

# GL Applied Data Science Program

## Data Collection and Visualization for Exploratory Data Analysis

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Introduction



<http://www.carolineuhler.com>

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Overview

## Overview of this week / module:

- Data collection and visualization for exploratory data analysis
- Network analysis
- Unsupervised learning - clustering

## Overview of this lecture:

- Data collection: Mammography case study
- Hypothesis testing
- Visualizing high-dimensional data for exploratory data analysis

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

## Case study: Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
  - Mammography: screening women for breast cancer by X-rays
- 
- \* Does mammography speed up detection by enough to matter?
  - \* How would you approach this problem? What is important when setting up a study / experiment?

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

## Case study: Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
  - Mammography: screening women for breast cancer by X-rays
- \* Does mammography speed up detection by enough to matter?
- \* How would you approach this problem? What is important when setting up a study / experiment?
- ⇒ Perform a **controlled randomized experiment** to minimize the problem of **confounding**

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
<b>Treatment</b>					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

**Reference:** D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
<b>Treatment</b>					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

**Reference:** D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

This file is meant for personal use by nehakinjal@gmail.com only.  
Which rates should be compared to show the efficacy of treatment?  
Sharing or publishing the contents in part or full is liable for legal action.

Which rates should be compared to show the efficacy of treatment?

- Seems natural to compare those who accepted screening to those who refused or the control group
  - But this is an **observational** comparison!
  - Becomes clear when comparing the death rates from all other causes
  - Instead compare the whole treatment group against the whole control group (i.e., compare the numbers 1.3 versus 2.0)
- \* **Intention-to-treat analysis**

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ( $= \frac{63}{31000}$ )
- Death rate from breast cancer in treatment group: 0.0013 ( $= \frac{39}{31000}$ )

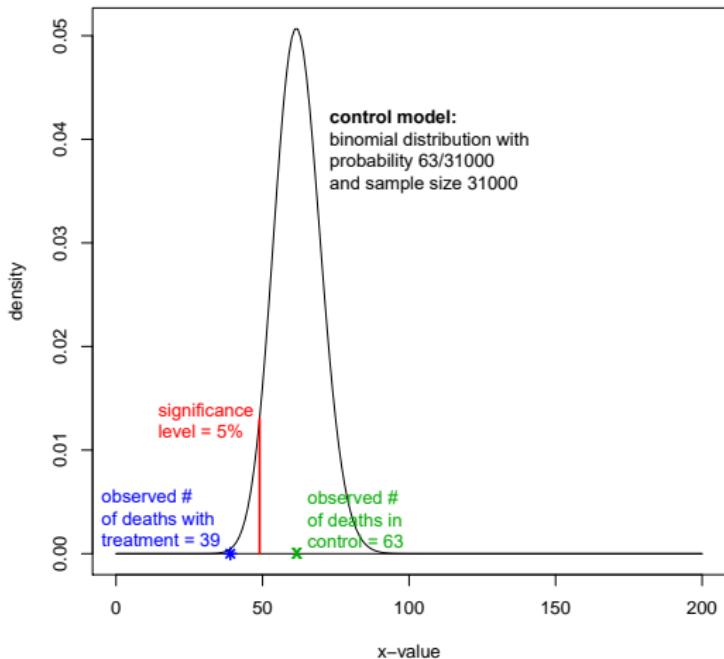
Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

⇒ Perform a **hypothesis test**

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Hypothesis testing



**p-value:** probability under the control model to observe  $\leq 39$  deaths is 0.0012; this is too unlikely to happen by chance; thus introducing mammography significantly reduced the number of breast cancer deaths

This file is meant for personal use by [nehakinal@gmail.com](mailto:nehakinal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Hypothesis testing applications outside of healthcare

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Hypothesis testing applications outside of healthcare

- Quality management in manufacturing environments: deciding whether new process, technique, method is likely to change number of defective products
- Finance: deciding which investment / instrument is likely to provide satisfactory return
- Advertising: deciding whether an advertising campaign, marketing technique, etc. is likely to increase sales
- Business: make informed decisions on which initiatives help grow your business

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

## Example research findings

*Giovannucci et al., Journal of the National Cancer Institute 87 (1995):*

Intake of tomato sauce ( $p$ -value of 0.001), tomatoes ( $p$ -value of 0.03), and pizza ( $p$ -value of 0.05) reduce the risk of prostate cancer;

But for example tomato juice ( $p$ -value of 0.67), or cooked spinach ( $p$ -value of 0.51), and many other vegetables are not significant.

