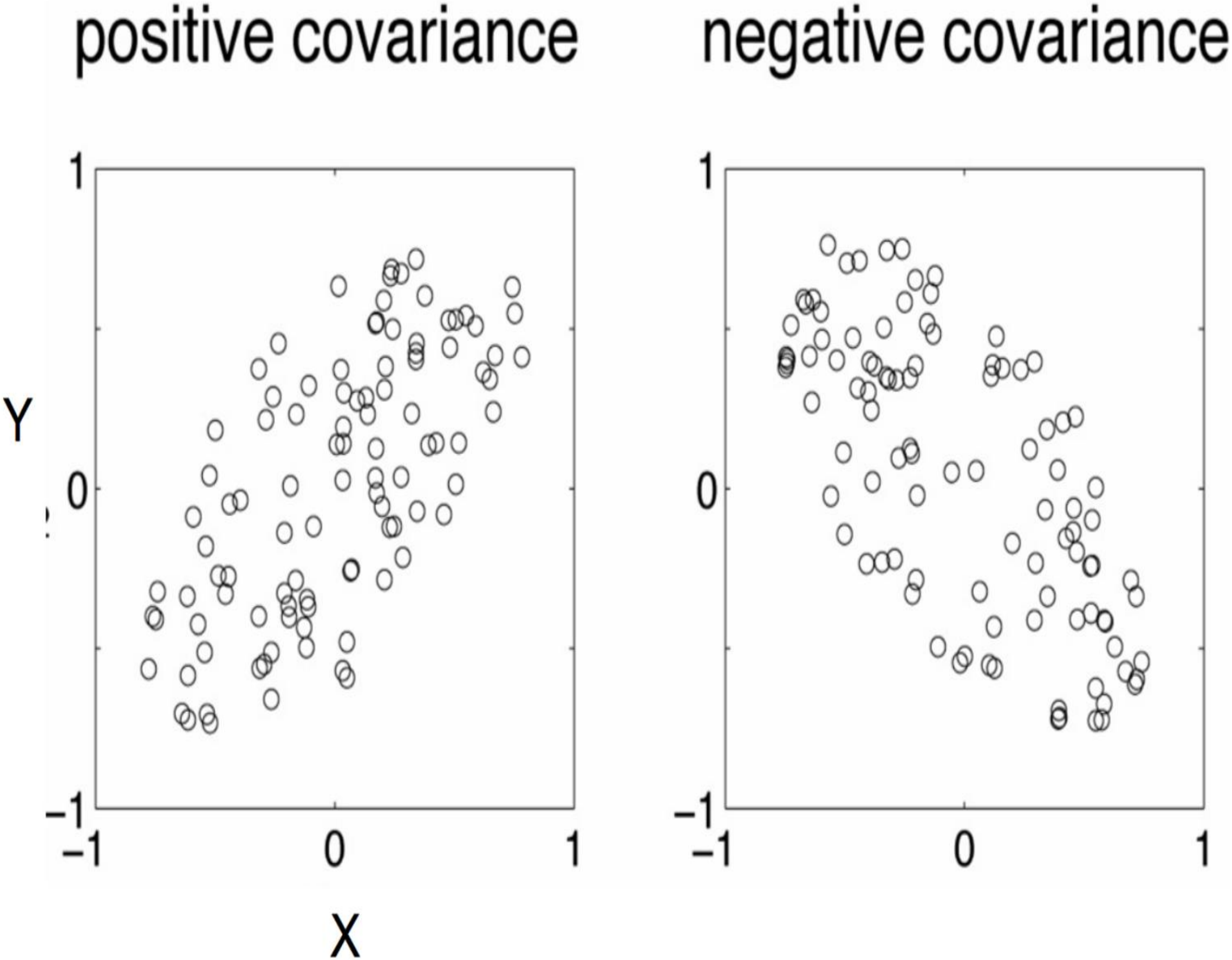
Covariance

Teacher:
prof. G.Shadmanova

Plan:

- **Covariance is measured between two dimensions**
- **Covariance sees if there is a relation between two dimensions**
- **Covariance between one dimension is the variance**



