QTM 347 Machine Learning

Lecture 18: PCA

Ruoxuan Xiong

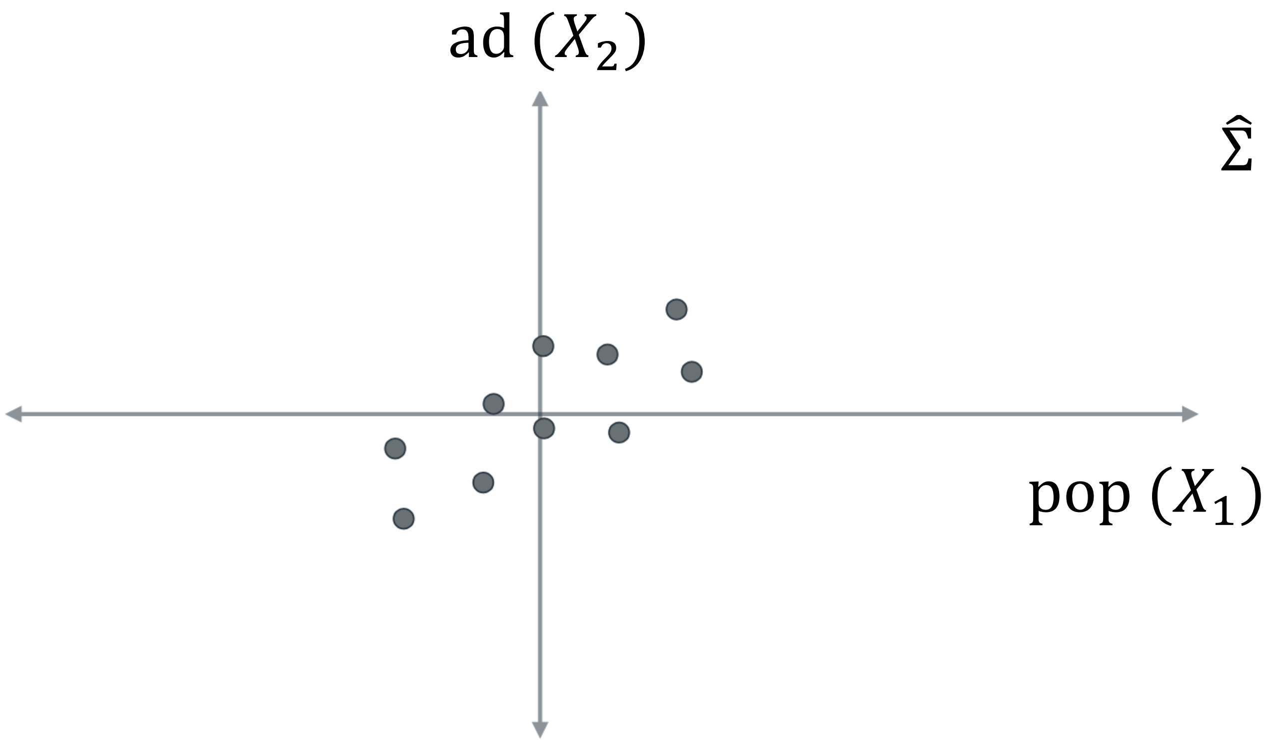
Suggested reading: ISL Chapter 6 and 12

Lecture plan

- PCA

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.