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Wonder-pill

- randomized group of 1000 people
- measure 100 variables before and after taking the pill: weight, blood pressure, etc.
- perform a hypothesis test with a significance level of 5%

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

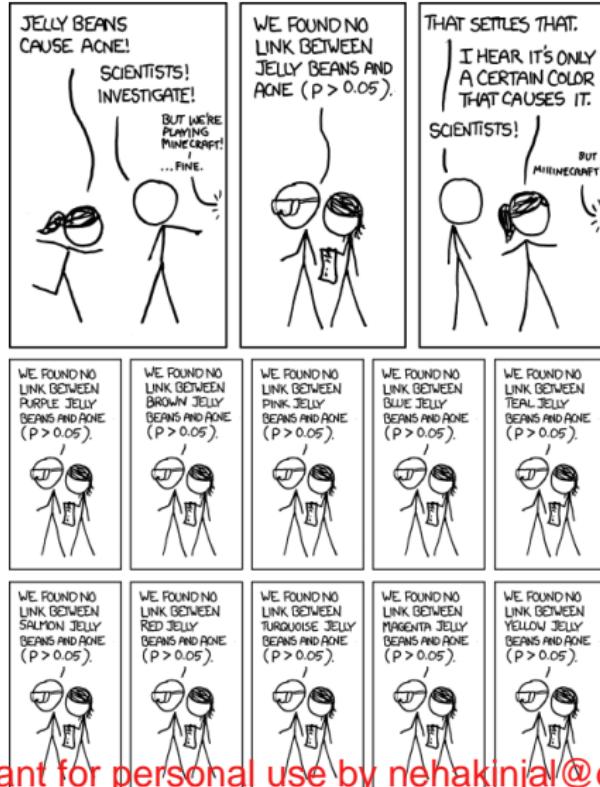
# Wonder-pill

- randomized group of 1000 people
- measure 100 variables before and after taking the pill: weight, blood pressure, etc.
- perform a hypothesis test with a significance level of 5%
- $V := \# \text{ false significant tests}$ :  $V \sim \text{Binomial}(100, 0.05)$

⇒ in average 5 out of 100 variables show a significant effect!

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

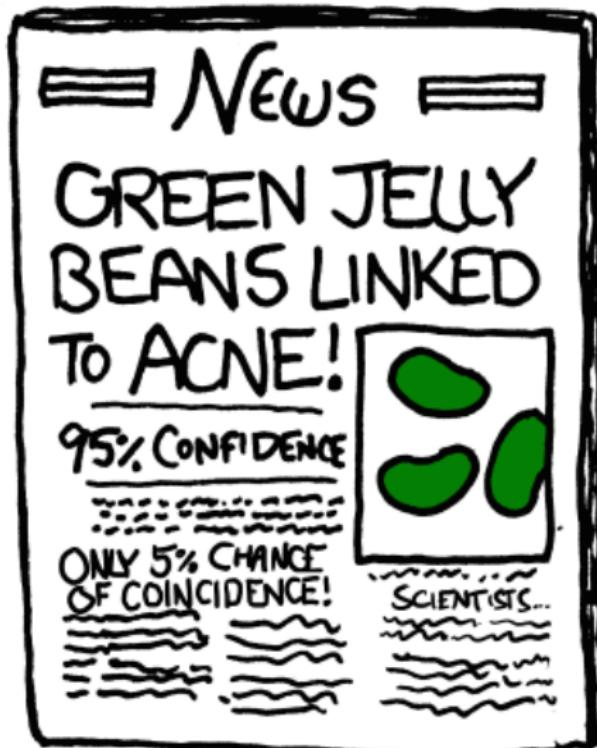
# Jelly Beans and Acne



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Problematic of selective inference