Positive: Both dimensions increase or decrease together

Negative: While one increase the other decrease

Covariance

- Used to find relationships between dimensions in high dimensional data sets

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The Sample mean

Covariance

Variance(x)= $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

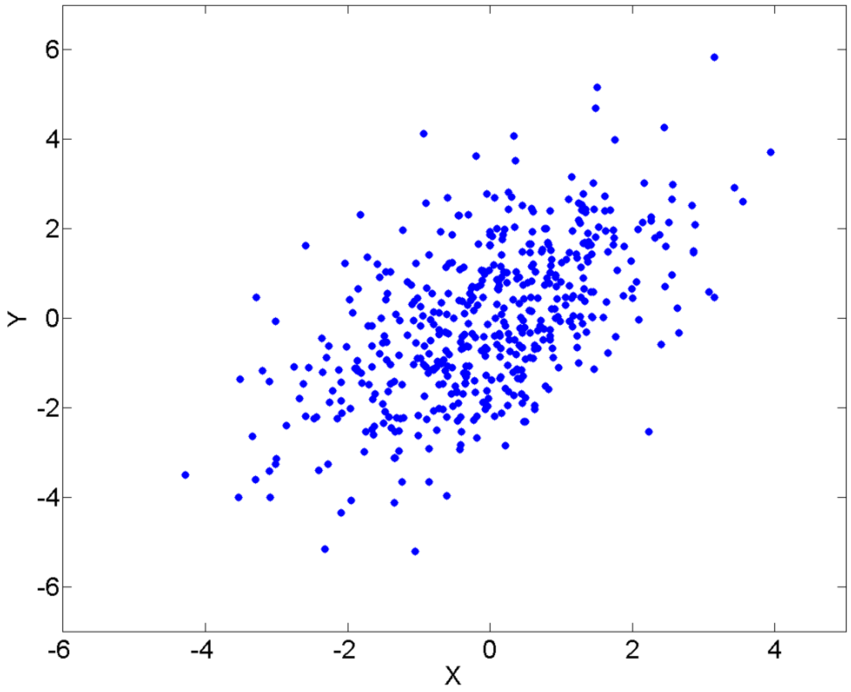
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- $\text{Covariance}(x, x) = \text{var}(x)$
- $\text{Covariance}(x, y) = \text{Covariance}(y, x)$

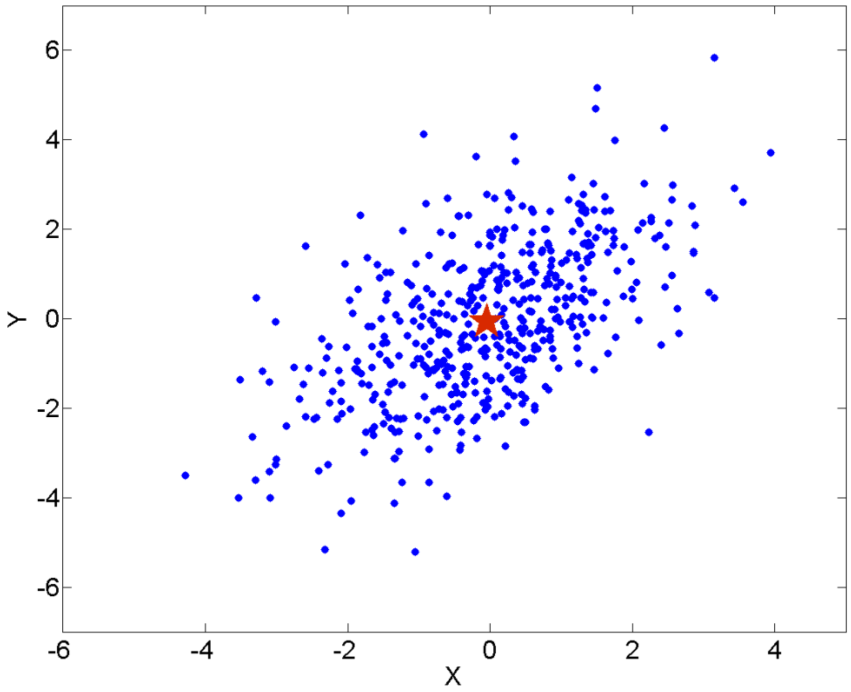
Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



