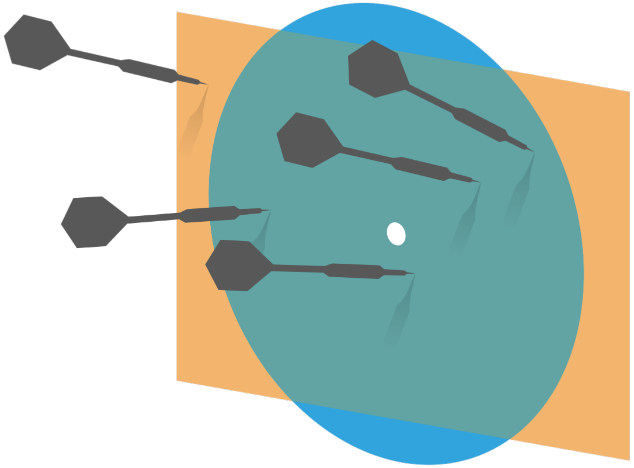
Probability and Statistics for Computer Science



$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$

Covariance is coming back in matrix!

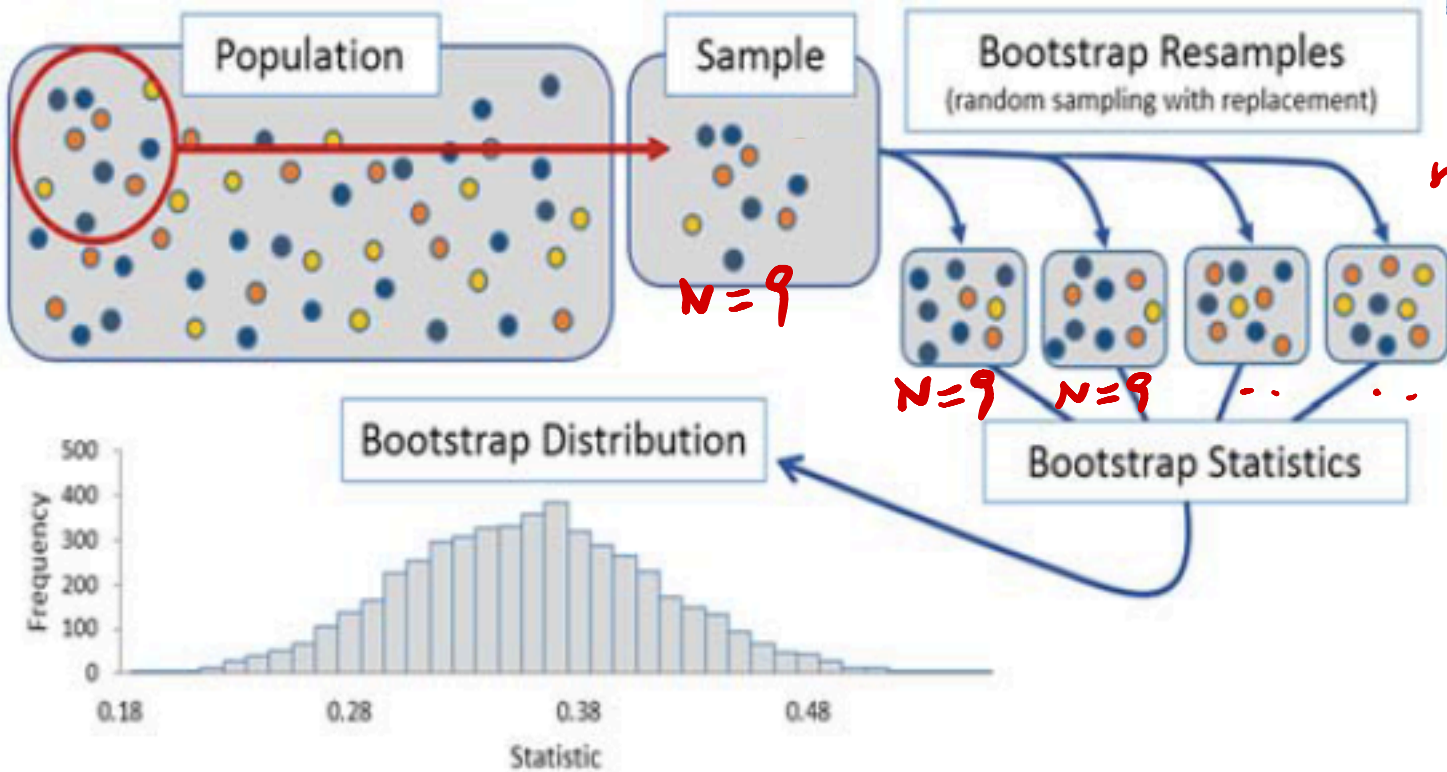
Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 3.25.2021