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

<http://imgs.xkcd.com/comics/significant.png>

# Different protection levels

Compute  $p$ -values using methods that control:

- family-wise error rate (FWER)  $\leq \alpha$ , where

$$\text{FWER} = \mathbb{P}(\text{at least one false significant result})$$

- false discovery rate (FDR)  $\leq \alpha$ , where

FDR = expected fraction of false significant results  
among all significant results

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Corrections for multiple testing (math optional)

## Bonferroni correction:

- Reject  $H_0$  when:  $m \cdot p\text{-value} \leq \alpha$   
where  $m$  is the total number of hypothesis tests performed
- Bonferroni correction implies FWER  $\leq \alpha$

## Holm-Bonferroni correction:

- Sort  $p$ -values in increasing order:  $p_{(1)} \leq \dots \leq p_{(m)}$
- Reject  $H_0$  when:  $(m - i + 1)p_{(i)} \leq \alpha$  (more power than Bonferroni)
- Holm-Bonferroni correction implies FWER  $\leq \alpha$

## Benjamini-Hochberg correction:

- Sort  $p$ -values in increasing order:  $p_{(1)} \leq \dots \leq p_{(m)}$
  - Reject  $H_0$  when:  $mp_{(i)}/i \leq \alpha$
- This file is meant for personal use by nehakinjal@gmail.com only.
- Sharing or publishing the contents in part or full is liable for legal action.

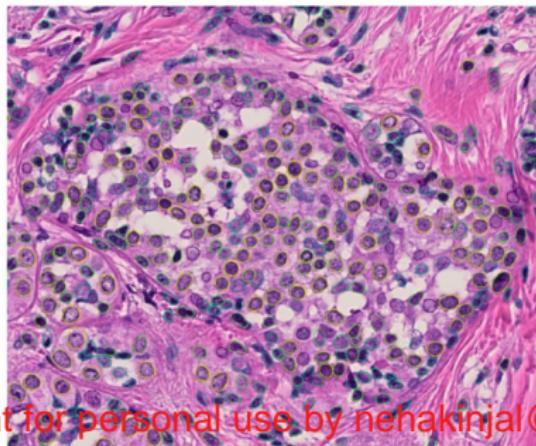
## Commonly accepted practice

- No correction for multiple testing when generating hypotheses (but report number of tests performed)
- $\text{FDR} \leq 10\%$  in exploratory analysis or screening
  - balance between high power and low # of false significant results
- $\text{FWER} \leq 5\%$  in confirmatory analysis
  - food and drug administration (FDA)

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Application: Microscopy Images

- Microscopy images of human tissue slices
- Crop cells ( $n$  cells) and summarize each cell by 100 different texture features (i.e.,  $D = 100$ )
- How can we visualize this data set to find clusters or abnormal cells?
- **Input:**  $x_1, \dots, x_n \in \mathbb{R}^D$ ,    **Output:**  $y_1, \dots, y_n \in \mathbb{R}^d$ , where  $d \ll D$

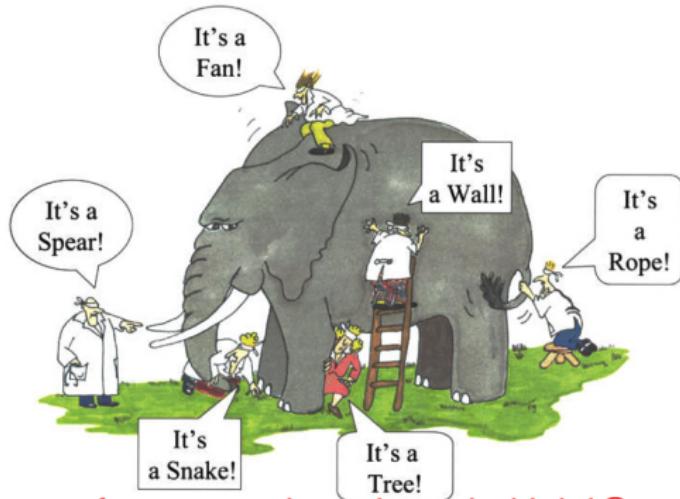


This file is meant for personal use by nehakinjal@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

