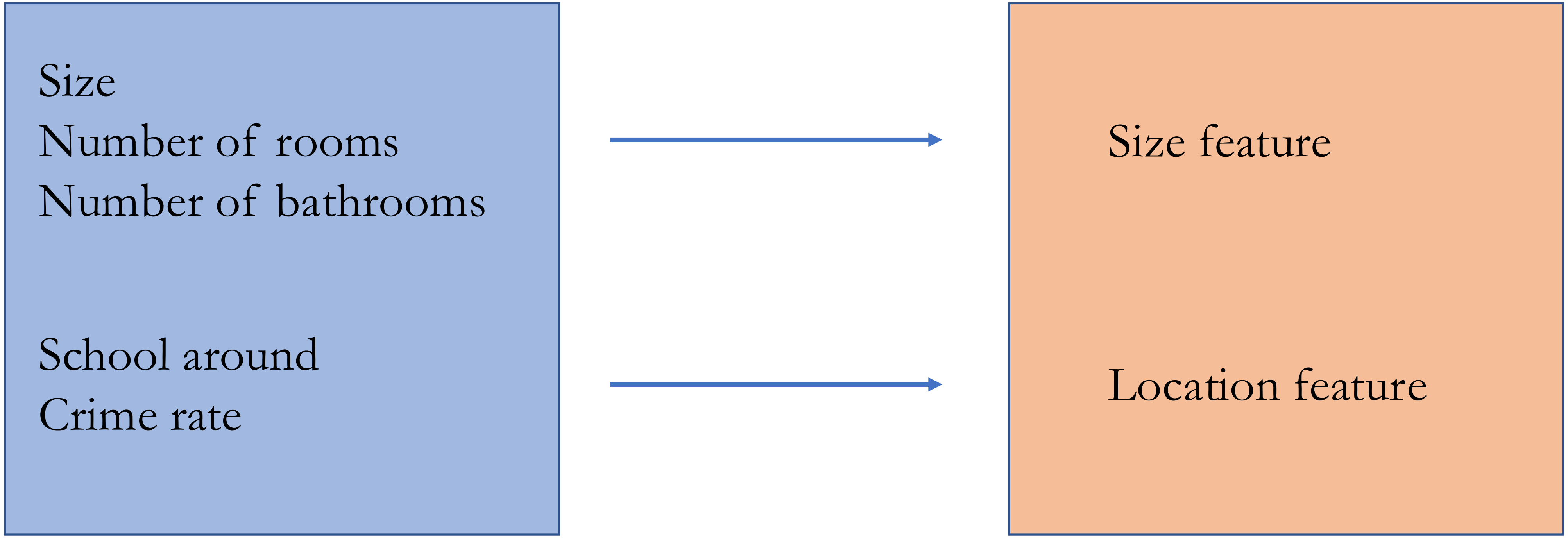
- Suppose we are taking multiple pictures from different angles
  - We have obtained data points from different angles
  - Which is the "important" direction?