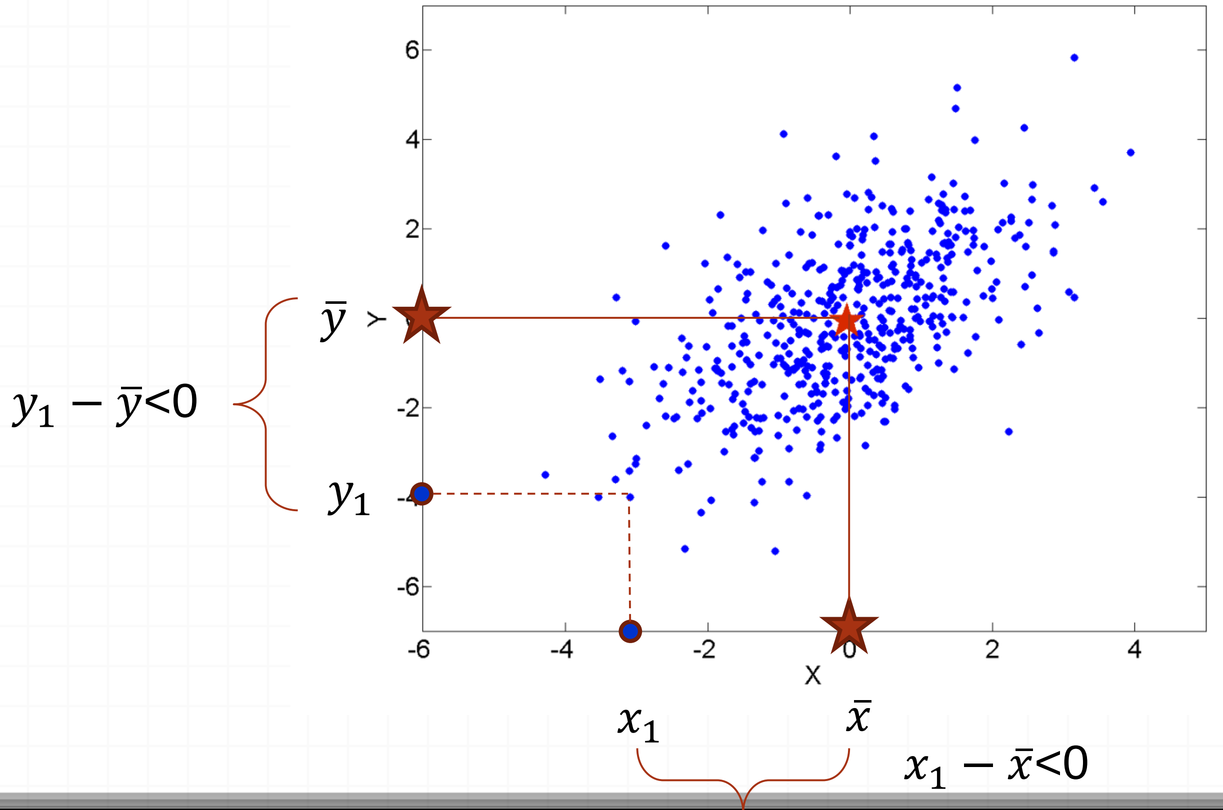
Covariance

$$\text{Covariance}(\text{x}, \text{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



