GL Applied Data Science Program

Data Collection and Visualization for Exploratory Data Analysis

Introduction

http://www.carolineuhler.com

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

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?

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?

$\Rightarrow$ Perform a **controlled randomized experiment** to minimize the problem of **confounding**

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 No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
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.

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 No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
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.

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

HIP study

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**

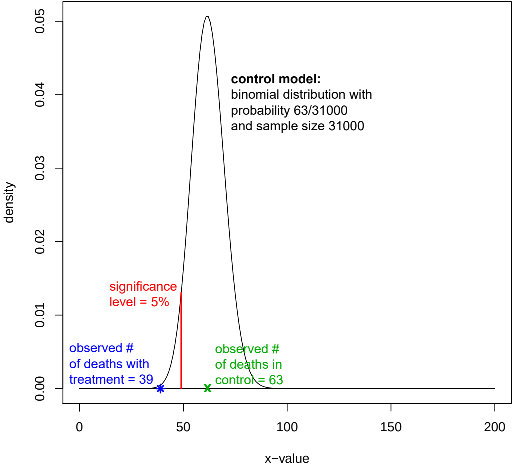
Hypothesis testing

- 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?

$\Rightarrow$ Perform a **hypothesis test**

Hypothesis testing



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

Hypothesis testing applications outside of healthcare

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

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.

Wonder-pill

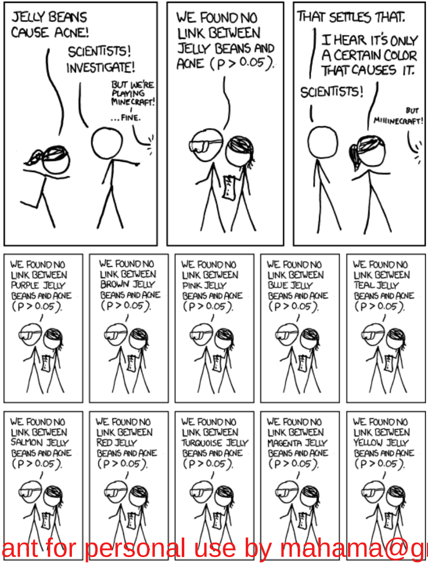
- randomized group of 1000 people
- measure 100 variables before and after taking the pill: weight, blood pressure, etc.
- perform a hypothesis test with a significance level of 5%

Wonder-pill

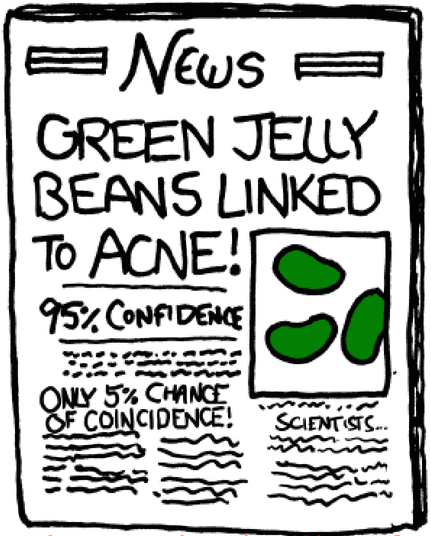
- randomized group of 1000 people
- measure 100 variables before and after taking the pill: weight, blood pressure, etc.
- perform a hypothesis test with a significance level of 5%
- $V := \#$ false significant tests: $\quad V \sim \mathrm{Binomial}(100, 0.05)$

$\Rightarrow$ in average 5 out of 100 variables show a significant effect!

Jelly Beans and Acne



Problematic of selective inference



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

$$\begin{aligned}\text{FDR} \ &= \text{expected fraction of false significant results} \\ &\quad\ \text{among all significant results}\end{aligned}$$

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 $\mathrm{FWER} \leq \alpha$

Holm-Bonferroni correction:

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

Benjamini-Hochberg correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $mp_{(i)}/i \leq \alpha$
- Benjamini-Hochberg correction implies $\mathrm{FDR} \leq \alpha$

Commonly accepted practice

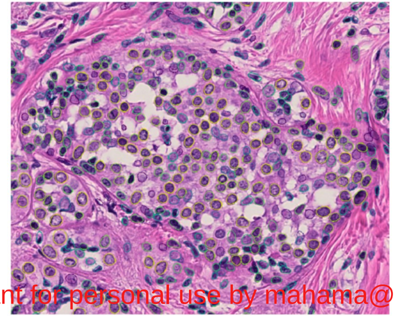
- No correction for multiple testing when generating hypotheses (but report number of tests performed)

- $\mathrm{FDR} \leq 10\%$ in exploratory analysis or screening
  - balance between high power and low # of false significant results

- $\mathrm{FWER} \leq 5\%$ in confirmatory analysis
  - food and drug administration (FDA)

