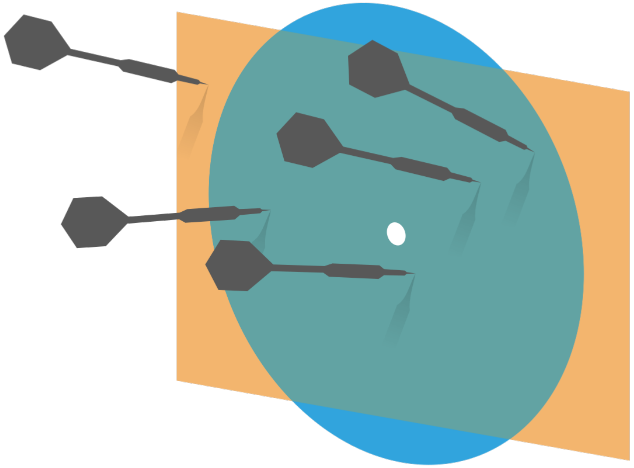
Probability and Statistics for Computer Science



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

Covariance is coming back in matrix!

Credit: wikipedia

Last time

- Review of Maximum likelihood Estimation (MLE)
- Bayesian Inference (MAP)

Objective

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

Beta distribution

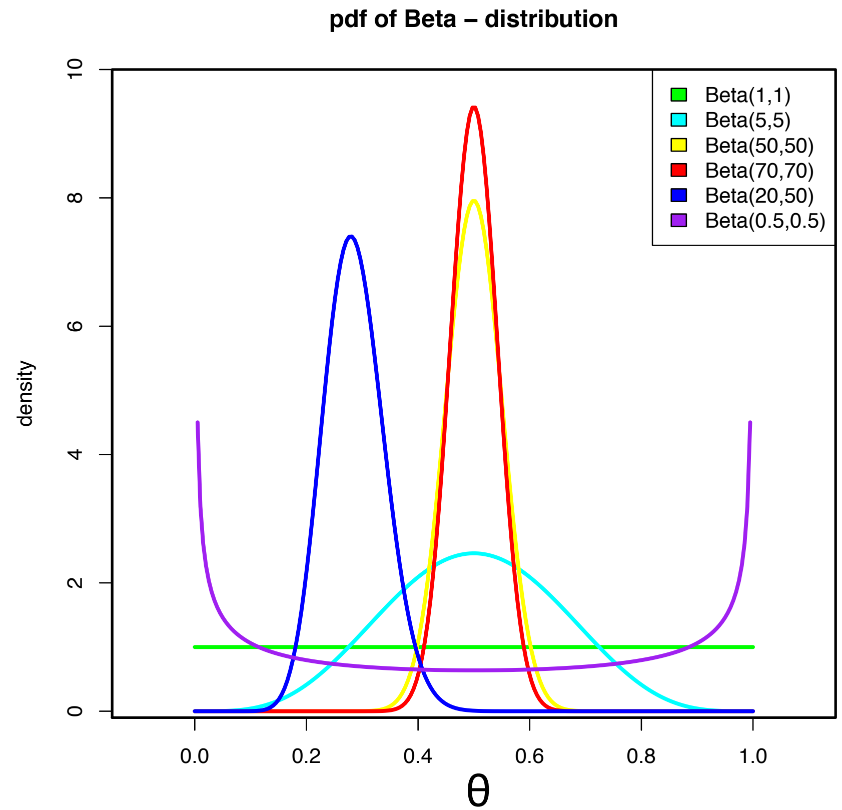
- A distribution is Beta distribution if it has the following pdf:

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

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

- Is an expressive family of distributions

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



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

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

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

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

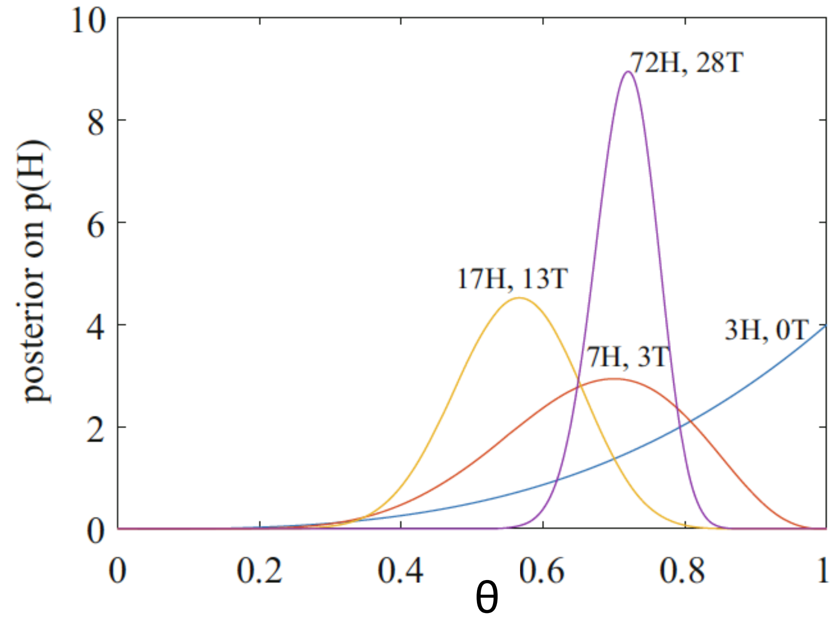
$$P(\theta|D) = K(\alpha + k, \beta + N - k)\theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$$

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

θ

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

θ

A data set with high dimensions

- Seed data set from the UCI Machine Learning site:

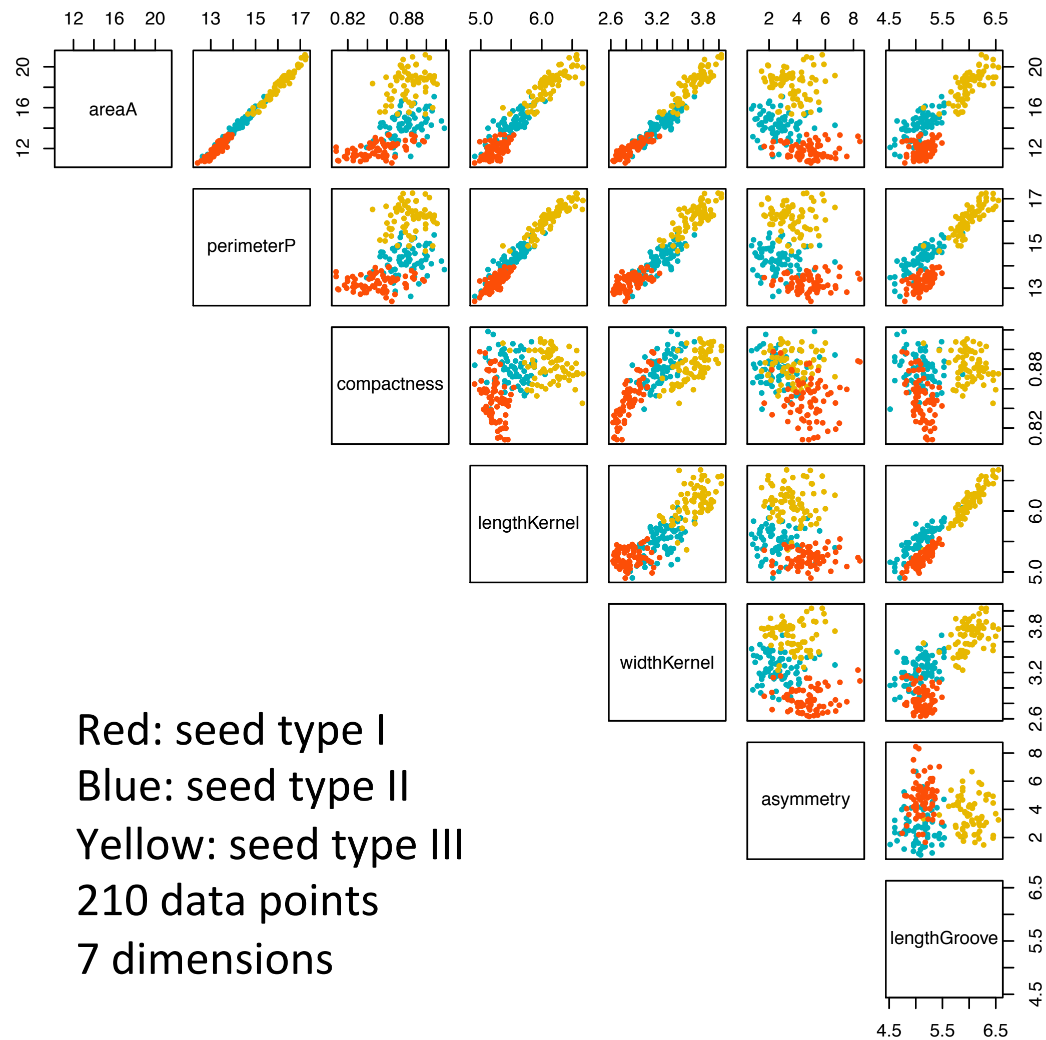
	areaA	perimeterP	compactness	lengthKernel	widthKernel	asymmetry	lengthGroove	Label
1	15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
3	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
4	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
5	16.14	14.99	0.9034	5.658	3.562	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