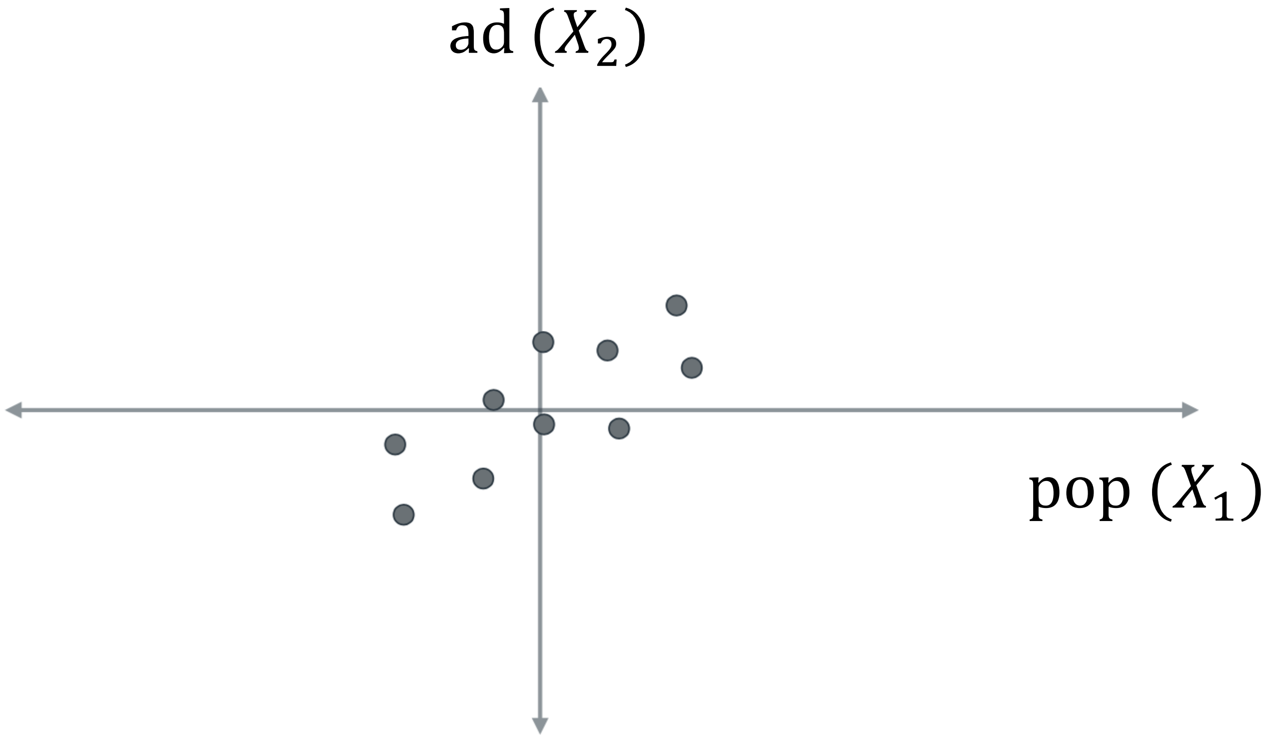
$$\hat{\Sigma} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{bmatrix}$$