Bootstrap for confidence interval of other sample statistics

$r = 4000$

r: # of replicates



N^N

100

100

Figure 1. Summary of Bootstrapping Process

Credit: E S. Banjanovic and J. W. Osborne, 2016, PAREonline

Last time

- Maximum likelihood Estimation (MLE II) $L(\theta) = P(D \mid \theta)$

$\hat{\theta} = \operatorname{argmax} L(\theta)$

- Bayesian Inference (MAP)

$\theta \rightarrow RV$

$P(\theta \mid D) \rightsquigarrow$ distri:

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

Objective

- Review of Bayesian inference
- Visualizing high dimensional data & Summarizing data
- The covariance matrix
- Refresh of some linear algebra

Beta distribution

- A distribution is Beta distribution if it has the following pdf:

$$P(\theta) = \begin{cases} K(\alpha, \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ 0 \quad \text{O.W.} \end{cases}$$

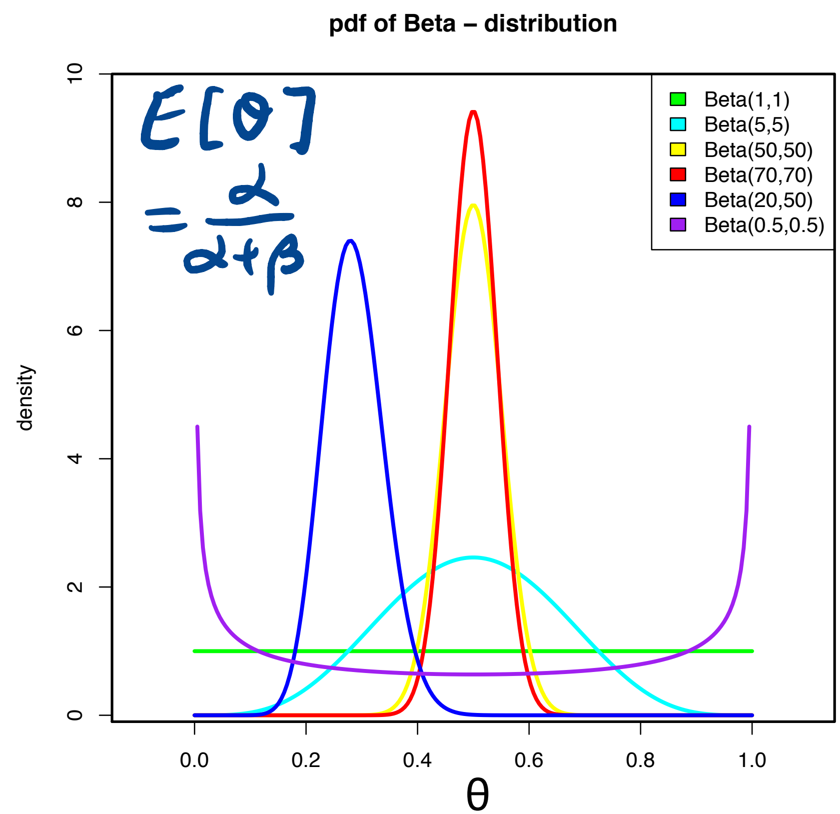
$0 \leq \Theta \leq 1$
$\alpha > 0$, $\beta > 0$

$$K(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

- Is an expressive family of distributions $p(\theta | D)$

- $Beta(\alpha = 1, \beta = 1)$ is uniform

$\hat{\theta}$ that maximize $p(\theta | D)$



Beta distribution as the conjugate prior for Binomial likelihood

- The likelihood is Binomial (*N*, *k*)

$$P(D|\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

- The Beta distribution is used as the prior

$$P(\theta) = K(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$\theta \in [0, 1]$

$\hat{\alpha} = \alpha + k - 1$

- So $P(\theta|D) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$

$\hat{\beta} = \beta + N - k$

- Then the posterior is $Beta(\boxed{\alpha+k,}\boxed{\beta+N-k)}$

$$P(\theta|D) = K(\alpha+k, \beta+N-k)\theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$$

