Lecture 8

Monday, February 7, 2022 9:32 AM

Principal Component Analysis

Admin

① HW due ~~today~~ Thurs.

Reading

① Blind source separation

② Multi- and megavariate data analysis, 2003, Eriksson.

Objectives

① Discuss datasets consisting of many variables from which measurements are arrising from a lower-dimensional underlying process.

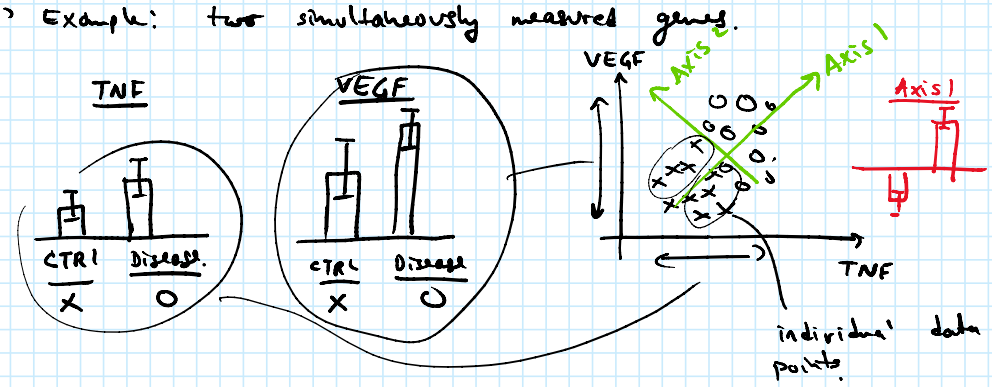
② Define principal component analysis (PCA)

③ Consider examples

Large datasets arrising from lower-dimensional processes

- In many cases, we consider datasets in which we collect many 'data points' that are highly correlated with one another.
  - → multi-dimensional gene expression data from healthy & diseased subjects
  - → Mutiple microphones recording from different parts of a room → "cocktail party problem"
  - → multiple ECG recordings from a pregnant woman.
- In each case, the measurements are highly related to one another, but represent contributions from underlying processes composed of
  - → together drive each sensor reading.

→ Example: two simultaneously measured genes.

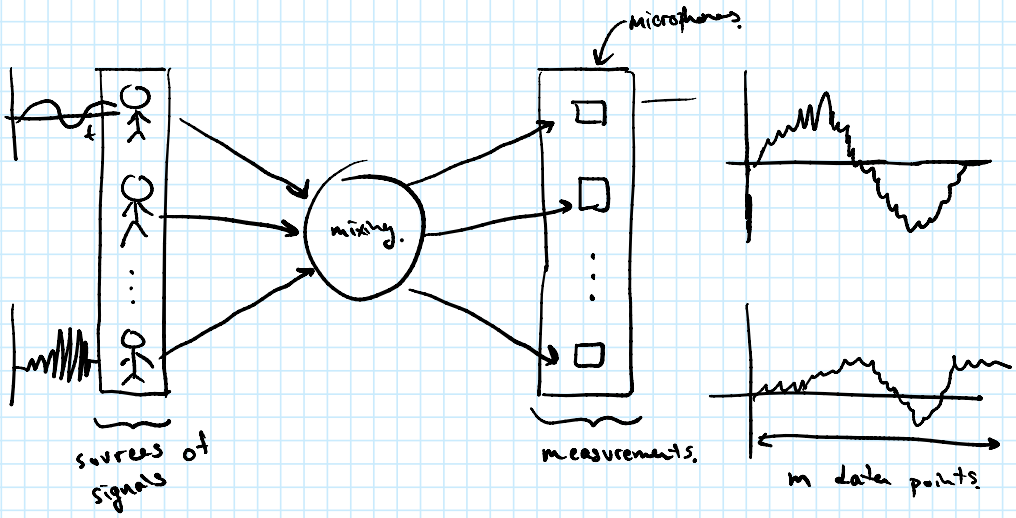


- Can identify a composite axis (Axis 1) that more clearly separates groups.