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

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

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

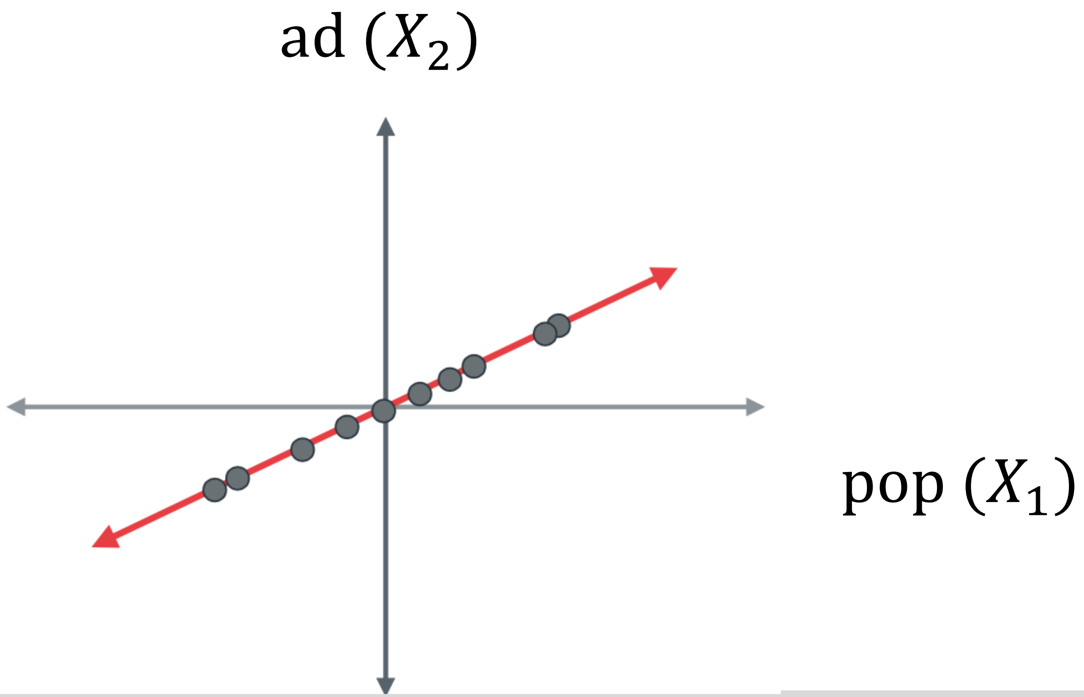
Eigenvalues

$\lambda_1 = 5$ $\lambda_2 = 1$

Projection to first principal component

3. Select the **first principal component**

- First principal component, which is corresponds to the following equation:
  - $z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}})$ and $Var(z_{i1}) = \lambda_1$



Unit norm eigenvectors (direction)

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

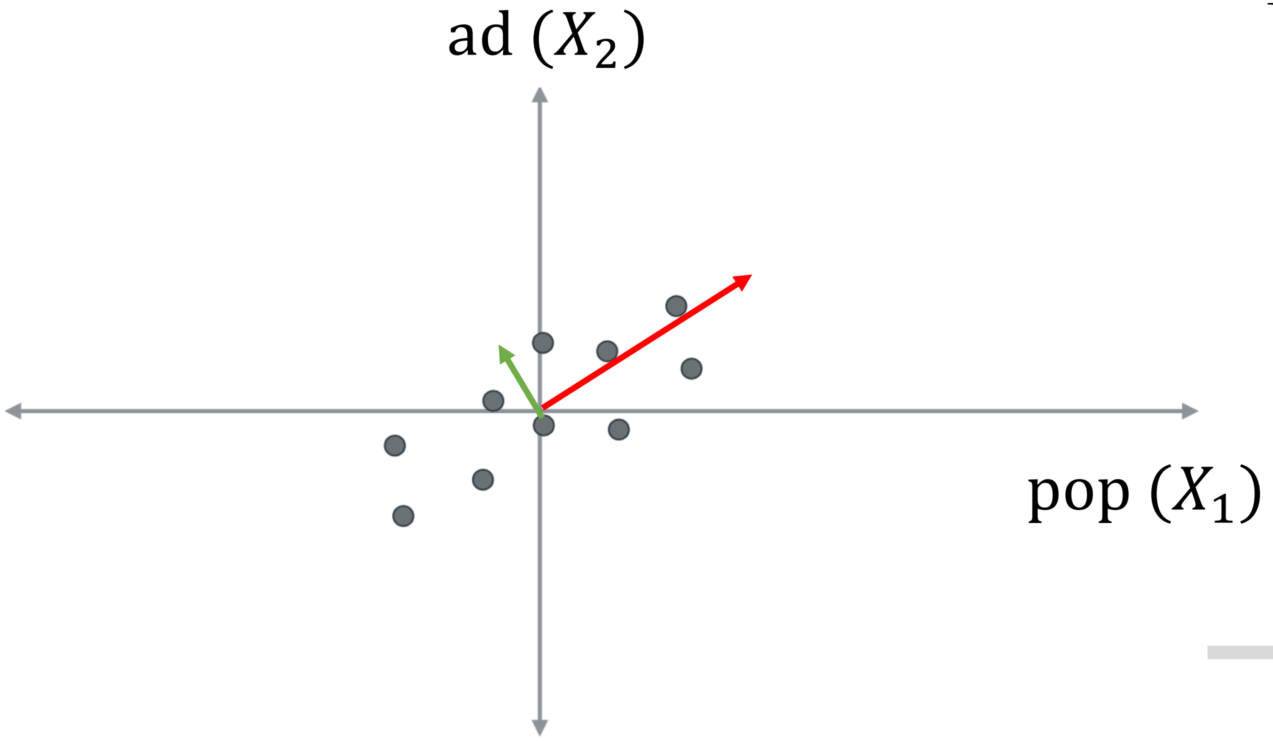
Eigenvalues (magnitude)