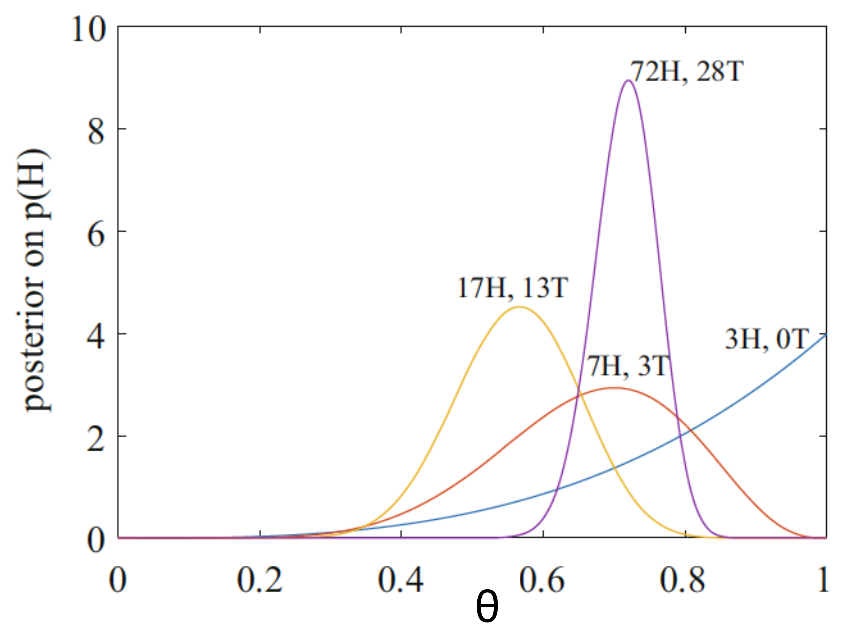
$\theta \in [0, 1]$

The update of Bayesian posterior

- Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed.

- Suppose we start with a uniform prior on the probability θ of heads

N	k	$\hat{\alpha}$	$\hat{\beta}$
		1	1
3	0	1	4
10	7	8	7
30	17	25	20
100	72	97	48



Maximize the Bayesian posterior (MAP)

- The posterior of the previous example is

$$P(\theta|D) = K(\alpha + k, \beta + N - k)\theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$$

- Differentiating and setting to 0 gives the MAP estimate

$$\hat{\theta} = \frac{\alpha - 1 + k}{\alpha + \beta - 2 + N}$$

Table of conjugate prior for different likelihood functions

$L(\theta) = P(D|\theta)$

Likelihood	Conjugate Prior
Bernoulli Geometric Binomial	Beta distri.
Poisson Exponential	Gamma distri.
Normal with known σ^2	Normal distri.

$P(\theta)$

Conjugate prior for other likelihood functions

- If the likelihood is Bernoulli or geometric, the conjugate prior is Beta
- If the likelihood is Poisson or Exponential, the conjugate prior is Gamma
- If the likelihood is normal with known variance, the conjugate prior is normal

Which distri. is the posterior ?

If the likelihood is Geometric and we use the corresponding conjugate prior.

A) Binomial

B) Beta

C) Poisson

D) Bernoulli

E) Normal

What are the dims of A?

$$A \cdot X = B$$

$M \times d \cdot d \times N \quad M \times N$

X is matrix of $d \times N$

B is matrix of $M \times N$

A is matrix of ?