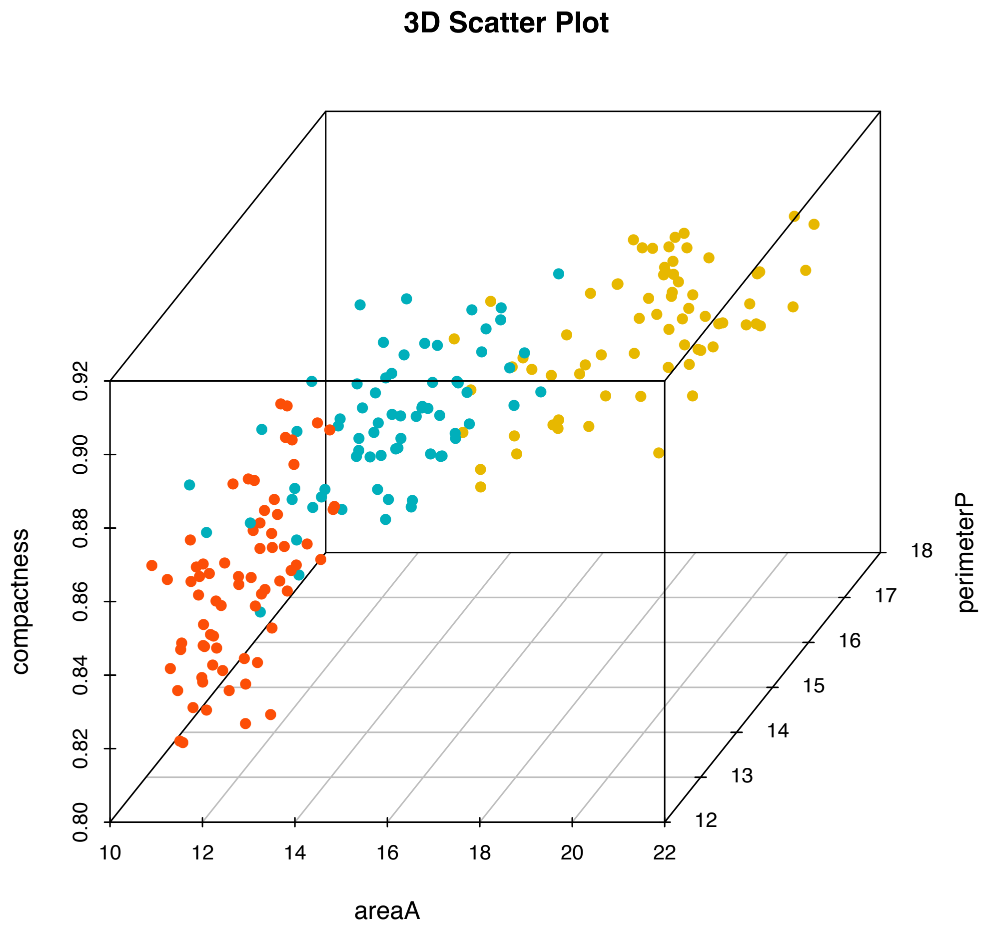
- Visualizing high dimensional data with scatter plot matrix
- Limited to small number of scatter plots