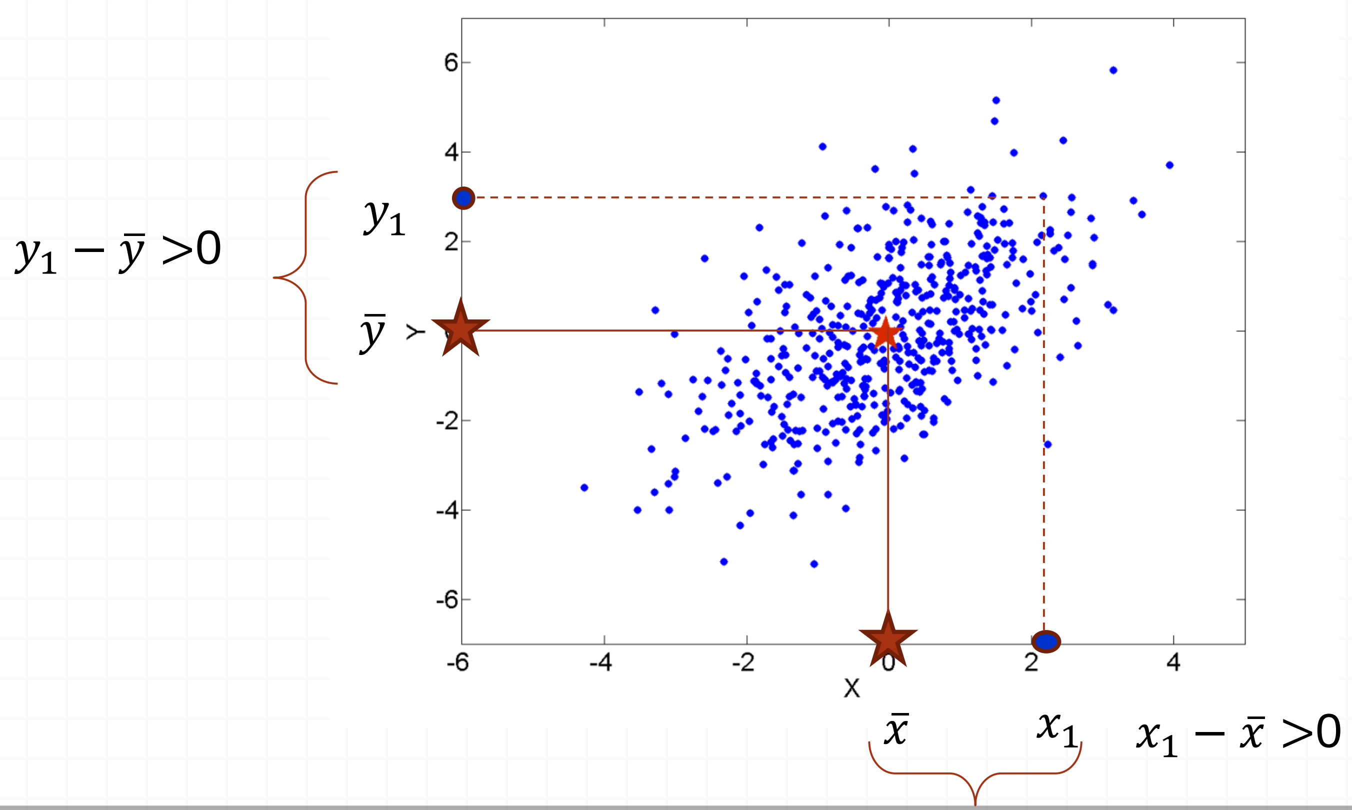
Covariance

$$\text{Covariance}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



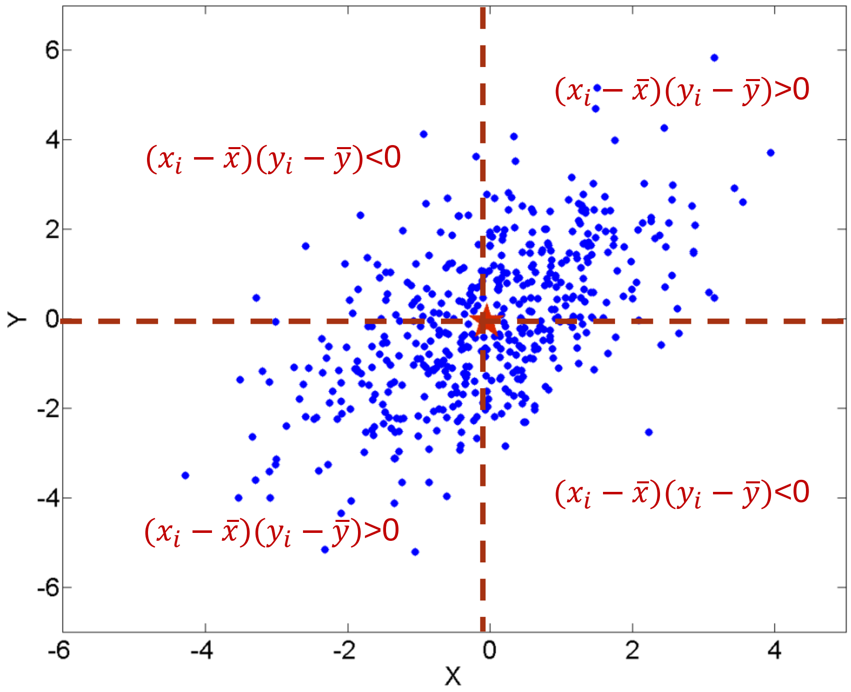
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