$M \times d$

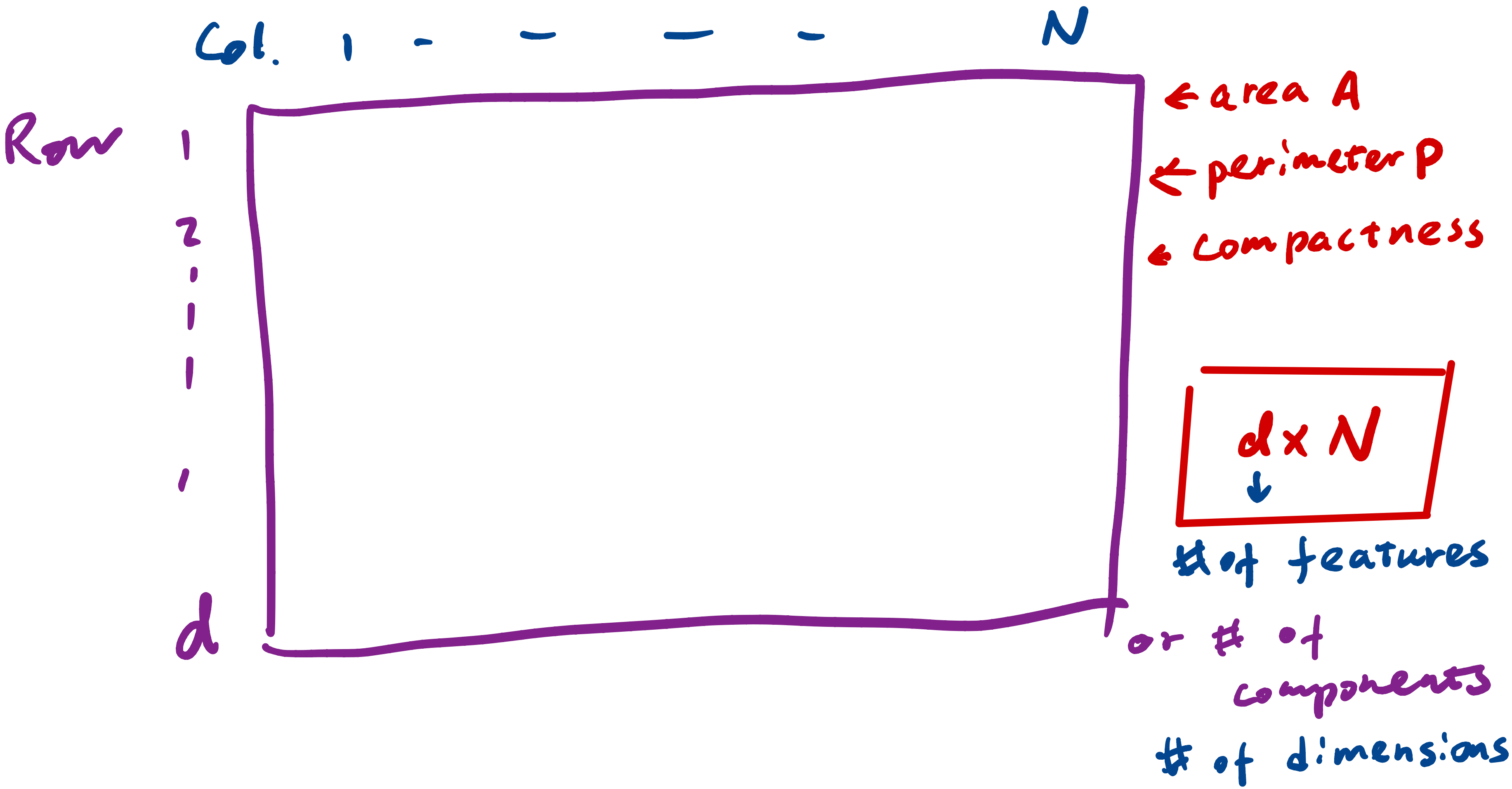
A data set with high dimensions

- Seed data set from the UCI Machine Learning site:

	areaA	perimeterP	compactness	lengthKernel	widthKernel	asymmetry	lengthGroove	Label
1	15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
3	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
4	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
5	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1
6	14.38	14.21	0.8951	5.386	3.312	2.462	4.956	1
7	14.69	14.49	0.8799	5.563	3.259	3.586	5.219	1