## 2 different approaches

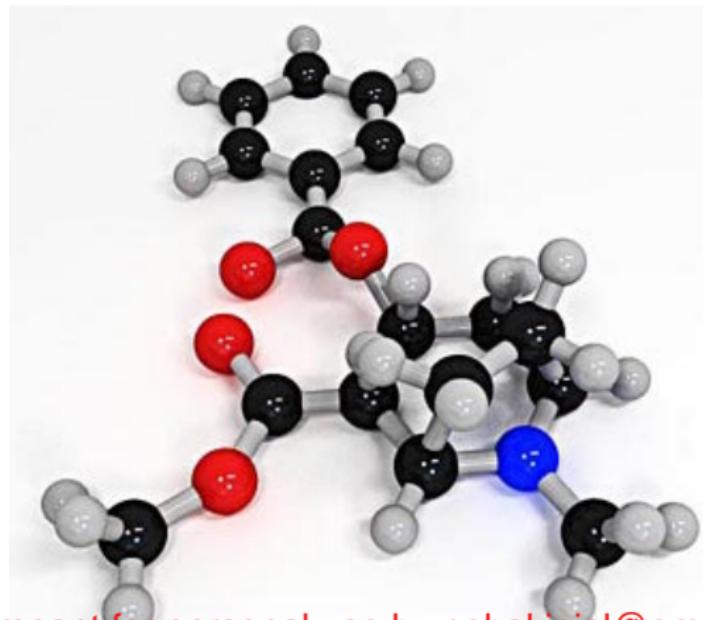
- Principal component analysis: projection that spreads data as much as possible
- Stochastic neighbor embedding: non-linear embedding that tries to keep close-by points close



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Principal Component Analysis

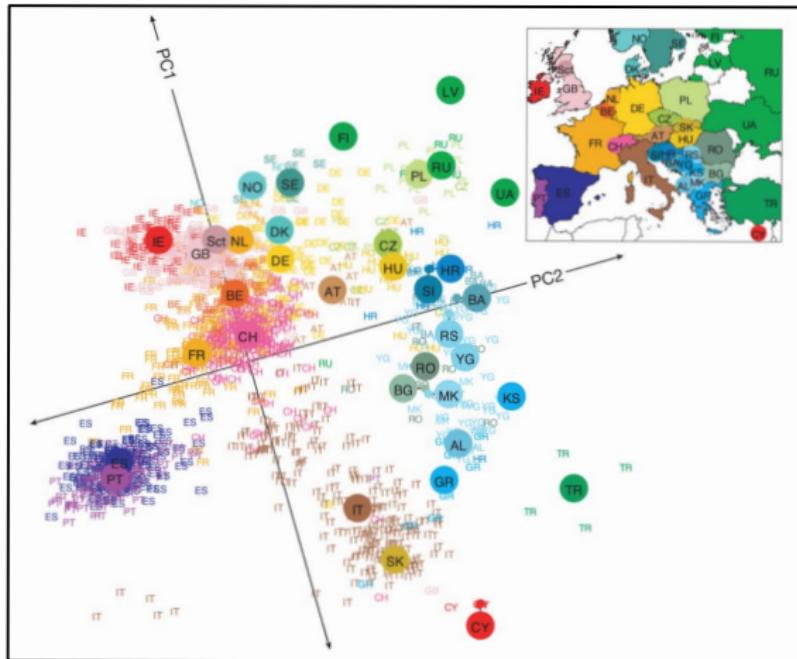
- **Goal:** Dimension reduction to a few dimensions
- **Intuition:** Find low-dimensional projection with largest spread