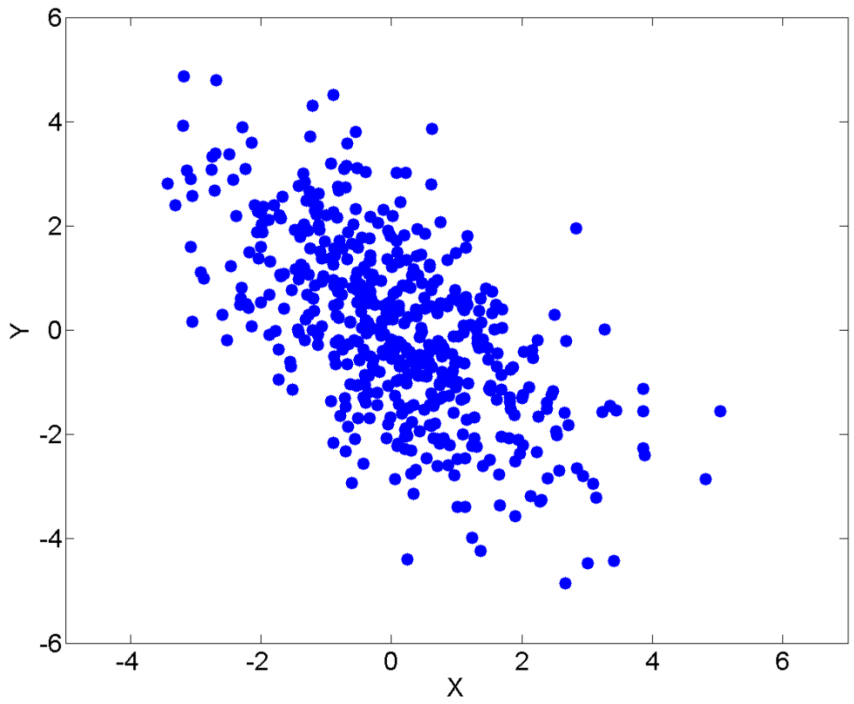
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



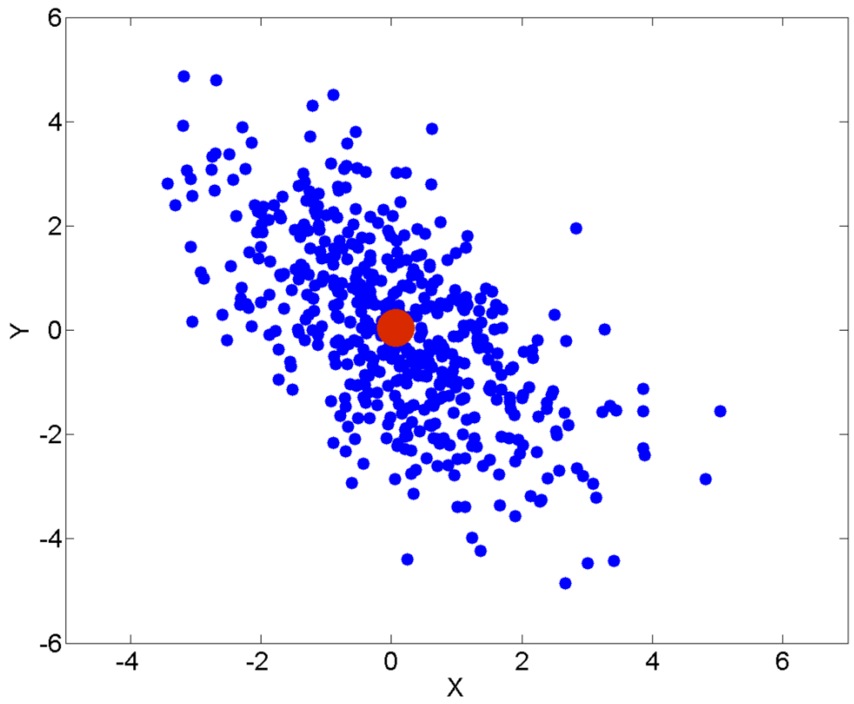
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