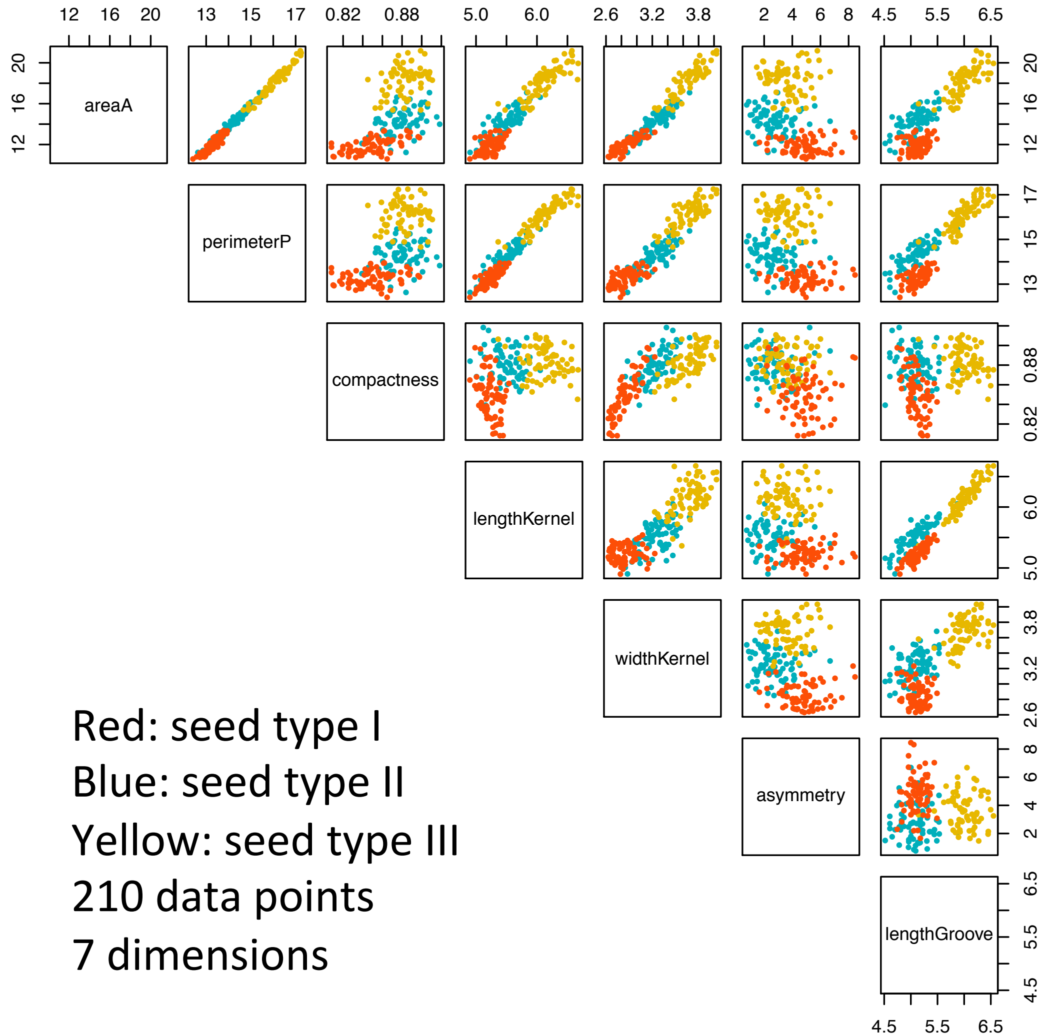
...

Matrix format of a dataset in the textbook



Scatterplot matrix

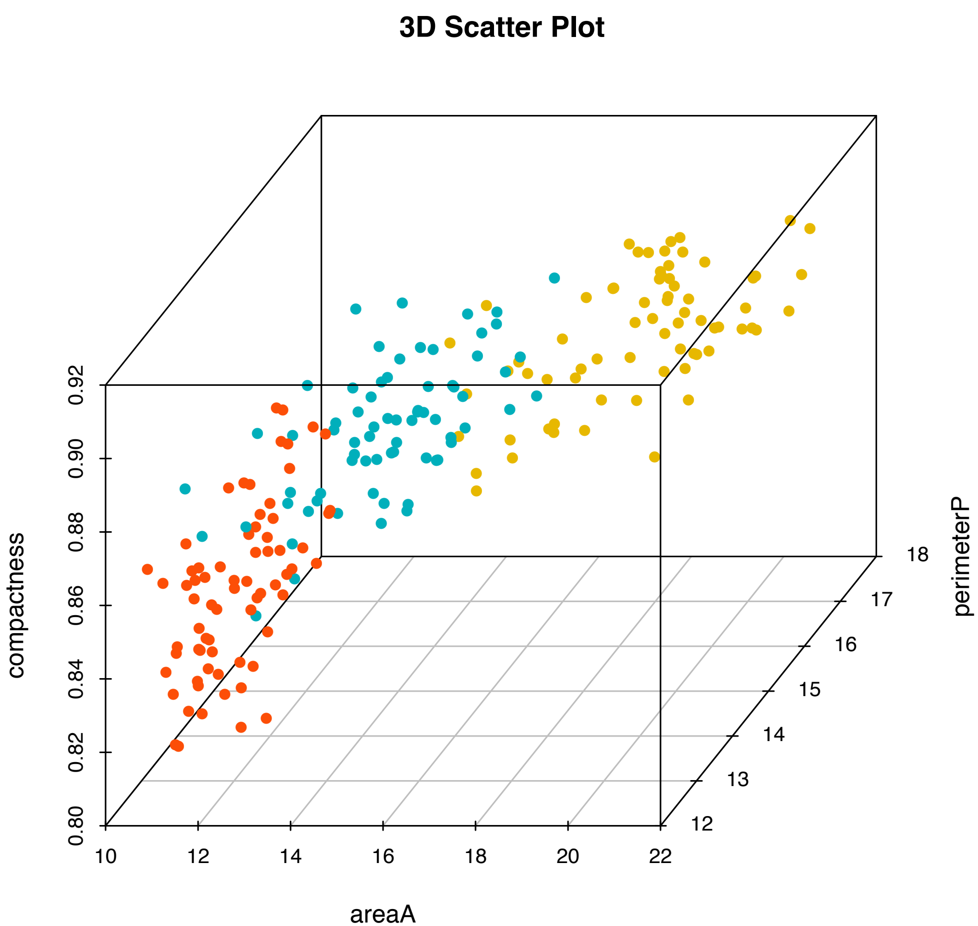
- Visualizing high dimensional data with scatter plot matrix
- Limited to small number of scatter plots