$$\lambda_1 = 5$$

How to perform PCA IV

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

- The second principal component Z_2 has **largest variance** subject to **being orthogonal** to first principal component Z_1
  - $z_{i2} = 0.544 \times (\text{pop}_i - \overline{\text{pop}}) - 0.839 \times \left(\text{ad}_i - \overline{\text{ad}}\right)$ and $Var(z_{i2}) = \lambda_2$



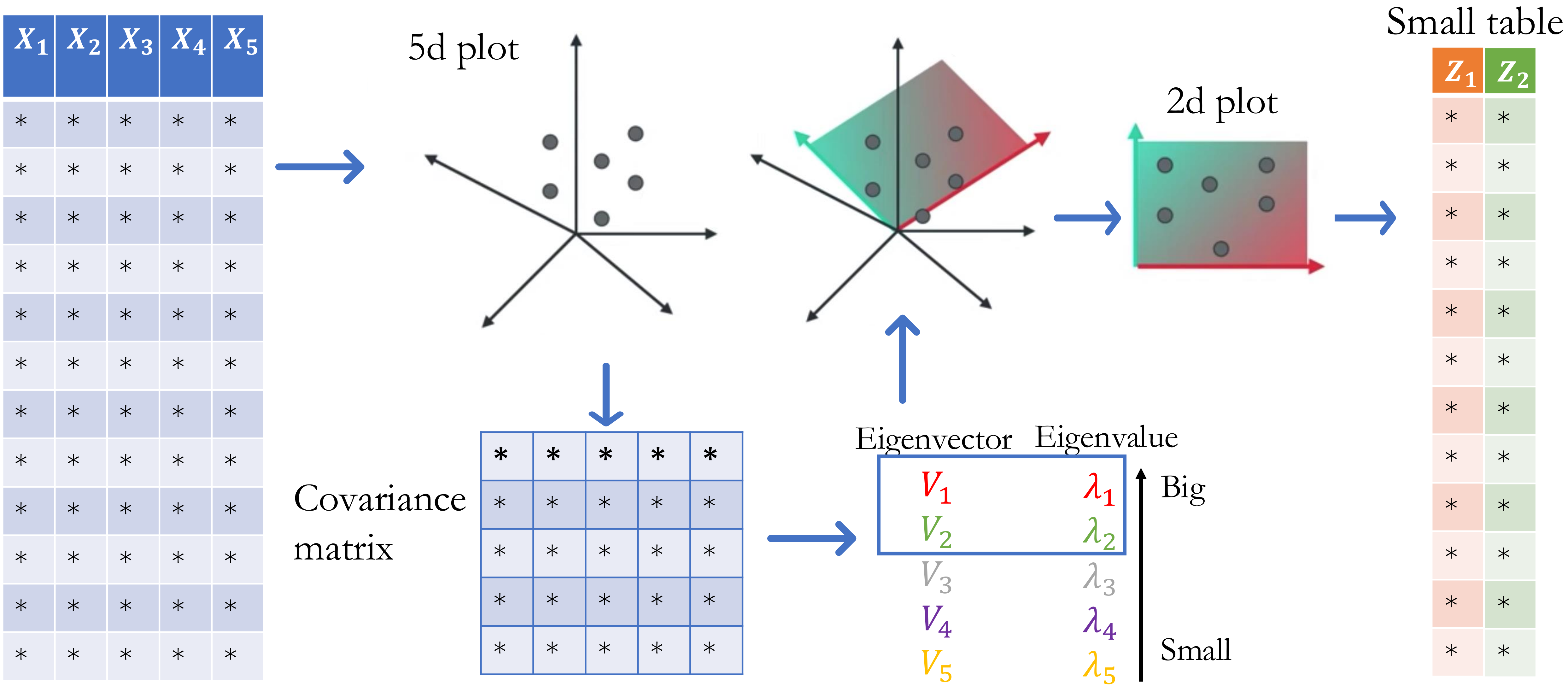
Unit norm eigenvectors (direction)