  ⇒ our approach can better distinguish sources of variation in underlying data.

Example — The cocktail party problem

Example – The cocktail party problem



Principal Component Analysis

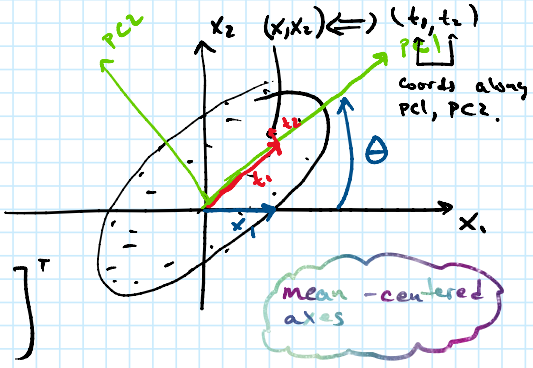
- problem: multi-dimensional dataset

$$X - m \times n$$

(m: # of observation; n: # of variables.)

- Goal is to represent the data in a lower-dimensional plane
  - → First described by Cauchy, formulated as finding lines/planes in data of closest fit by Pearson.
  - → Representation in lower-dimensional plane will reveal groups of differences between samples, sensors, that would not otherwise be obvious.

- We can re-express the data in a new coord. system by rotating.



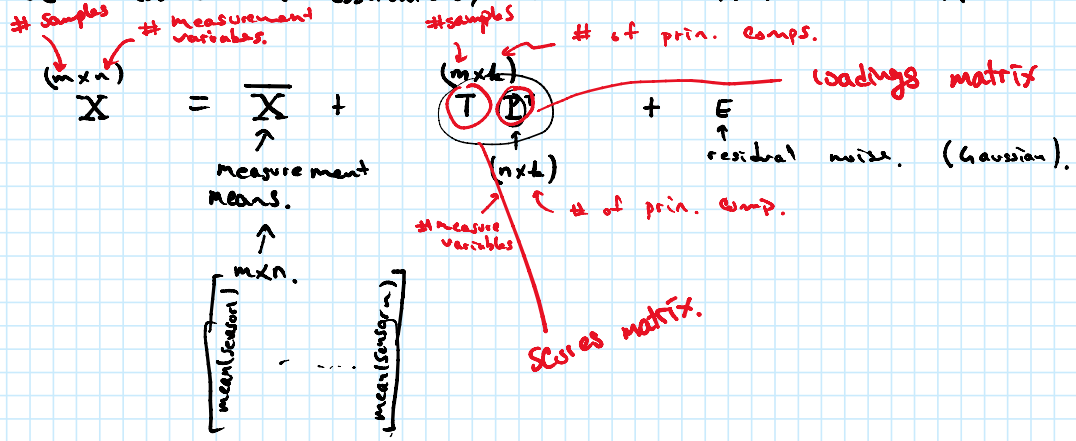
$$x_1 = t_1 \cos\theta + t_2 \sin\theta$$
$$x_2 = -t_1 \sin\theta + t_2 \cos\theta$$

$$\underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{original data}} = \underbrace{\begin{bmatrix} t_1 & t_2 \end{bmatrix}}_{\text{PCA coords}} \underbrace{\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^T}_{T^T}$$

T^T: orthogonal rotation matrix.

- We can choose our rotation, θ, to:
  1) maximize variation (variance) in the data explained by PC1.
  2) reduce covariance between variables (PC1, PC2, ...)

- We measure m observations, want to transform them



- We often pre-treat the data
  - subtract column means $\Rightarrow \bar{X} = 0$
  - scale by by standard dev.

$\left.\begin{matrix} \bar{X} = 0 \\ \text{z-score} \end{matrix}\right\} \tilde{X}_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$

- How do we compute T, P
  - → In PCA, we choose $E[t_{\square,i}\ t_{\square,j}] = \text{Cov}(\text{col } i, \text{col } j) = 0, i \neq j$ (two cols.)