Red: seed type I
Blue: seed type II
Yellow: seed type III
210 data points
7 dimensions

3D scatter plot

- We can also view the data set in 3 dimensions
- But it's still limited in terms of number of dimensions we can see.



Summarizing multidimensional data

- Location and spread parameters of a data set
- Notation
  - Write $\{\mathbf{x}\}$ for a dataset consisting of N data items
  - Each item x_i is a **d**-dimensional vector; column
  - Write jth component of x_i as $x_i^{(j)}$; row
  - Matrix for the data set $\{\mathbf{x}\}$ is **d** by **N** dimension

Mean of a multidimensional data

- We compute the mean of {x} by computing the mean of each component separately and stacking them to a vector

$$\text{mean of jth component} = \frac{\sum_i x_i^{(j)}}{N}$$

- We write the mean of {x} as

$$mean(\{x\}) = \frac{\sum_i x_i}{N}$$

Example of mean of a multidimensional data set

N=3

	1	2	3	Mean
area A feature 1	1	2	3	2
	-1	0	1	0
	3	7	5	5

Mean-Centering a data matrix

Raw

			mean
1	2	3	2
-1	0	1	0
3	7	5	5

Mean Centered

-1	0	1
-1	0	1
-2	2	0

Covariance

- The **covariance** of random variables X and Y is

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- Note that

$$cov(X,X) = E[(X - E[X])^2] = var[X]$$

Correlation coefficient is normalized covariance

- The correlation coefficient is

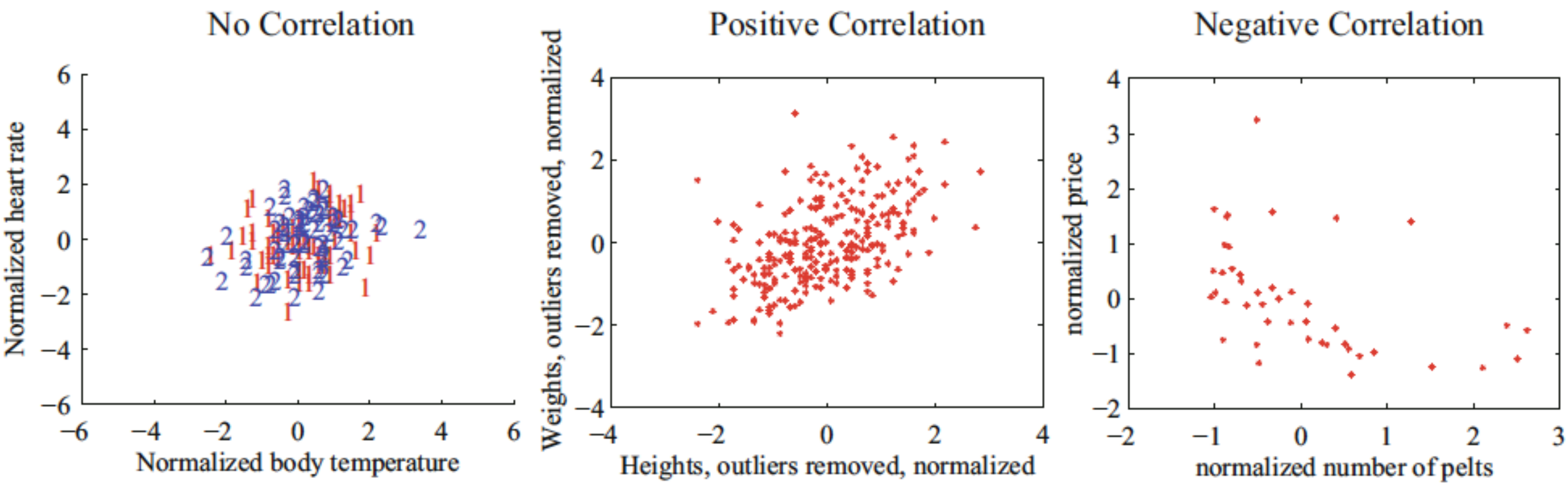
$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

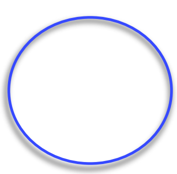
- When X, Y takes on values with equal probability to generate data sets $\{(x,y)\}$, the correlation coefficient will be as seen in Chapter 2.