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

Eigenvalues (magnitude)

$\lambda_1 = 5$ $\lambda_2 = 1$

Summarizing PCA



More on PCA

- **Mean**: Variables should be centered to have mean zero
  - First principal component (PC) reflects the direction of max variance, instead of the mean of the data
- **Variance**: Choose case by case whether to scale variables to have unit variance
  - Results typically *depend on* whether variables have been individually scaled
    - Small-scale variables will have small variance
  - *Whether to scale* depends on whether variables are *measured on the same unit*

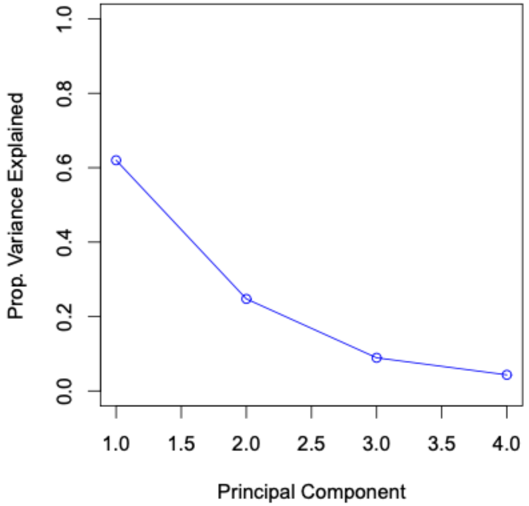
  - Example 1: Variables are expression levels of genes (no need to scale the genes)

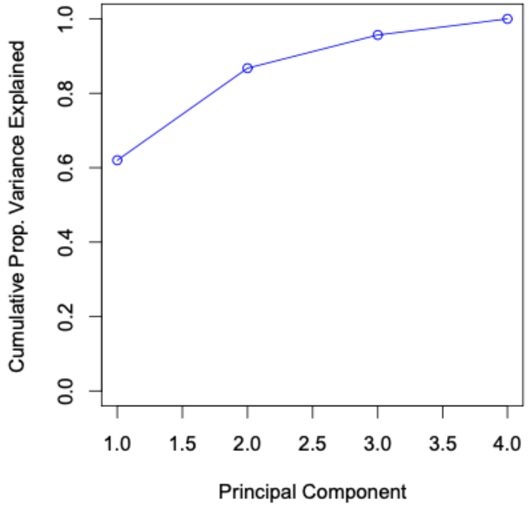
  - Example 2: Variables include ad spending and population size (scale the variables)

Choosing the number of PCs

- **Choosing the number of PCs**:
  - How much information is lost by projecting observations on the first M PCs?
  - Equivalently, how much variance of the data is not contained in the first M PCs?
  - Choose the smallest number that explains a sizable amount of the variation
- Eigenvalues of feature covariance matrix: $\lambda_1, \lambda_2, \cdots, \lambda_p$
- **Scree plot** shows the variance explained by each PC (an ad hoc method):