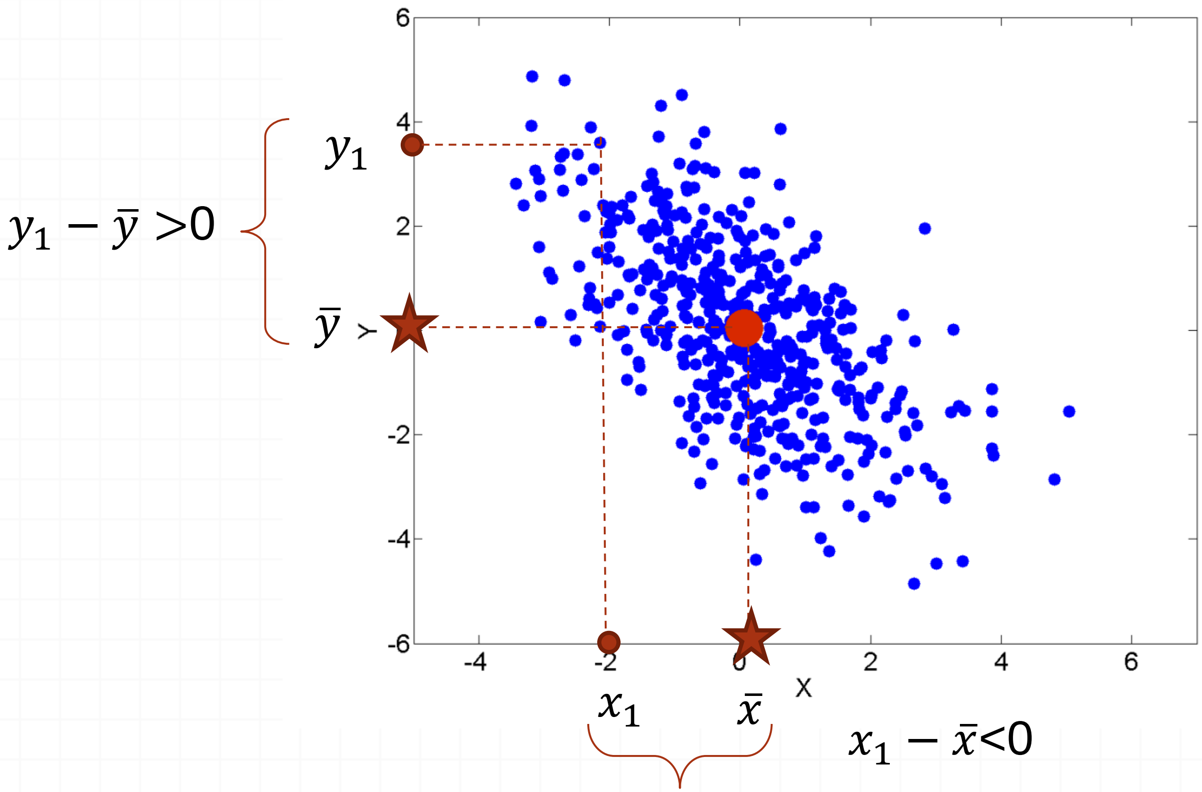
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



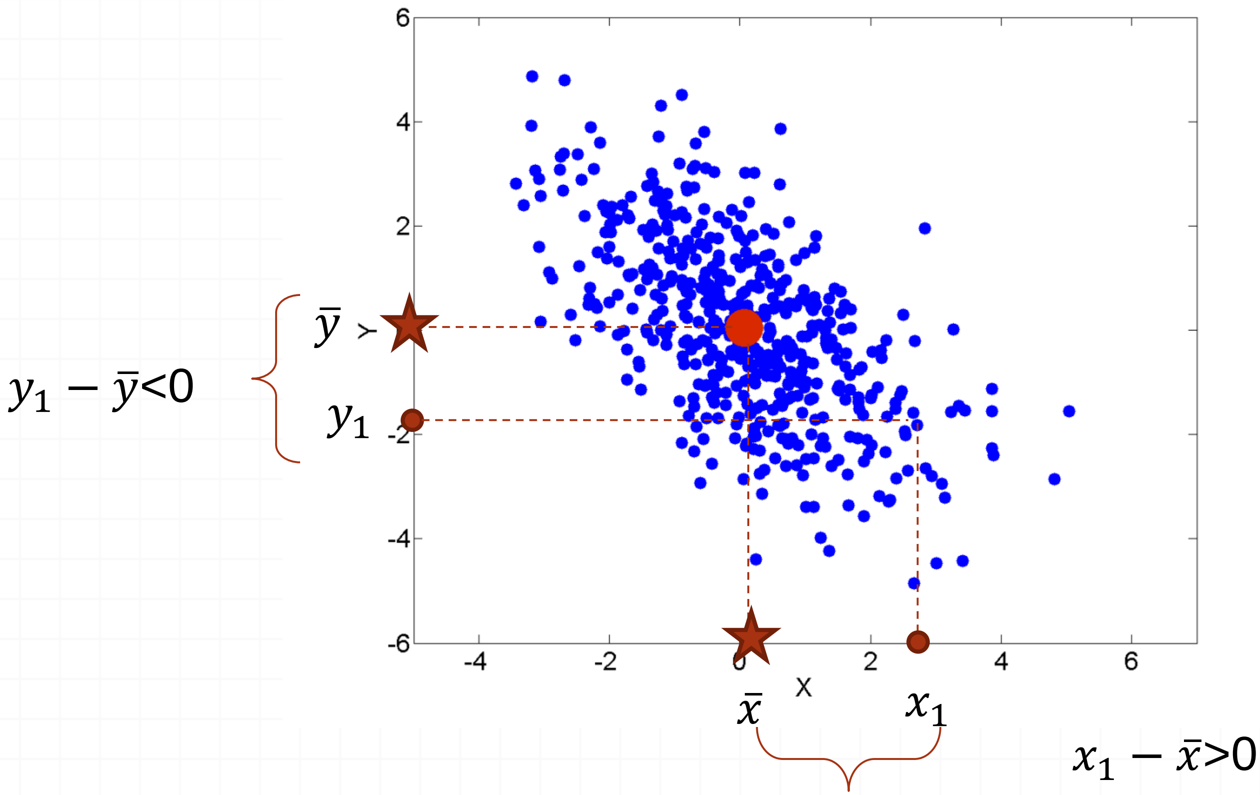
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