  - Principal component analysis finds this "important" direction

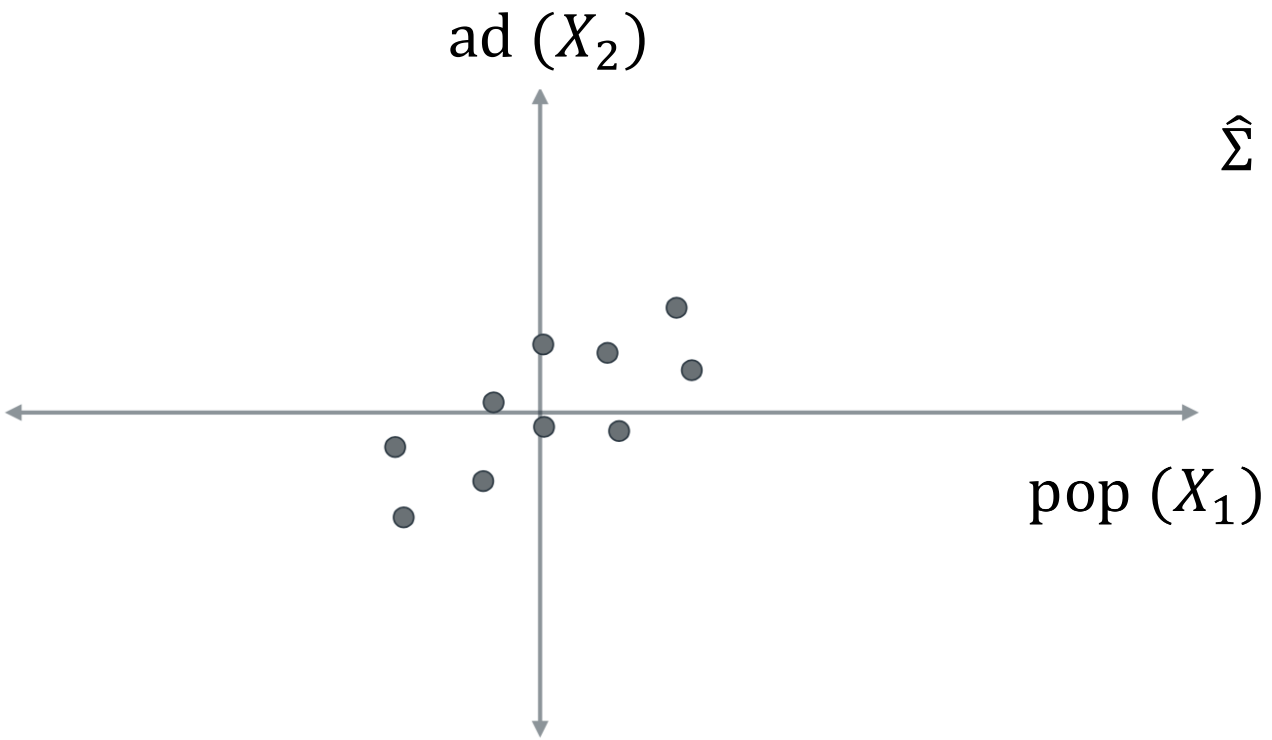
Example

- Five features ($p = 5$) in the Boston housing data
- Reduce them to $M = 2$ features



How to perform PCA I

1. Estimate the **covariance matrix** $\hat{\Sigma}$ of $X_1, X_2, \cdots, X_p$
   - $\hat{\Sigma}$ is a $p \times p$ matrix, the (i, j)-th entry being the covariance of X_i, X_j
   - **Example**: population size (pop) and ad spending (ad) for 100 cities



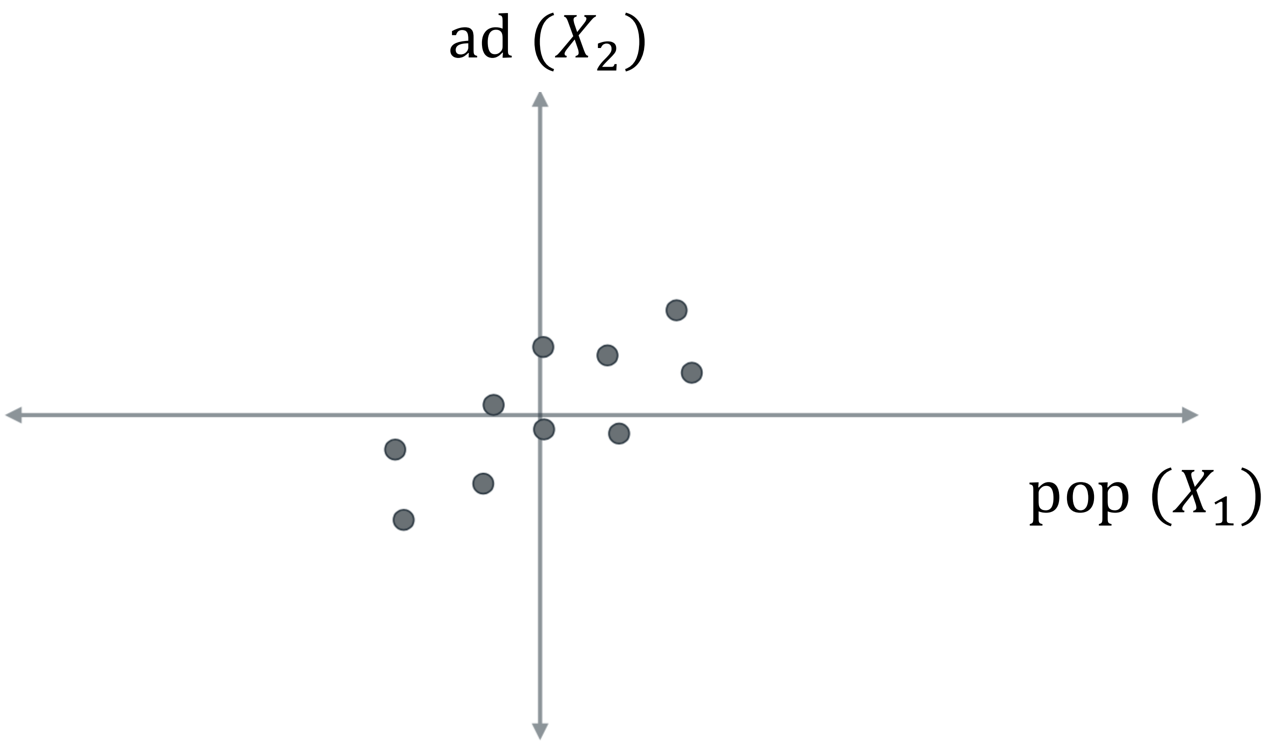
$$\hat{\Sigma} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 3.816 & 1.826 \\ 1.826 & 2.184 \end{bmatrix}$$