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# PCA application



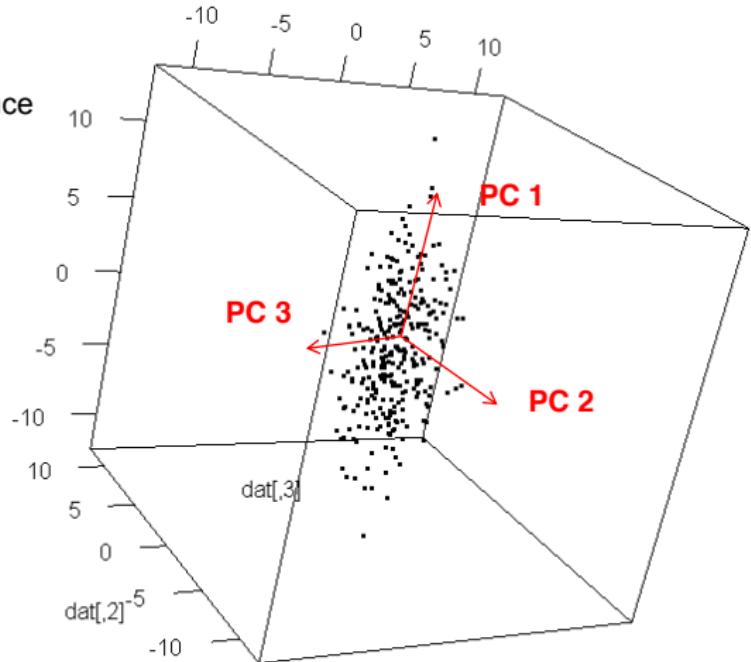
Reference: J. Novembre et al., *Genes mirror geography within Europe*, Nature 456 (2008).  
This file is meant for personal use by [nihakimjai@gmail.com](mailto:nihakimjai@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

# Definition 1: Maximize projection variance

Start with centered data  $X \in \mathbb{R}^{n \times p}$

- PC 1 is direction of largest variance
- PC 2 is
  - perpendicular to PC 1
  - again largest variance
- PC 3 is
  - perpendicular to PC 1, PC 2
  - again largest variance
- etc.

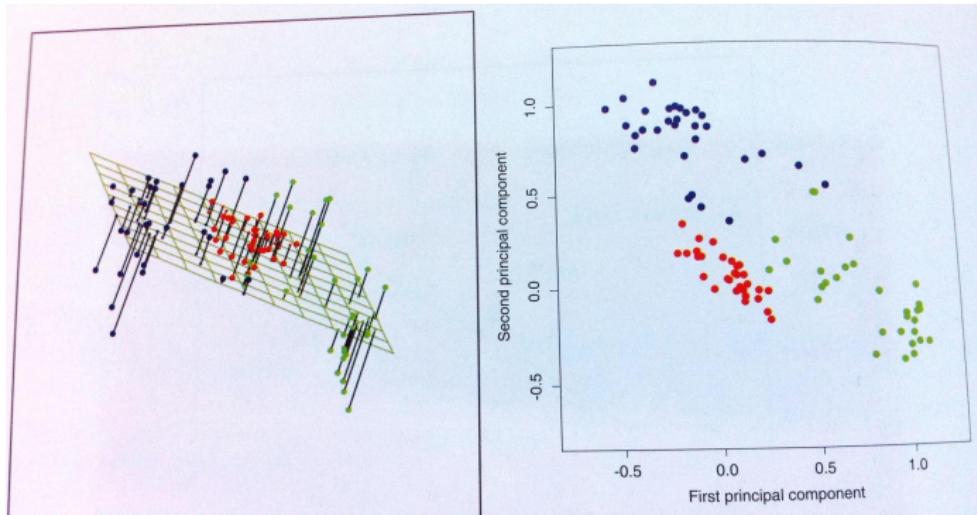


This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

## Definition 2: Minimize projection residuals

- PC 1: Straight line with smallest orthogonal distance to all points
- PC 1 & PC 2: Plane with smallest orthogonal distance to all points
- etc.