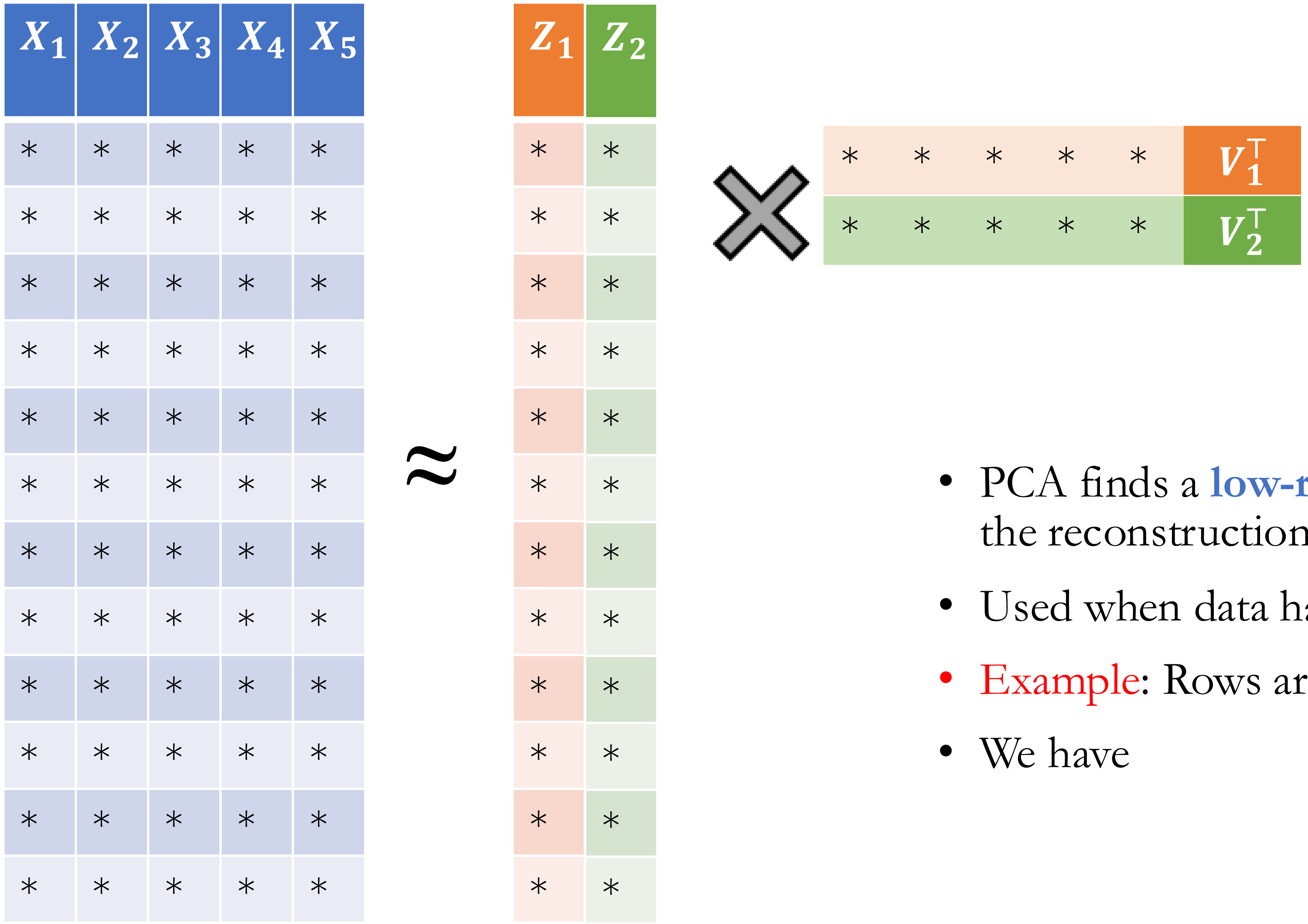
$$\frac{\lambda_1}{\lambda_1+\lambda_2+\cdots+\lambda_p}, \frac{\lambda_2}{\lambda_1+\lambda_2+\cdots+\lambda_p}, \cdots, \frac{\lambda_p}{\lambda_1+\lambda_2+\cdots+\lambda_p}$$





The first PC explains 62%
The next PC explains 24.7%

PCA for low-rank matrix factorization



- PCA finds a **low-rank matrix factorization** that minimizes the reconstruction error
- Used when data has **inherent low-dimensional structure**
- Example: Rows are users and columns are movies
- We have

$$\boldsymbol{X} \approx [\boldsymbol{Z}_1 \quad \boldsymbol{Z}_2] \begin{bmatrix} \boldsymbol{V}_1^\top \\ \boldsymbol{V}_2^\top \end{bmatrix}$$