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 {**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 {**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

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

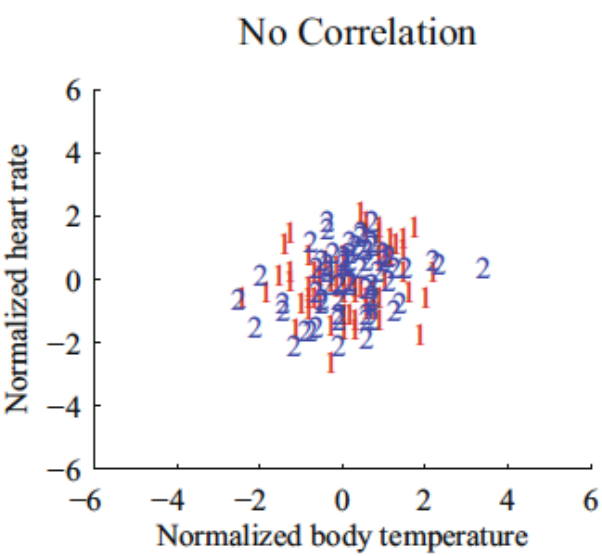
- The correlation coefficient is

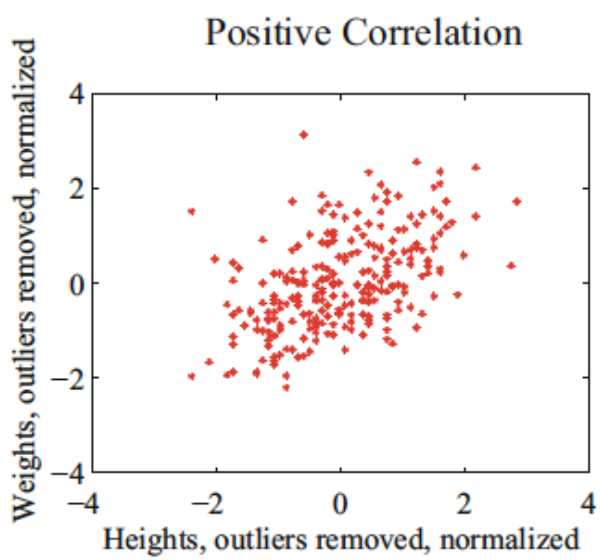
$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