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

## Definition 3: Spectral decomposition

- Covariance matrix (or correlation matrix)  $R = \frac{1}{n}X^T X$  is symmetric and positive semidefinite
- **Spectral Decomposition Theorem:** Every real symmetric matrix  $R$  can be decomposed as

$$R = V\Lambda V^T,$$

where  $\Lambda$  is diagonal and  $V$  is orthogonal

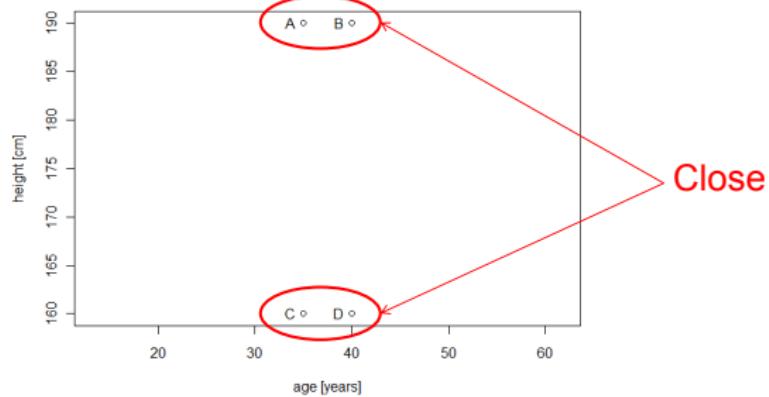
- Columns of  $V$  (= eigenvectors of  $R$ ) are the PCs
- Diagonal entries of  $\Lambda$  (= eigenvalues of  $R$ ) are variances along PCs

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Covariance versus correlation - to scale or not to scale

- Using covariance will find the variable with largest spread as 1. PC
- Use correlation, if different units are compared

Person	Age (years)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160

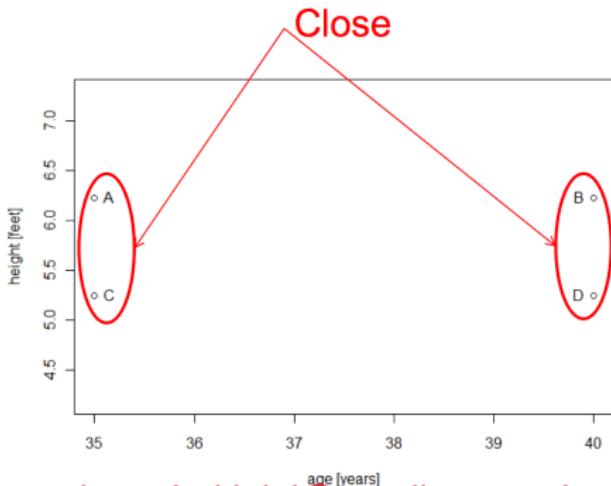


This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Covariance versus correlation - to scale or not to scale

- Using covariance will find the variable with largest spread as 1. PC
- Use correlation, if different units are compared

Person	Age (years)	Height (feet)
A	35	6.232
B	40	6.232
C	35	5.248
D	40	5.248

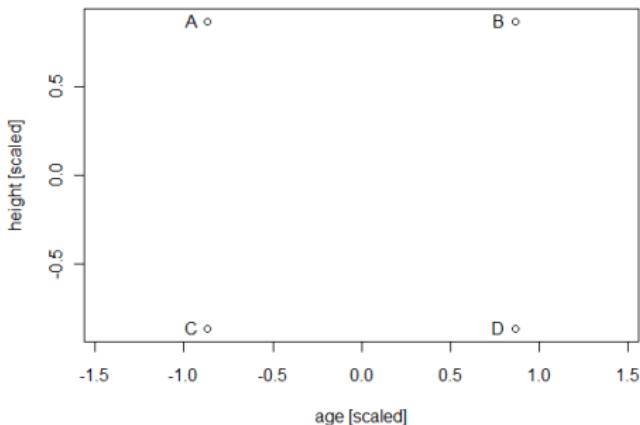


This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

## Covariance versus correlation - to scale or not to scale

- Using covariance will find the variable with largest spread as 1. PC
- Use correlation, if different units are compared