How to perform PCA II

2. Calculate the **eigenvalues** and **eigenvectors** of the covariance

- **Covariance matrix:** $\hat{\Sigma} = \begin{bmatrix} 3.816 & 1.826 \\ 1.826 & 2.184 \end{bmatrix}$



Unit norm eigenvectors

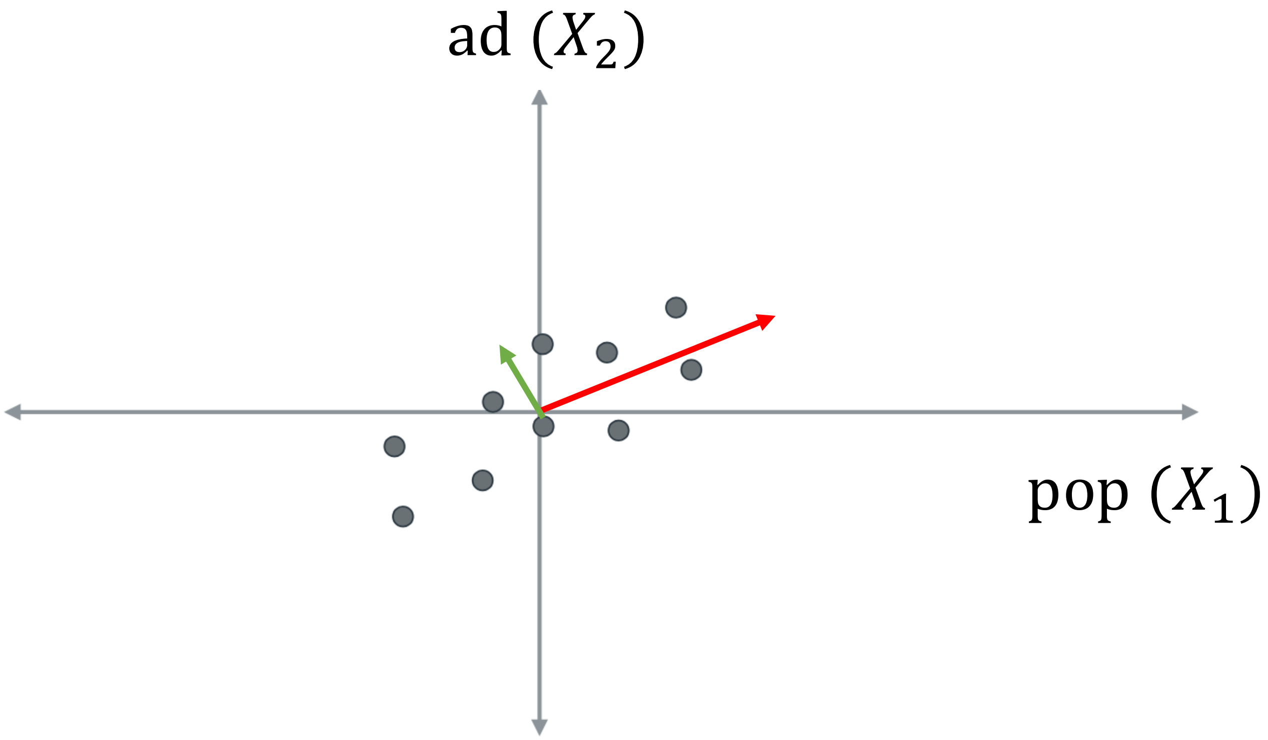
$\begin{pmatrix} 0.839 \\ 0.544 \end{pmatrix}$ $\begin{pmatrix} 0.544 \\ -0.839 \end{pmatrix}$

Eigenvalues

5 1

How to perform PCA III

3. Select the **first principal component**



Unit norm eigenvectors (direction)