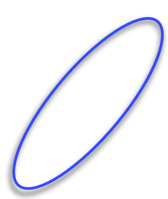
$$corr(\{x,y\}) = \frac{\sum \hat{x}\,\hat{y}}{N}$$

Covariance seen from scatter plots

Zero Covariance	Positive Covariance	Negative Covariance







Credit: Prof.Forsyth

Covariance for a pair of components in a data set

- For the jth and kth components of a data set {x}

$$cov(\{x\};j,k) = \frac{\sum_i (x_i^{(j)} - mean(\{x^{(j)}\}))(x_i^{(k)} - mean(\{x^{(k)}\}))^T}{N}$$

$$corr(\{x,y\}) = \frac{\sum \hat{x}\hat{y}}{N}$$

Covariance of a pair of components

Data set $\{\mathbf{x}\}$ 7×8 $d \times N$

$cov(\{\mathbf{x}\}; 3, 5)$ $N = 8$

	1	2	3	4	5	6	7	8
1	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*

compact-ness

width-kernel

Take each row (component) of a pair and subtract it by the row mean, then do the inner product of the two resulting rows and divide by the number of columns

Covariance of a pair of components

Data set $\{\mathbf{x}\}$ 7×8

$cov(\{\mathbf{x}\}; 3, 5)$

	1	2	3	4	5	6	7	8
1	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*

7×7

How many pairs of rows are there for which we can compute the covariance?

A) 49
B) 64
C) 56

Covariance matrix

Data set $\{\mathbf{x}\}$ 7×8

$cov(\{\mathbf{x}\}; 3, 5)$

	1	2	3	4	5	6	7	8
1	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*

Covmat($\{\mathbf{x}\}$) 7×7

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

Properties of Covariance matrix

$$cov(\{x\}; j, j) = var(\{x^{(j)}\})$$

Covmat($\{\mathbf{X}\}$) 7×7

- The **diagonal** elements of the covariance matrix are just **variances** of each jth components

- The off diagonals are covariance between different components

	1	2	3	4	5	6	7
1	* σ_1^2	*	*	*	*	*	*
2	*	* σ_2^2	*	*	v_0	*	*
3	*	*	* σ_3^2	*	*	*	*
4	*	*	*	* σ_4^2	*	*	*
5	*	*	*	*	* σ_5^2	*	*
6	*	*	*	*	*	* σ_6^2	*
7	*	*	*	*	*	*	* σ_7^2

$$corr = \frac{cov}{\sigma_x \sigma_y} = \frac{v_0}{\sigma_2 \sigma_5} = \frac{v_0}{\sqrt{cov(2,2)}\sqrt{cov(5,5)}},$$

Properties of Covariance matrix

$$cov(\{x\}; j, k) = cov(\{x\}; k, j)$$

Covmat($\{\mathbf{x}\}$) 7×7

- The covariance matrix is **symmetric**!
- And it's **positive semi-definite**, that is all $\lambda_i \geq 0$ → eigenvalues
- Covariance matrix is diagonalizable

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

Properties of Covariance matrix

- If we define X_c as the mean centered matrix for dataset {x} $\sim d \times N$

$$Covmat(\{x\}) = \frac{X_c X_c^T}{N}$$

- The covariance matrix is a d×d matrix

$\text{Covmat}(\{\mathbf{x}\})$ 7×7

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

d =7

Example: covariance matrix of a data set

(I)

$$A_0 = \begin{array}{cc} \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix} & \begin{array}{c} x^{(1)} \\ x^{(2)} \end{array} \end{array}$$

What are the dimensions of the covariance matrix of this data?