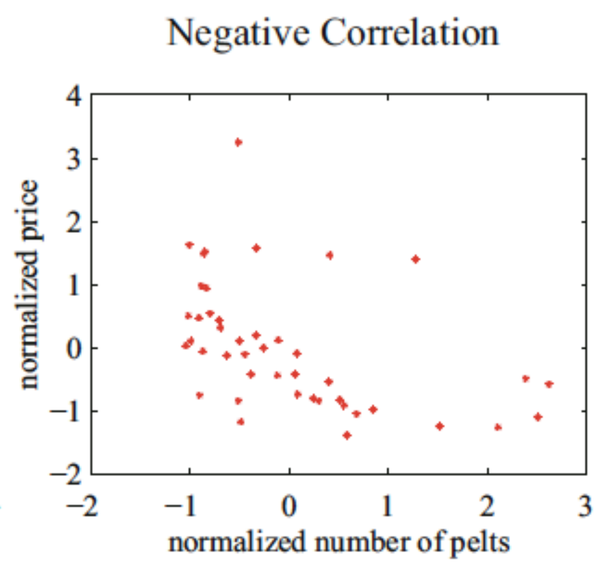
- When X, Y takes on values with equal probability to generate data sets $\{(x,y)\}$, the correlation coefficient will be as seen in Chapter 2.

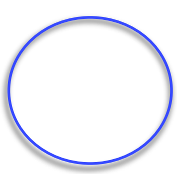
Covariance seen from scatter plots

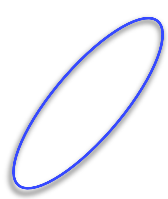
Zero	Positive	Negative
Covariance	Covariance	Covariance











Credit: Prof.Forsyth

Covariance for a pair of components in a data set

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

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

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	*	*	*	*	*	*	*	*

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	*	*	*	*	*	*	*	*

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

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

- The diagonal elements of the covariance matrix are just variances of each jth components
- The off diagonals are covariance between different components

Properties of Covariance matrix

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

Covmat($\{\mathbf{x}\}$) 7×7

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

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

Properties of Covariance matrix

- If we define $\mathbf{x}_c$ as the mean centered matrix for dataset {x}

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

- 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)

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix} \begin{matrix} X^{(1)} \\ X^{(2)} \end{matrix}$$

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

Example: covariance matrix of a data set

Mean centering

(I)

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

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

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

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

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

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

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

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

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

Dimension Reduction

- In stead of showing more dimensions through visualization, it's a good idea to do dimension reduction in order to see the major features of the data set.
- For example, principal component analysis help find the major components of the data set.
- PCA is essentially about finding eigenvectors of covariance matrix

Refresh of some linear algebra

Why linear algebra?

- We are now into part **IV** of the course. The contents will be basic machine learning techniques.
- Linear algebra is essential for a lot of machine Learning methods!

Eigenvalues and eigenvectors review

- If A is an **n×n** square matrix, an eigenvalue λ and its corresponding eigenvector v (of dimension n×1) satisfy $Av = \lambda v$.

- To solve for λ, we solve the characteristic equation

  $$|A - \lambda I| = 0$$

- Given a value of λ, we solve v by solving

  $$(A - \lambda I)\, v = 0$$

- Note if v is an eigenvector, then so is any multiple kv.

Eigenvalues and eigenvectors example

- Find the eigenvalues and eigenvectors

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

Eigenvalues and eigenvectors example

- Find the eigenvectors

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

Eigenvalues and eigenvectors example (2)

- Find the eigenvalues and eigenvectors of

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

Eigenvalues and eigenvectors example

- Find the eigenvectors of

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

Diagonalization of a symmetric matrix

- If A is an n×n symmetric square matrix, the eigenvalues are real.
- If the eigenvalues are also distinct, their eigenvectors are orthogonal
- We can then scale the eigenvectors to unit length, and place them into an orthogonal matrix $U = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots.\ \mathbf{u}_n]$
- We can write the diagonal matrix $\Lambda = U^T A U$ such that the diagonal entries of Λ are $\lambda_1, \lambda_2 \ldots \lambda_n$ in that order.

Diagonalization example

- For

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

Q. Are these two vectors orthogonal?

$V_1 = [3\ 6]$, $V_2 = [-2\ 1]$

A. Yes

B. No

Q. Is this true?

When two zero-mean vectors of data are orthogonal, they are uncorrelated

A. Yes

B. No

Assignments

- Read Chapter 10 of the textbook
- 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!