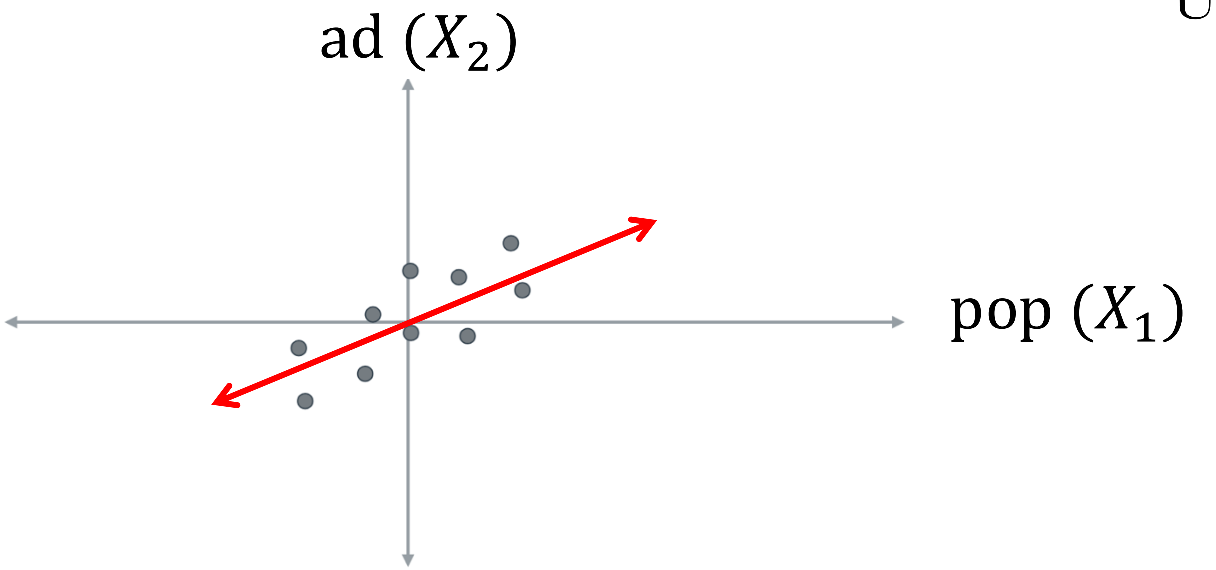
$$\begin{pmatrix} \mathbf{0.839} \\ \mathbf{0.544} \end{pmatrix} \quad \begin{pmatrix} 0.544 \\ -0.839 \end{pmatrix}$$

Eigenvalues (magnitude)

5 ✓ 1

First principal component

- **Geometric interpretation:** $(\phi_{11}, \phi_{21}) = (0.839, 0.544)$ is the solution to
  - Maximize $\text{Var}\left(\phi_{11} \times (\text{pop}_i - \overline{\text{pop}}) + \phi_{21} \times (\text{ad}_i - \overline{\text{ad}})\right)$
  - Subject to the constraint $\phi_{11}^2 + \phi_{21}^2 = 1$



Unit norm eigenvectors (direction)

$$\begin{pmatrix} 0.839 \\ 0.544 \end{pmatrix}$$

Eigenvalues (magnitude)

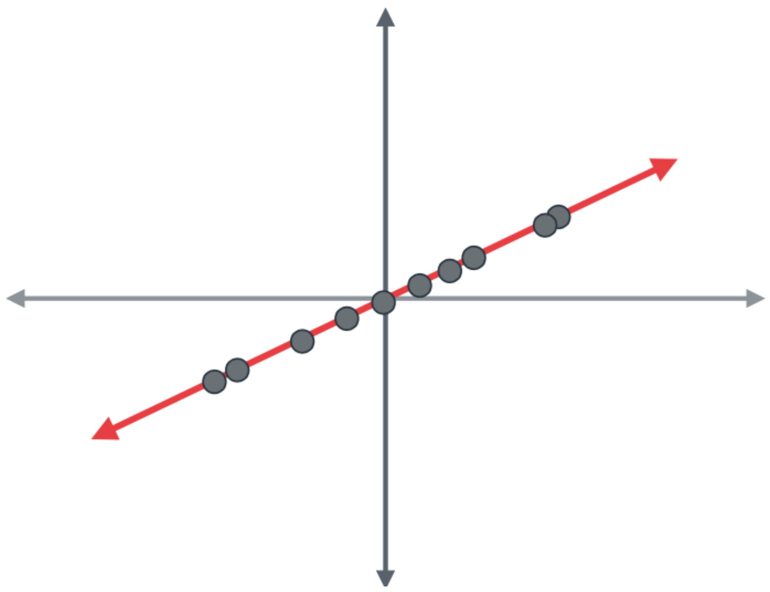
5

- $\text{Var}\left(0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}})\right) = 5$

Projection to first principal component

- First principal component, which is a line, corresponds to the following equation:
  - $z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times \left(\text{ad}_i - \overline{\text{ad}}\right)$

ad (X_2)



pop (X_1)

Unit norm eigenvectors (direction)