Rationale behind low-rank approximation

- For any jth PC, we have $\boldsymbol{XV_j} = \boldsymbol{Z_j}$, or equivalently, for each unit i, $Z_{ij} = V_{1j}X_{i1} + V_{2j}X_{i2} + \cdots + V_{pj}X_{ip}$, where V_{kj} is the kth entry in $\boldsymbol{V_j}$

- Right multiply $\boldsymbol{XV_j} = \boldsymbol{Z_j}$ by $\boldsymbol{V_j}$, and sum over j, we have $\sum_{j=1}^{p} \boldsymbol{XV_jV_j^\top} = \sum_{j=1}^{p} \boldsymbol{Z_jV_j^\top}$

- As $\boldsymbol{X}$ does not depend on j, we can take $\boldsymbol{X}$ out from the sum and $\sum_{j=1}^{p} \boldsymbol{XV_jV_j^\top} = \boldsymbol{X}\sum_{j=1}^{p} \boldsymbol{V_jV_j^\top} = \boldsymbol{X}$
  - Here we use an important property of eigenvectors: $\sum_{j=1}^{p} \boldsymbol{V_jV_j^\top} = \boldsymbol{I_p}$ (identity matrix)

$$\boldsymbol{X} = \sum_{j=1}^{p} \boldsymbol{Z_jV_j^\top} = [\boldsymbol{Z_1} \quad \cdots \quad \boldsymbol{Z_p}]\begin{bmatrix}\boldsymbol{V_1^\top} \\ \vdots \\ \boldsymbol{V_p^\top}\end{bmatrix} \approx [\boldsymbol{Z_1} \quad \boldsymbol{Z_2}]\begin{bmatrix}\boldsymbol{V_1^\top} \\ \boldsymbol{V_2^\top}\end{bmatrix}$$

  - Third to last eigenvectors are truncated when $\boldsymbol{Var(Z_j)}$ is small for large $j = 3, \cdots, p$

EMORY

Missing values and matrix completion

- In data streaming services (e.g., Netflix, Amazon), most of the rating matrix is missing --- users only rated a tiny fraction of all movies/items
- We use the approximation

$$\boldsymbol{X} \approx [\boldsymbol{Z}_1 \quad \boldsymbol{Z}_2] \begin{bmatrix} \boldsymbol{V}_1^\top \\ \boldsymbol{V}_2^\top \end{bmatrix}$$

  - Most entries in $\boldsymbol{X}$ are *missing*
  - $[\boldsymbol{Z}_1 \quad \boldsymbol{Z}_2]$: *latent user features (e.g., cliques)*
  - $\begin{bmatrix} \boldsymbol{V}_1^\top \\ \boldsymbol{V}_2^\top \end{bmatrix}$: *latent movie features (e.g., genres)*
  - Estimate $\boldsymbol{Z}$ and $\boldsymbol{V}$ using observed entries in $\boldsymbol{X}$
  - An *iterative* algorithm:
    1. Impute missing entries by $\bar{X}$ (mean)
    2. Apply PCA or similar methods to estimate $\boldsymbol{Z}$ and $\boldsymbol{V}$
    3. Use estimated $\boldsymbol{Z}$ and $\boldsymbol{V}$ to impute missing entries in $\boldsymbol{X}$
    4. Repeat Steps 2 and 3 until convergence

	Jerry Maguire	Oceans	Road to Perdition	A Fortunate Man	Catch Me If You Can	Driving Miss Daisy	The Two Popes	The Laundromat	Code 8	The Social Network	···
Customer 1	·	·	·	·	4	·	·	·	·	·	···
Customer 2	·	·	3	·	·	·	3	·	·	3	···
Customer 3	·	2	·	4	·	·	·	·	2	·	···
Customer 4	3	·	·	·	·	·	·	·	·	·	···
Customer 5	5	1	·	·	4	·	·	·	·	·	···
Customer 6	·	·	·	·	·	2	4	·	·	·	···
Customer 7	·	·	5	·	·	·	·	3	·	·	···
Customer 8	·	·	·	·	·	·	·	·	·	·	···
Customer 9	3	·	·	·	5	·	·	1	·	·	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

EMORY