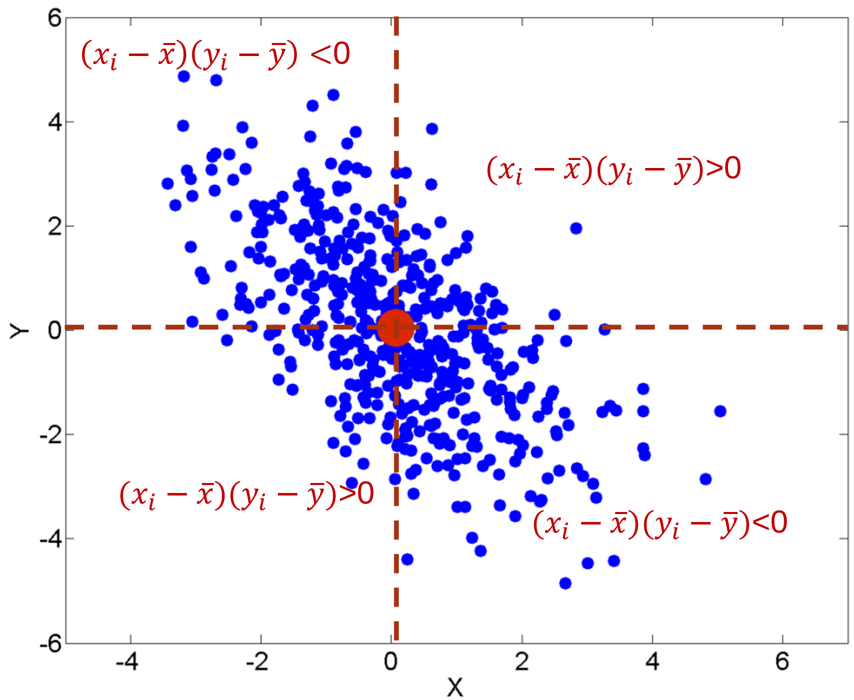
$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