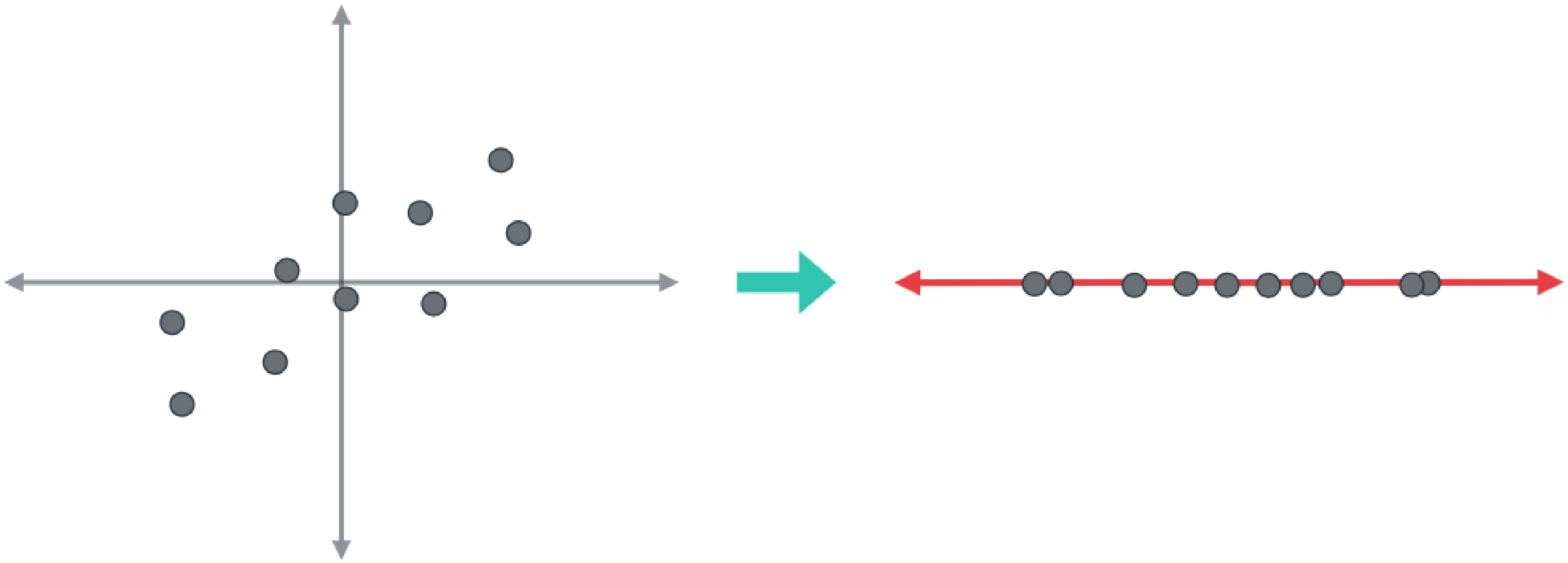
$$\begin{pmatrix} 0.839 \\ 0.544 \end{pmatrix}$$

Eigenvalues (magnitude)