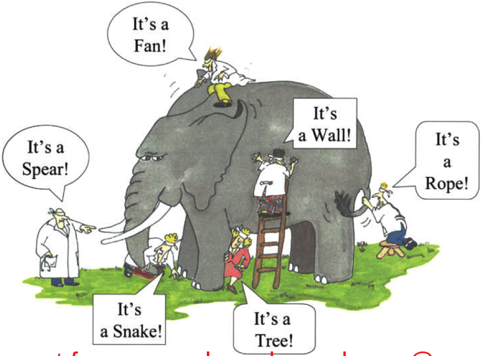
Application: Microscopy Images

- 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:** $x_1, \ldots, x_n \in \mathbb{R}^D$,   **Output:** $y_1, \ldots, y_n \in \mathbb{R}^d$, where $d << D$



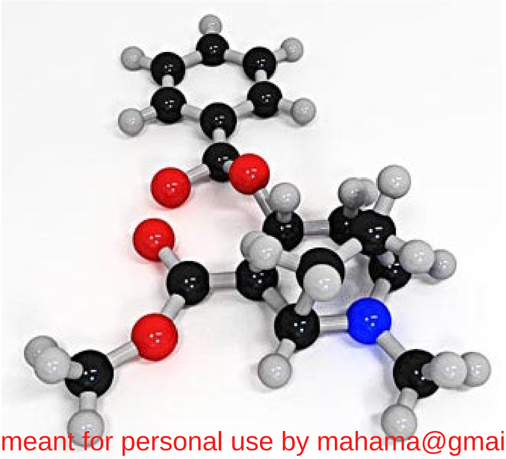
2 different approaches

- 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

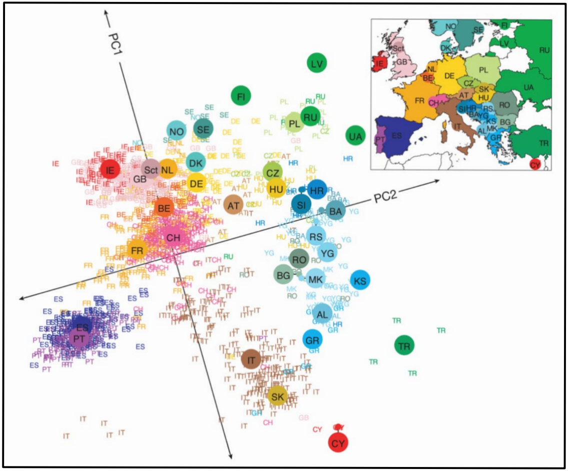


Principal Component Analysis

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



PCA application

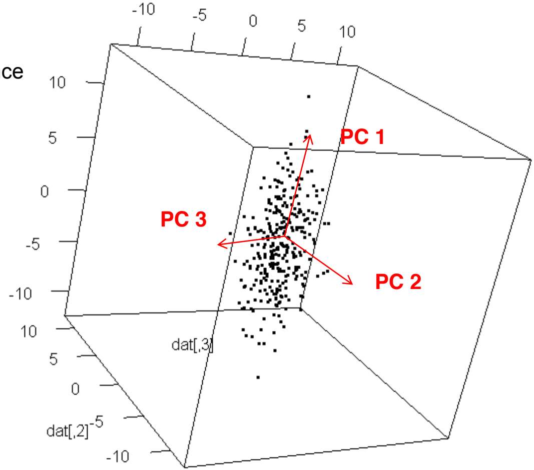


Reference: J. Novembre et al., *Genes mirror geography within Europe*, Nature 456 (2008).

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.



