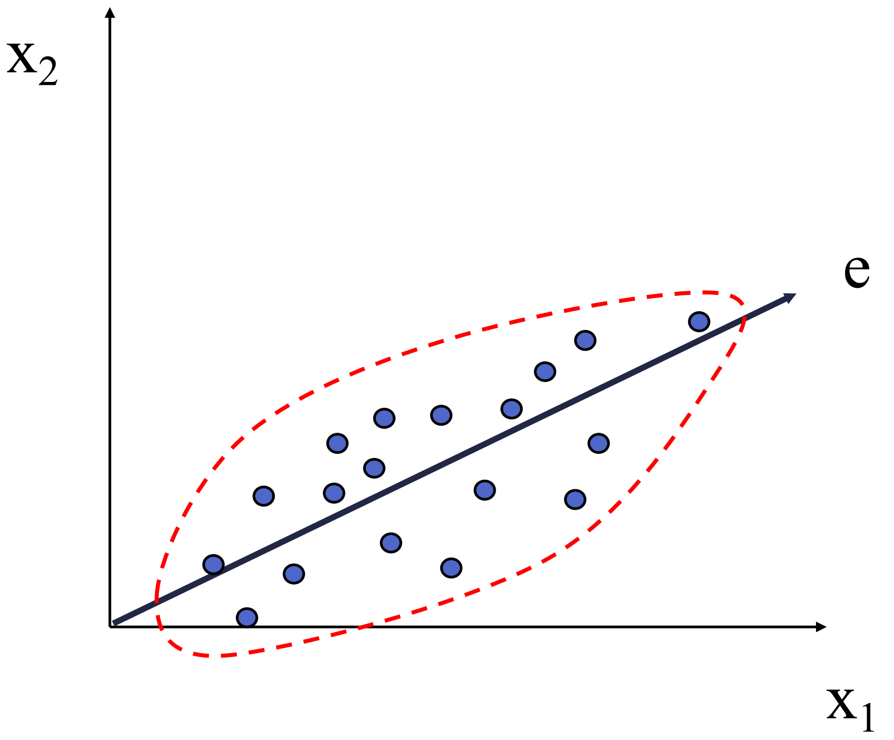
# Departemen Teknik Informatika

# Principal Component Analysis (PCA

- Transformasi data ke representasi baru dengan cara menemukan proyeksi yang menangkap variasi data terbesar
- Data asli diproyeksikan ke dimensi yang lebih kecil sehingga mengurangi ukuran dimensi.
- Mendapatkan vektor eigen dari matriks kovarians, dan vektor eigen ini digunakan untuk memproyeksikan data asli ke dimensi baru



# Principal components

- Principal component 1 (PC1)
  - The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation
- Principal component 2 (PC2)
  - the direction with maximum variation left in data,  orthogonal to the PC1
- In general, only few directions manage to capture most of the variability in the data.

# Tahapan PCA

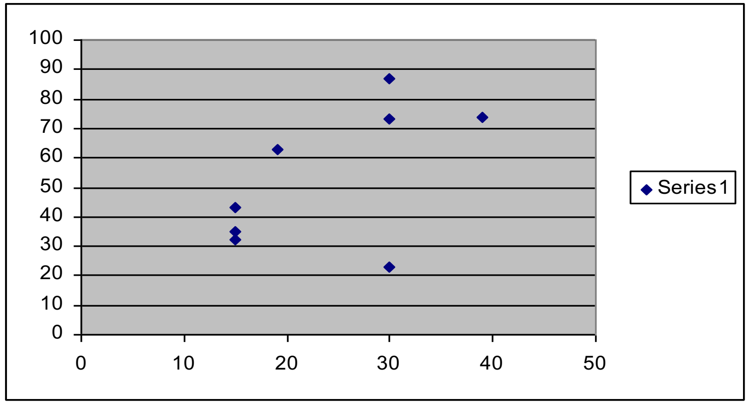
- Input: Vektor data $X$ dengan $n$-dimensi
- Tujuan: Mencari $p$ komponen utama ($p \le n$ ) yang paling baik untuk merepresentasikan data
- Tahapan:
  - Hitung $\overline{X}$ sebagai nilai rata-rata setiap fitur
  - Hitung data orisinal X dikurangi nilai rata-rata $\overline{X}$
    $X' = X - \overline{X}$
  - Hitung *Covariance matrix* C dari *X'*
  - Hitung *eigenvectors* dan *eigenvalues* dari C
    - *eigenvectors* dari C adalah **e** sedemikian hingga C**e** = λ**e**,
    - λ adalah *eigenvalue* dari C
    - C**e** = λ**e** ⇔ (C-λI)**e** = 0
  - Memilih $p$ eigenvectors dengan eigenvalues terbesar
  - Transformasi data asli dengan cara mengalikan matrik X' dengan $p$ eigenvectors