5

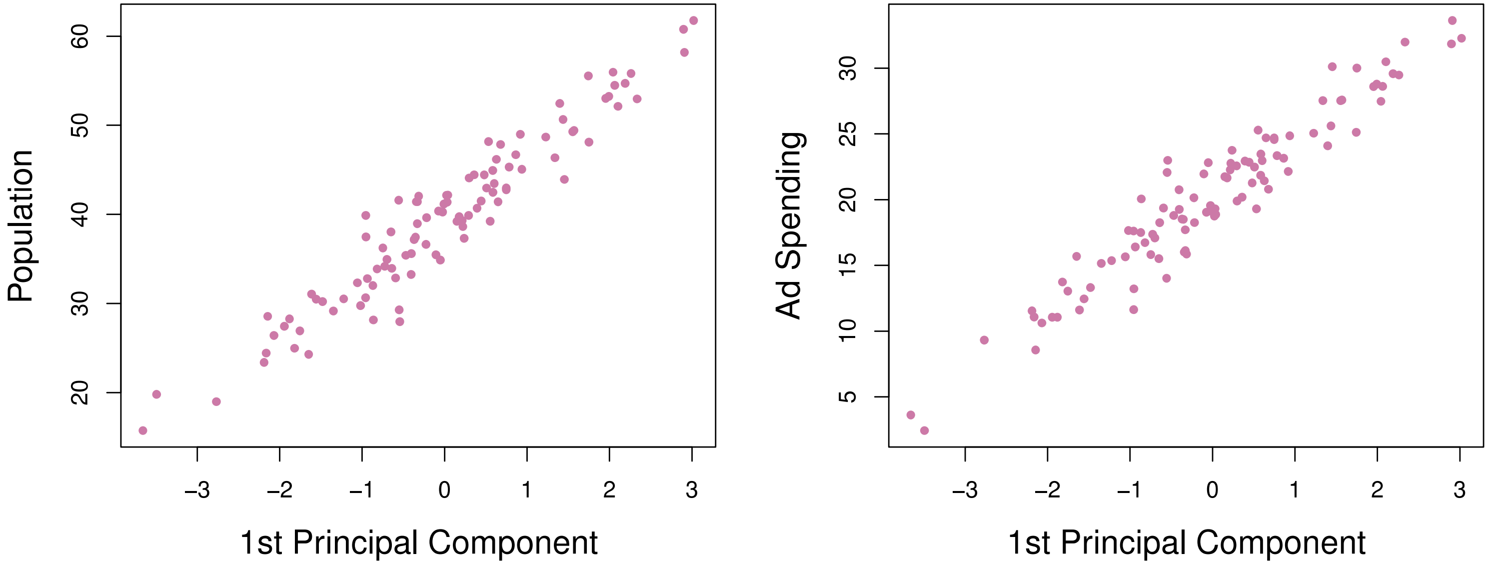
Projection reduces dimension

- Projecting to the first principal component leads to
  - $z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}})$
- This projection is the most accurate projection of the data to one dimension
  - The projected observations are as close as possible to the original observations



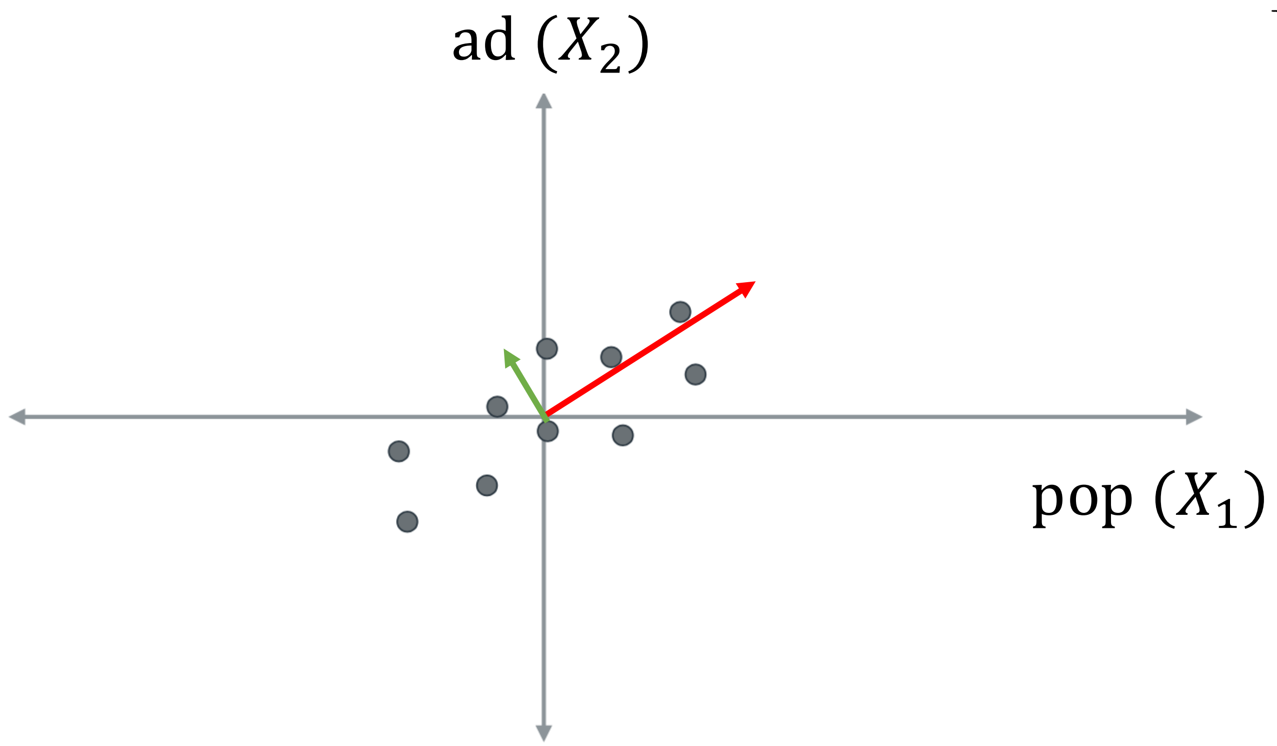
Illustration

- Illustrating first principal component scores
  - $z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}})$
- The plots show a **strong** relationship between z_{i1} and both $\mathbf{pop}_i$ and $\mathbf{ad}_i$ features