Person	Age (years)	Height (feet)
A	-0.87	0.87
B	0.87	0.87
C	-0.87	-0.87
D	0.87	-0.87



This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

## Stochastic neighbor embedding (tSNE)

- probabilistic approach to place samples from high-dimensional space into low-dimensional space so as to preserve the identity of neighbors
- find embedding so that original high-dimensional sample distribution is approximated well by resulting low-dimensional sample distribution (tSNE uses *Kullback-Leibler divergence* to measure "distance" between distributions and minimizes this objective function)

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

## Stochastic neighbor embedding (tSNE)

- probabilistic approach to place samples from high-dimensional space into low-dimensional space so as to preserve the identity of neighbors
- find embedding so that original high-dimensional sample distribution is approximated well by resulting low-dimensional sample distribution (tSNE uses *Kullback-Leibler divergence* to measure "distance" between distributions and minimizes this objective function)
- gives rise to **non-linear embedding** where close-by points remain close-by and far away points remain far away, so that **clusters are preserved**
- non-linearity can reduce the problem of crowding often observed in PCA: moderate distance in high-dim. space can be faithfully modeled by much larger distance in low-dim. space

This file is meant for personal use by [nakinjal@gmail.com](mailto:nakinjal@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

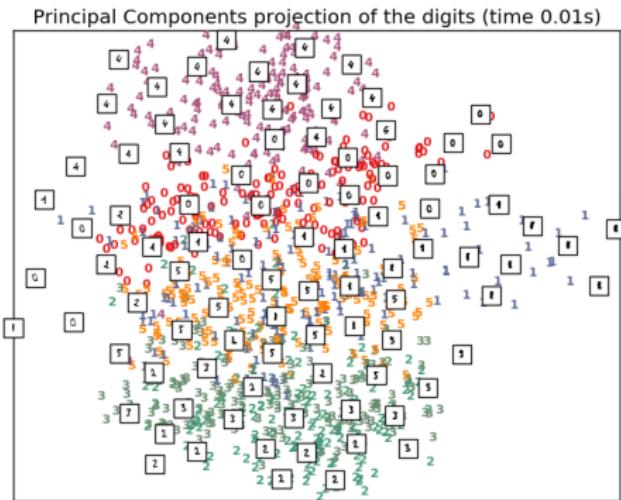
# Case study: Digit recognition

- $\sim 1800$  hand-written digits (i.e.,  $n \approx 180$  for each number)
- each (centered) digit was put in a  $8 \times 8$ -grid (i.e.,  $D = 64$ )
- measure grey value in each part of the grid, i.e. 64 grey values
- **Input:**  $x_1, \dots, x_n \in \mathbb{R}^D$ ,    **Output:**  $y_1, \dots, y_n \in \mathbb{R}^d$ , where  $d \ll D$

A selection from the 64-dimensional digits dataset

0	1	2	3	4	5	0	1	2	3	4	5	0	5
5	5	0	4	1	3	5	1	0	0	2	2	0	1
4	4	1	5	0	5	2	0	0	1	3	2	1	4
3	1	4	0	5	3	1	5	4	4	2	2	2	5
2	3	4	5	0	1	2	3	4	5	0	1	2	3
0	4	1	3	5	1	0	0	2	2	1	0	1	2
1	5	0	5	2	2	0	0	1	3	2	1	3	4
0	5	7	4	5	4	4	1	2	2	5	5	4	4
5	0	1	2	3	4	5	0	1	2	3	4	5	0
3	5	4	0	0	2	2	2	0	1	2	3	3	3
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
3	5	4	0	0	2	2	2	0	1	2	3	3	3
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5	0	1
5	1	0	0	2	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	1	3	1	3	4	4
3	1	5	4	2	2	2	5	5	4	4	0	0	1
0	1	2	3	4	5	0	1	2	3	4	5		