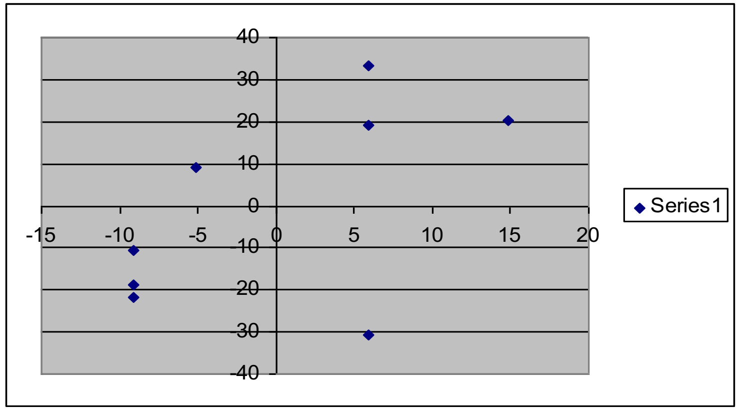
$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - \overline{x_2} \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

# Contoh Penerapan PCA

| X1 | X2 | X1' | X2' |
|---|---|---|---|
| 19 | 63 | -5.1 | 9.25 |
| 39 | 74 | 14.9 | 20.25 |
| 30 | 87 | 5.9 | 33.25 |
| 30 | 23 | 5.9 | -30.75 |
| 15 | 35 | -9.1 | -18.75 |
| 15 | 43 | -9.1 | -10.75 |
| 15 | 32 | -9.1 | -21.75 |
| 30 | 73 | 5.9 | 19.25 |

Mean X1=24.1
Mean X2=53.8



# Covariance Matrix

- C=

| | |
|---|---|
| 75 | 106 |
| 106 | 482 |

Covariance Matrix dengan 3 fitur

$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{pmatrix}$$

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

- Eigenvectors:
  - e1=(-0.98,-0.21), $\lambda$1=51.8
  - e2=(0.21,-0.98), $\lambda$2=560.2
- *Eigenvector* kedua e2 paling penting!

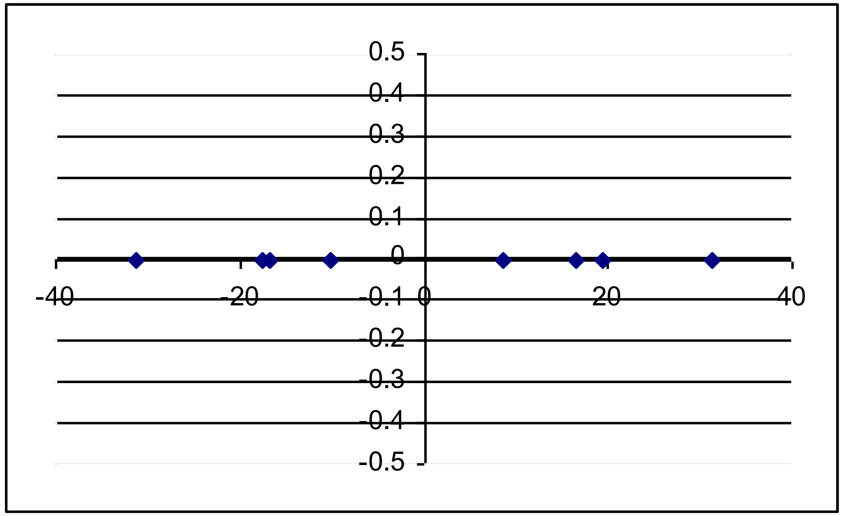
Bagaimana menghitung **e** dan $\lambda$:

- Hitung *det*(*C*-$\lambda$*I*)
- Temukan akar *det*(*C*-$\lambda$*I*)=0, nilai akar ➔ eigenvalues $\lambda$
- Selesaikan (*C*- $\lambda$*I*) **e** = 0 pada setiap $\lambda$ utk memperoleh eigenvectors **e**

# PCA dengan 1 PC (Principle Component)

| X1 | X2 | X1' | X2' |
|---|---|---|---|
| 19 | 63 | -5.1 | 9.25 |
| 39 | 74 | 14.9 | 20.25 |
| 30 | 87 | 5.9 | 33.25 |
| 30 | 23 | 5.9 | -30.75 |
| 15 | 35 | -9.1 | -18.75 |
| 15 | 43 | -9.1 | -10.75 |
| 15 | 32 | -9.1 | -21.75 |
| 30 | 73 | 5.9 | 19.25 |

- Menggunakan e2=(0.21,-0.98)
- Transformasi data yang didapatkan sbb:



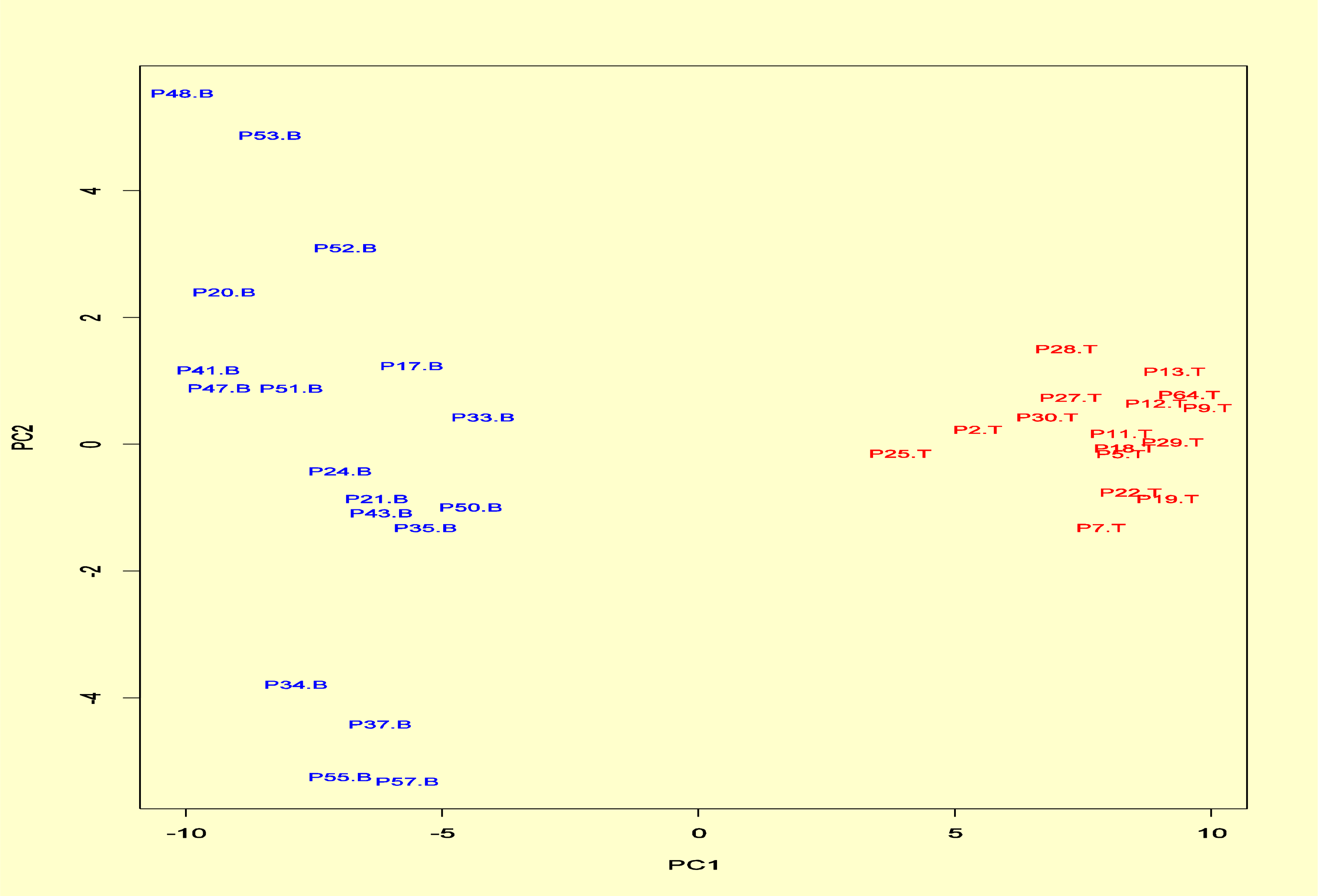

| yi |
|---|
| -10.14 |
| -16.72 |
| -31.35 |
| 31.374 |
| 16.464 |
| 8.624 |
| 19.404 |
| -17.63 |

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - \overline{x_2} \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

$$y_i = \begin{pmatrix} 0.21 & -0.98 \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = 0.21 * x_{i1} - 0.98 * x_{i2}$$

Data 34 pasien, dimensi dari 100 genes menjadi 2

# PCA pada Leukemia data dengan 100 genes serta 2 precursor B dan T



# PCA pada Citra Wajah

# Singular Value Decomposition (SVD)

$$\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \Lambda_{[r \times r]} (\mathbf{V}_{[m \times r]})^T$$

- **SVD :** metode dekomposisi sebuah matrik menjadi 3 matrik
- **A**: n x m matrik (contoh: n dokumen, m kata)
- **U**: n x r matrik orthogonal (contoh: n documen, r konsep)
- $\Lambda$: r x r matrik diagonal (r : rank dari matrik)
- **V**: m x r matrik orthogonal (contoh: m kata, r konsep)

# SVD Properties

**THEOREM** [Press+92]: always possible to decompose matrix **A** into **A** = **U** $\Lambda$ **V**$^T$, where:

- **U**, $\Lambda$, **V**: unique
- **U**, **V**: column orthonormal (orthogonal to each other)
  - **U**$^T$**U** = **I**; **V**$^T$**V** = **I** (**I:** identity matrix)
- $\Lambda$: singular value are positive, and sorted in decreasing order