How to perform PCA IV

4. Select the **second principal component** (if necessary)



Unit norm eigenvectors (direction)

$$\begin{pmatrix} 0.839 \\ 0.544 \end{pmatrix} \quad \begin{pmatrix} 0.544 \\ -0.839 \end{pmatrix}$$

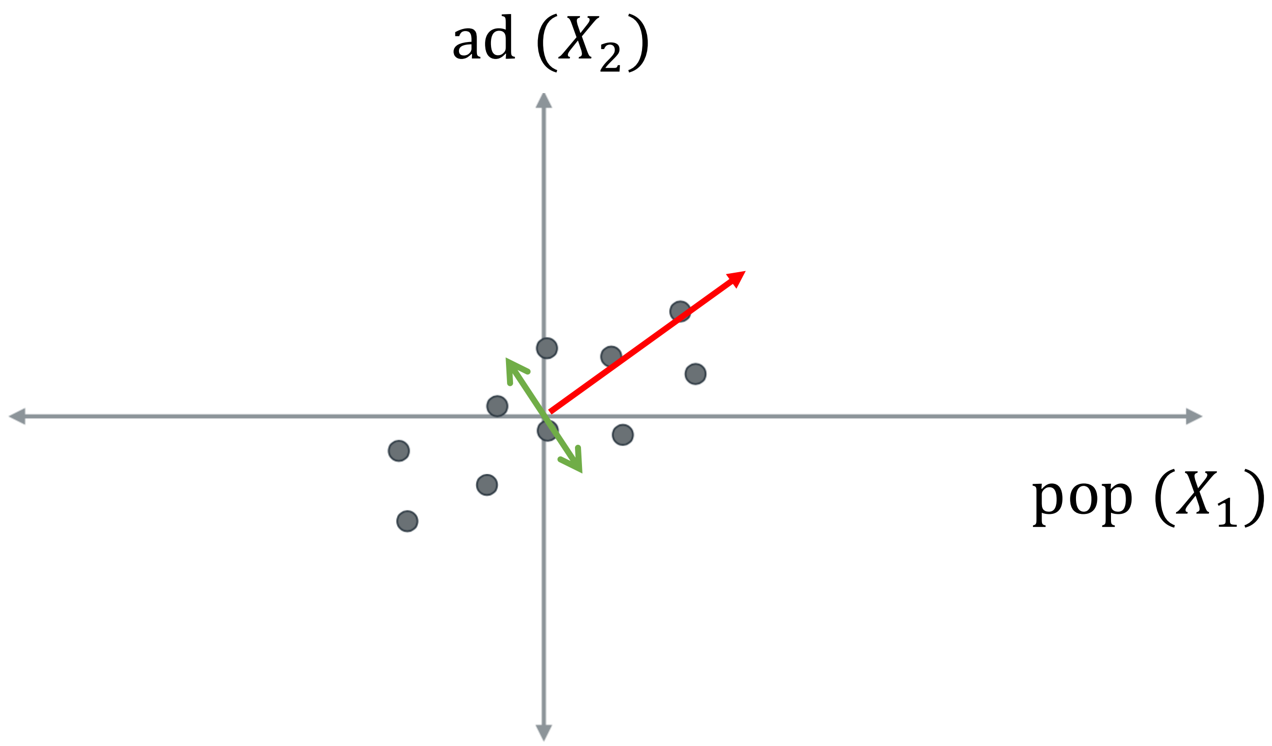
Eigenvalues (magnitude)

5 ✓ 1 ✓

EMORY

Second principal component

- The second principal component Z_2 is a linear combination of variables that is **orthogonal to** first principal component Z_1 and has **largest variance** subject to **being orthogonal**



Unit norm eigenvectors (direction)