- To do so, compute cov matrix of $\tilde{X}$

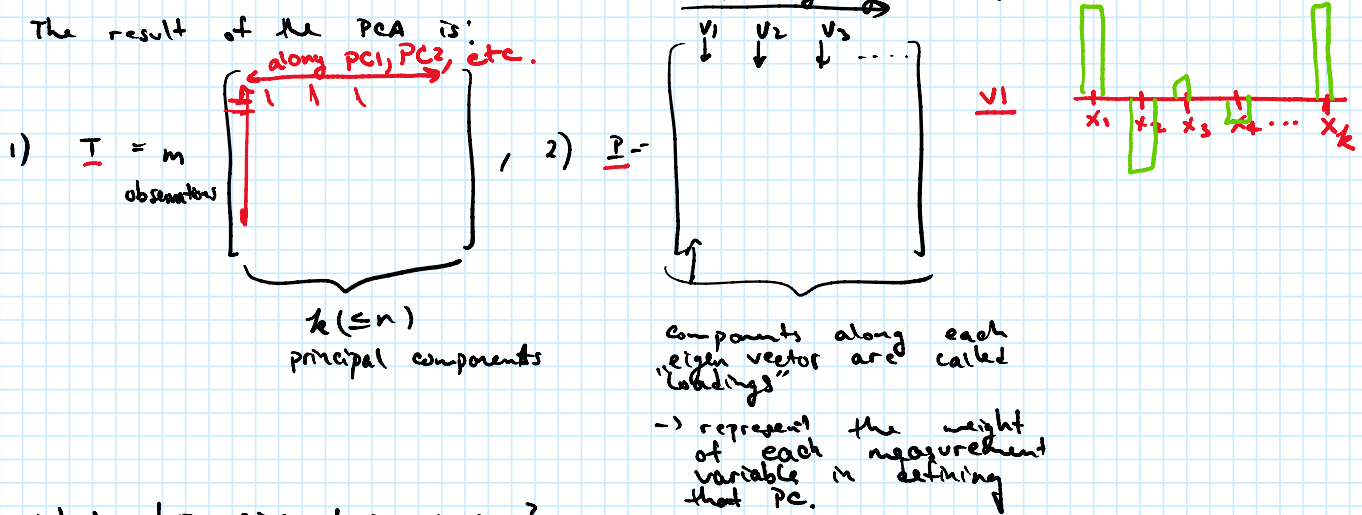
$C_{i,j} = \frac{1}{m} \sum_{k=1}^{m} x_{k,i}\, x_{k,j} \Rightarrow \underset{\sim}{C} = \frac{1}{m} \tilde{X}^T \tilde{X}$ ← want to diagonalize.

(col i, col j)

$\underset{\sim}{C} V = \lambda V$ ← eigenvectors

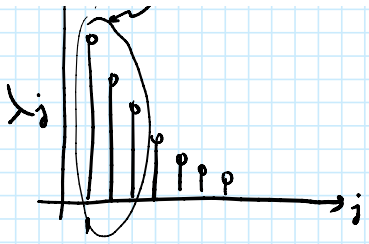
(components along $x_1, x_2, \ldots$; eigenvalues)

- The result of the PCA is:



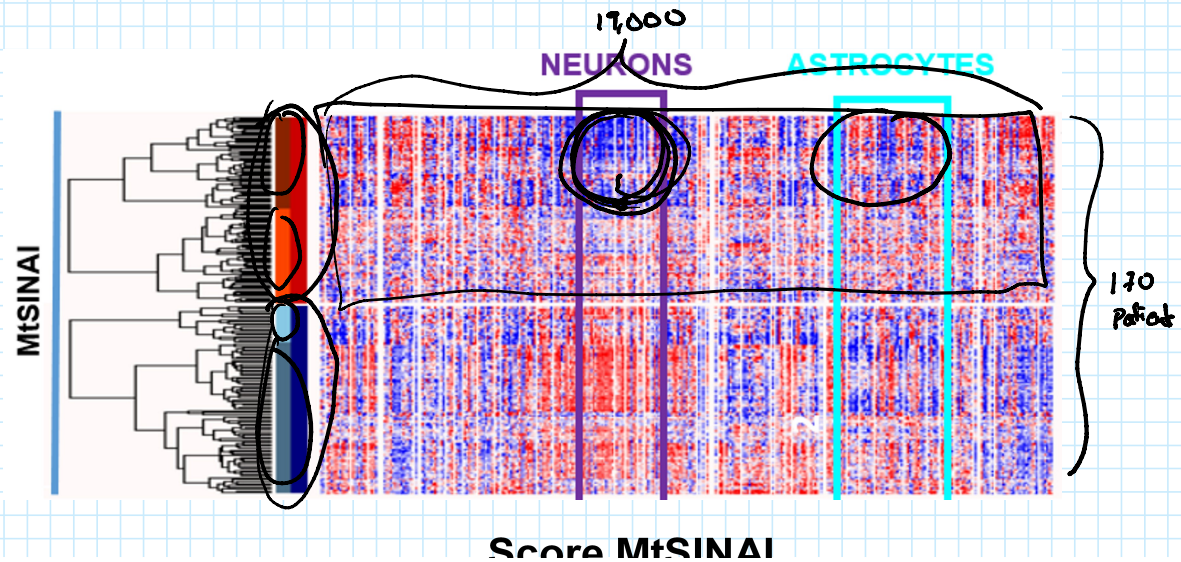
- What do eigenvalues mean?

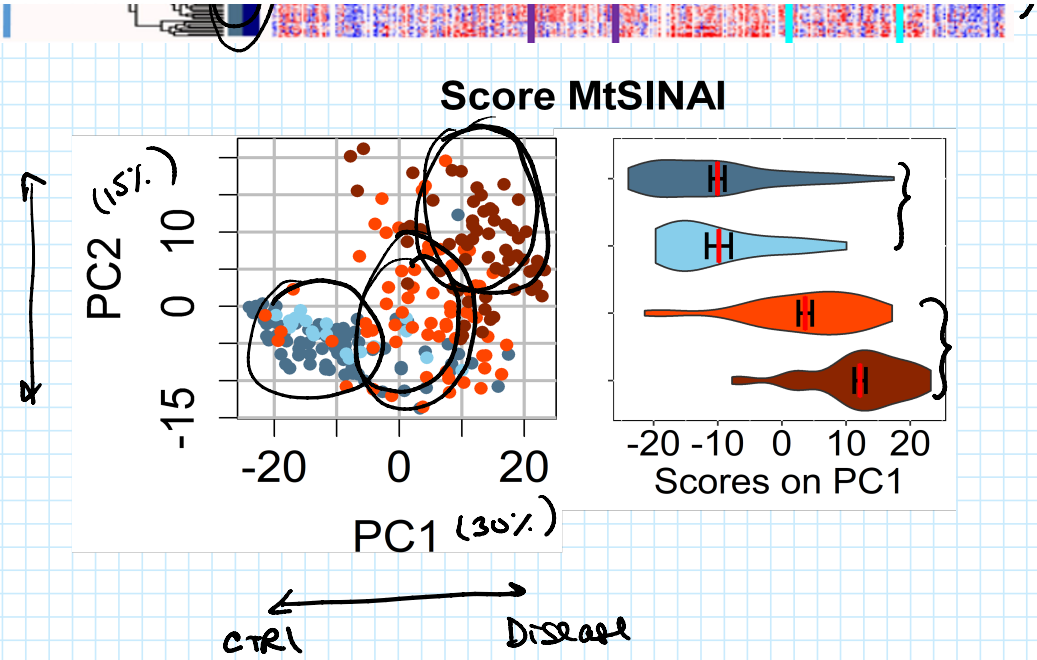
contributes a lot of variance.


λ_j
j

SamplePCA (R) (Class Discovery package)

MATLAB PCA




Score MtSINAI
PC2 (15%)
10
0
-15
-20
0
20
PC1 (30%)
-20 -10 0 10 20
Scores on PC1
CTRL
Disease