A) 2 by 2
B) 5 by 5
C) 5 by 2
D) 2 by 5

$$\begin{bmatrix} \mathrm{cov}(1,1) & \mathrm{cov}(1,2) \\ \mathrm{cov}(2,1) & \mathrm{cov}(2,2) \end{bmatrix}$$

Example: covariance matrix of a data set

Mean centering

(I)

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

Mean

$$\begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 2 & 1 & 0 & -1 & -2 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

Example: covariance matrix of a data set

(I) Mean centering

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 2 & 1 & 0 & -1 & -2 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

(II) $A_2 = A_1 A_1^T$

Inner product of each pairs:

A_2 [1,1] = 10

A_2 [2,2] = 4

A_2 [1,2] = 0

$$\mathrm{cov}(\{x\}) = \frac{x_c \, x_c^T}{N}$$

Example: covariance matrix of a data set

(I)

Mean centering

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 2 & 1 & 0 & -1 & -2 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

(II) $A_2 = A_1 A_1^T$

Inner product of each pairs:

A_2 [1,1] = 10

A_2 [2,2] = 4

A_2 [1,2] = 0

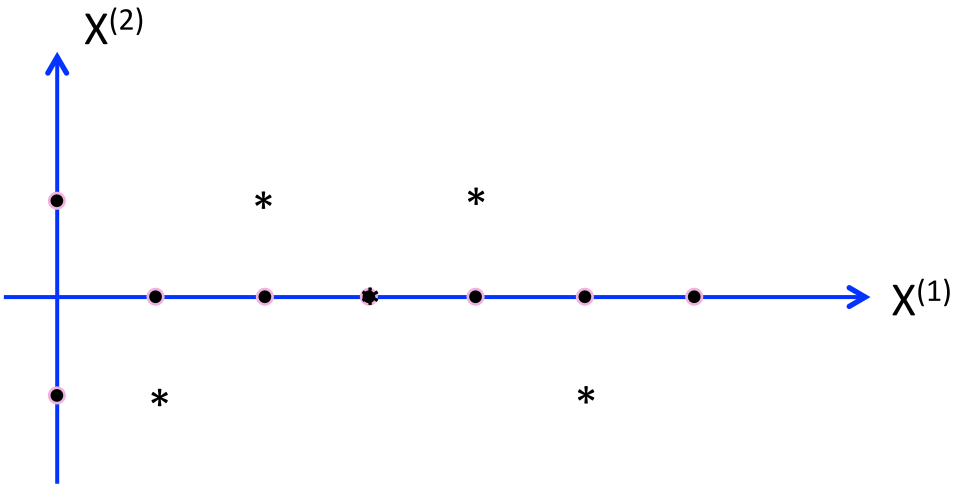
(III)

Divide the matrix with N – the number of items

$$\text{Covmat}(\{\mathbf{x}\}) = \frac{1}{N} A_2 = \frac{1}{5} \begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} = \boxed{\begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix}}$$

What do the data look like when Covmat({x}) is diagonal?

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$



$$\text{Covmat}(\{\mathbf{x}\}) = \frac{1}{N} A_2 = \frac{1}{5}\begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} = \boxed{\begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix}}$$

Translation properties of mean and covariance matrix

- Translating the data set translates the mean

$$mean(\{x\} + c) = mean(\{x\}) + c$$

- Translating the data set leaves the covariance matrix unchanged

$$Covmat(\{x\} + c) = Covmat(\{x\})$$

Translation properties of covariance matrix

- Proof:

$$\mathrm{covmat}(\{x\}) = \frac{X_c X_c^T}{N}$$

if we translate $\{x\}$, X_c doesn't change.

because

$$x + c - \mathrm{mean}(\{x\} + c)$$
$$= x - \mathrm{mean}(\{x\}) = X_c$$

Linear transformation properties of mean and covariance matrix

- Linearly transforming the data set linearly transforms the mean

$$mean(\{A\mathbf{x}\}) = A\ mean(\{\mathbf{x}\})$$

- Linearly transforming the data set linearly changes the covariance matrix quadratically

$$Covmat(\{A\mathbf{x}\}) = A\ Covmat(\{\mathbf{x}\})A^T$$

$var(k\{x\}) = k^2 var(\{x\})$

Proof of linear transformation of covariance matrix

$$\text{covmat}(\{x\}) = \frac{X_c X_c^T}{N}$$

$$\text{covmat}(\{Ax\}) = \frac{(AX)_c (AX)_c^T}{N}$$

$$= \frac{A X_c \, (A X_c)^T}{N}$$

$$= \frac{A X_c \cdot X_c^T A^T}{N}$$

$$= A \cdot \text{Covmat}(\{x\}) A^T$$

$$(M_a \cdot M_b)^T = M_b^T \cdot M_a^T$$

* suppose $X = X_c$

data is centered.

(1) $\therefore AX = AX_c$

(2) if X is centered

AX is centered

$\therefore (AX_c)_c = AX_c$

Assignments

- Read Chapter 10 of the textbook
- Finish Week9 module including the quiz.
- Next time: PCA

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See You!