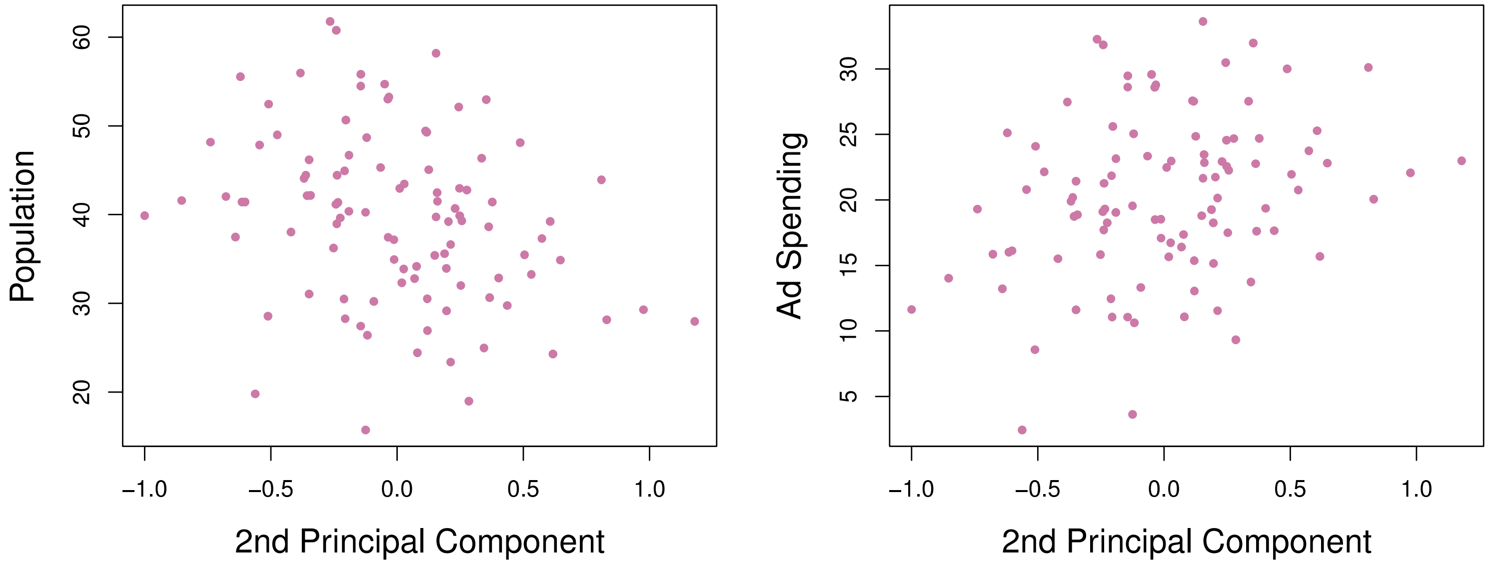
$$\begin{pmatrix} 0.839 \\ 0.544 \end{pmatrix} \quad \begin{pmatrix} \mathbf{0.544} \\ \mathbf{-0.839} \end{pmatrix}$$

Eigenvalues (magnitude)

5 1

Projection to second principal component

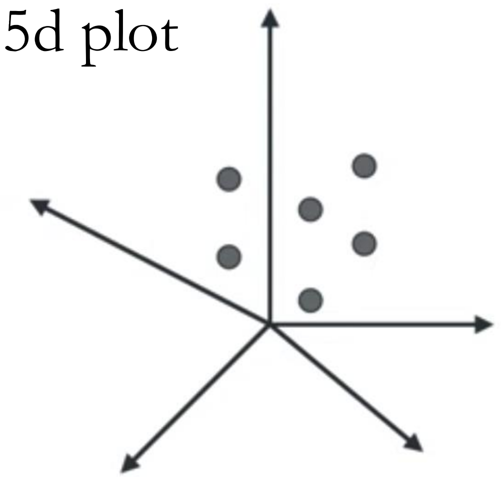
- Illustrating second principal component scores
  - $z_{i2} = 0.544 \times (\text{pop}_i - \overline{\text{pop}}) - 0.839 \times (\text{ad}_i - \overline{\text{ad}})$
- The plots show a **weak** relationship between z_{i2} and the pop_i and ad_i features

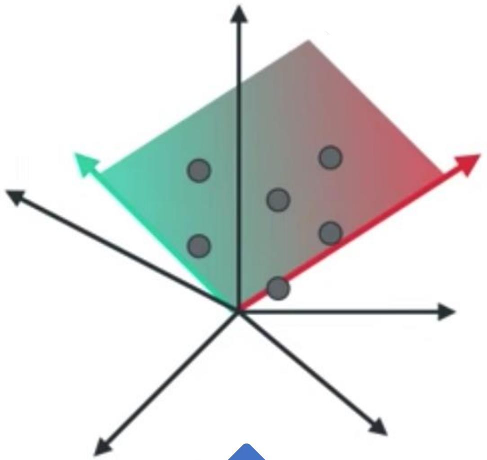


Summarizing PCA

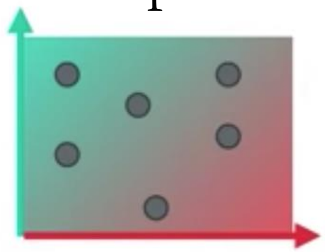
X_1	X_2	X_3	X_4	X_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

5d plot





2d plot



Covariance matrix

*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Eigenvector	Eigenvalue
V_1	λ_1
V_2	λ_2
V_3	λ_3
V_4	λ_4
V_5	λ_5

Big

Small

Small table

Z_1	Z_2
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*