# Case study: Digit recognition

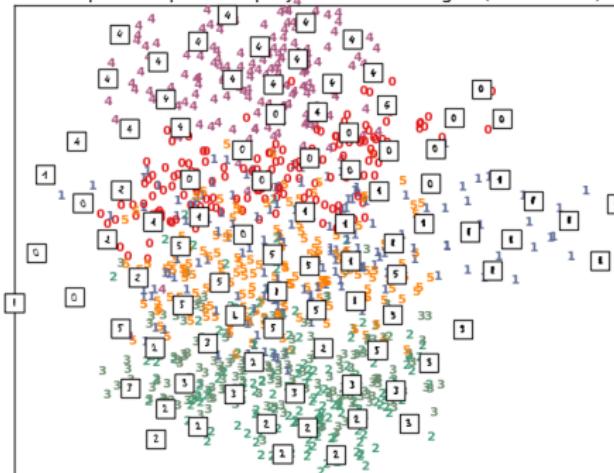


This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

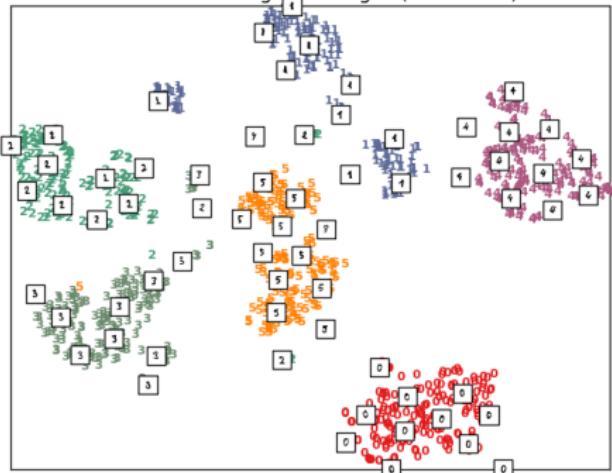
Sharing or publishing the contents in part or full is liable for legal action.

# Case study: Digit recognition

Principal Components projection of the digits (time 0.01s)



t-SNE embedding of the digits (time 5.70s)



- tSNE seems to find meaningful clusters
- Note: tSNE embedding is result of non-convex optimization problem: depends on starting configuration and computation takes longer

For code and figures see

This file is meant for personal use by [nehakinjal@gmail.com](mailto:nehakinjal@gmail.com) only.

[http://scikit-learn.org/stable/auto\\_examples/manifold/lle\\_digits.html](http://scikit-learn.org/stable/auto_examples/manifold/lle_digits.html)

Sharing or publishing the contents in part or full is liable for legal action.

## References

- For a statistics textbook, including controlled experiments and observational studies (chapters 1 and 2) and hypothesis testing (chapter 26-29):
  - D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.
- For selective inference and correcting for multiple hypothesis testing: Lecture by Yoav Benjamini, THE expert for multiple testing issues:  
<http://simons.berkeley.edu/talks/yoav-benjamini-2013-12-11a>
- For PCA and other projection methods:
  - T. Hastie, R. Tibshirani & J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- For tSNE:
  - L. van der Maaten & G. E. Hinton. *Visualizing Data using t-SNE*. JMLR, 2008.
  - G. E. Hinton & S. T. Roweis. *Stochastic Neighbor Embedding*. NIPS, 2002.  
This file is meant for personal use by hchakraborty@gmail.com only  
Sharing or publishing the contents in part or full is liable for legal action.