Definition 2: Minimize projection residuals

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

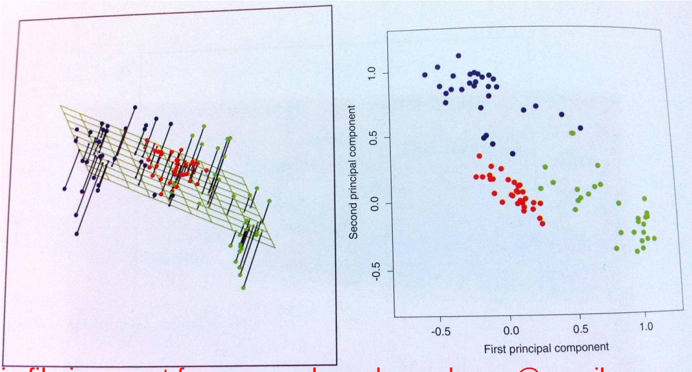


Figure from *Elements of Statistical Learning* by Hastie and Tibshirani

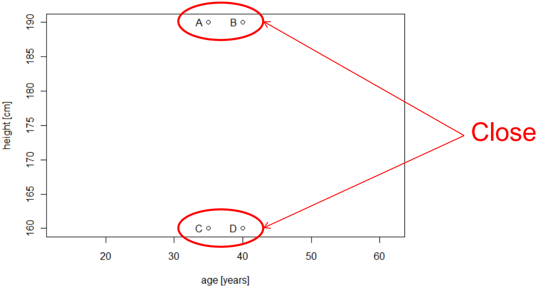
Definition 3: Spectral decomposition

- Covariance matrix (or correlation matrix) $R = \frac{1}{n}X^TX$ is symmetric and positive semidefinite
- **Spectral Decomposition Theorem:** Every real symmetric matrix R can be decomposed as
$$R = V\Lambda V^T,$$
where Λ is diagonal and V is orthogonal
- Columns of V (= eigenvectors of R) are the PCs
- Diagonal entries of Λ (= eigenvalues of R) are variances along PCs

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

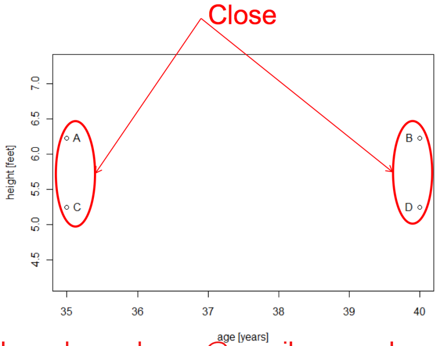
Person	Age (years)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160



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

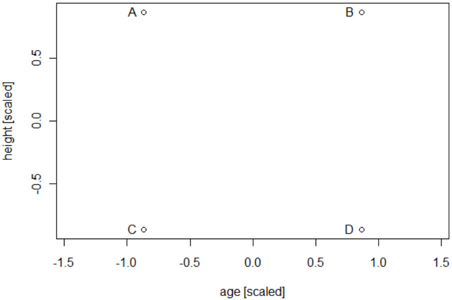
Person	Age (years)	Height (feet)
A	35	6.232
B	40	6.232
C	35	5.248
D	40	5.248



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



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)

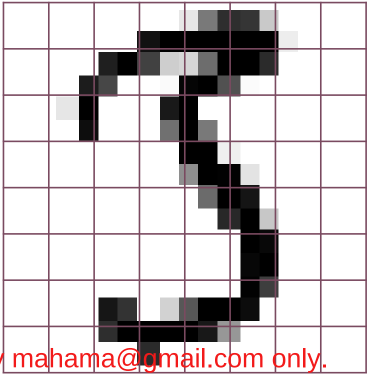
Stochastic neighbor embedding (tSNE)

- 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

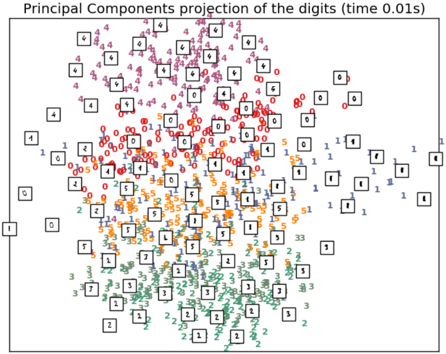
Case study: Digit recognition

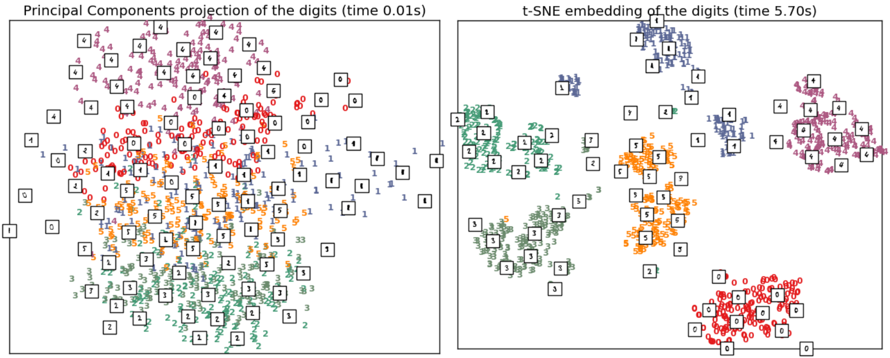
- ~ 1800 hand-written digits (i.e., $n \approx 180$ for each number)
- each (centered) digit was put in a 8×8-grid (i.e., $D = 64$)
- measure grey value in each part of the grid, i.e. 64 grey values
- **Input:** $x_1, \ldots, x_n \in \mathbb{R}^D$, **Output:** $y_1, \ldots, y_n \in \mathbb{R}^d$, where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

A selection from the 64-dimensional digits dataset



Case study: Digit recognition





- 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

`http:scikit-learn.orgstableauto examplesmanifoldplot lle digits.html`

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.