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \emptyset \\ \emptyset & \lambda_2 \end{bmatrix} \times \begin{bmatrix} \text{---} & v_1 & \text{---} \\ \text{---} & v_2 & \text{---} \end{bmatrix}$$

# Contoh Penerapan SVD

| term / document | data | information | retrieval | brain | lung |
|---|---|---|---|---|---|
| CS-TR1 | 1 | 1 | 1 | 0 | 0 |
| CS-TR2 | 2 | 2 | 2 | 0 | 0 |
| CS-TR3 | 1 | 1 | 1 | 0 | 0 |
| CS-TR4 | 5 | 5 | 5 | 0 | 0 |
| MED-TR1 | 0 | 0 | 0 | 2 | 2 |
| MED-TR2 | 0 | 0 | 0 | 3 | 3 |
| MED-TR3 | 0 | 0 | 0 | 1 | 1 |

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Tugas:

#1: Menemukan konsep pada teks

#2: Mengurangi dimensi

# Contoh Penerapan SVD

#1: Menemukan konsep pada teks

$$A = U \Lambda V^T$$

doc-to-concept

CS-concept

MED-concept

'strength' of CS-concept

$$
\begin{array}{c c}
 & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \text{CS} \\ \\ \\ \\ \text{MED} \\ \\ \\ \end{array} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
=
\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\times
\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
\times
\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

term-to-concept

# Contoh Penerapan SVD

#2: Mengurangi dimensi

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & \cancel{0} \\ 0.36 & \cancel{0} \\ 0.18 & \cancel{0} \\ 0.90 & \cancel{0} \\ 0 & \cancel{0.53} \\ 0 & \cancel{0.80} \\ 0 & \cancel{0.27} \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & \cancel{5.29} \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0.71} & \cancel{0.71} \end{bmatrix}
$$

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 9.64 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}
$$

# Contoh Penerapan SVD

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

#2: Mengurangi dimensi

# SVD Step

- https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm

# Referensi


- *C. Faloutsos*, “Powerful tools for Data Mining fractals, power laws, SVD”, Carnegie Mellon University, 2004.
- *Jure Leskovec*, “Dimensionality reduction: PCA, SVD, MDS, ICA, and friends”, Machine Learning recitation, 2006.