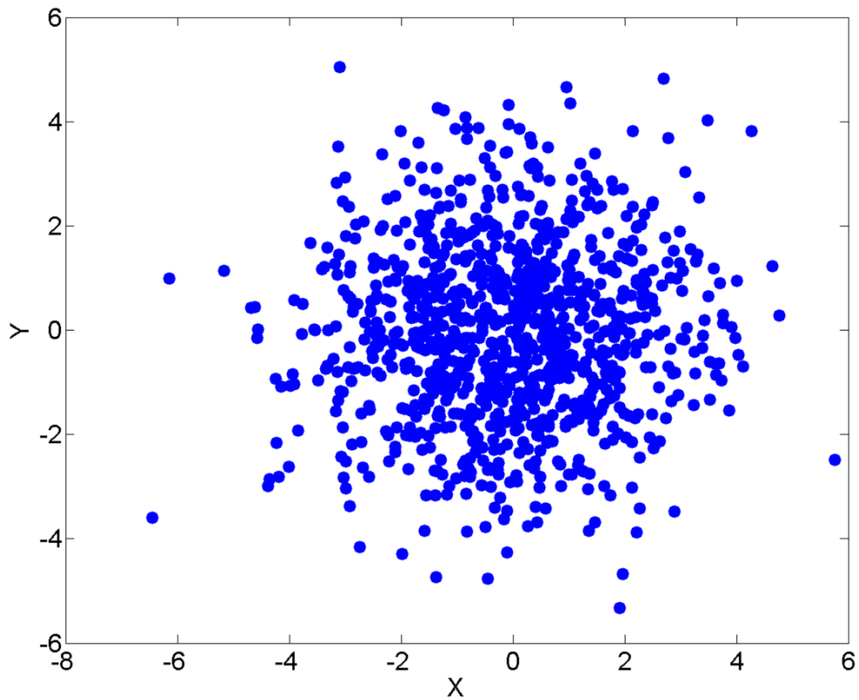
$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Negative Relation



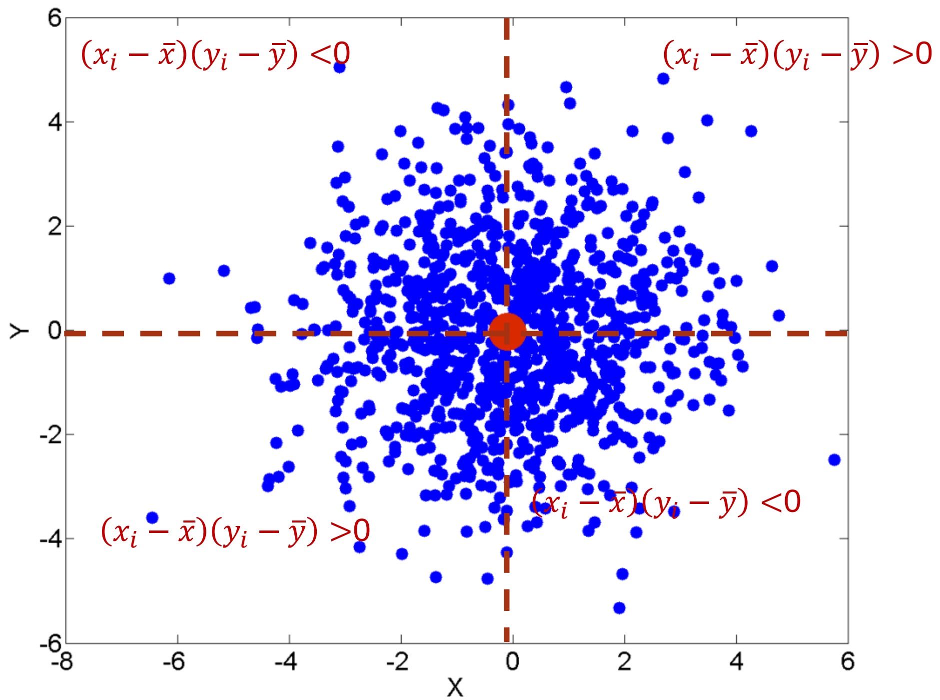
Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

ID	Education	Income				
1	8	15000				
2	12	23000				
3	11	16000				
4	18	65000				
5	16	39000				
6	8	11000				
7	14	28000				
8	12	38000				
9	16	32000				

Understanding the covariance

- Covariance should represent the degree of association
- To the extent that differences from the mean are in the same direction, covariance should be larger
- The stronger the association is, the larger the covariance should be

Application

ID	Education	Income	Education: Comparison to the mean	Income: Comparison to the mean
1	8	15000	-4.4	-14300
2	12	23000	-0.4	-6300
3	11	16000	-1.4	-13300
4	18	65000	5.6	35700
5	16	39000	3.6	9700
6	8	11000	-4.4	-18300
7	14	28000	1.6	-1300
8	12	38000	-0.4	8700
9	16	32000	3.6	2700
10	9	26000	-3.4	-3300
Mean	12.4	29300		

Covariance

Two variables, x and y.
When one is above the mean,
is the other one also above the mean?
x & y are the variables that we are interested in
x_i & y_i are the values of the variables for one individual
n is the sample size

$$s^2_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

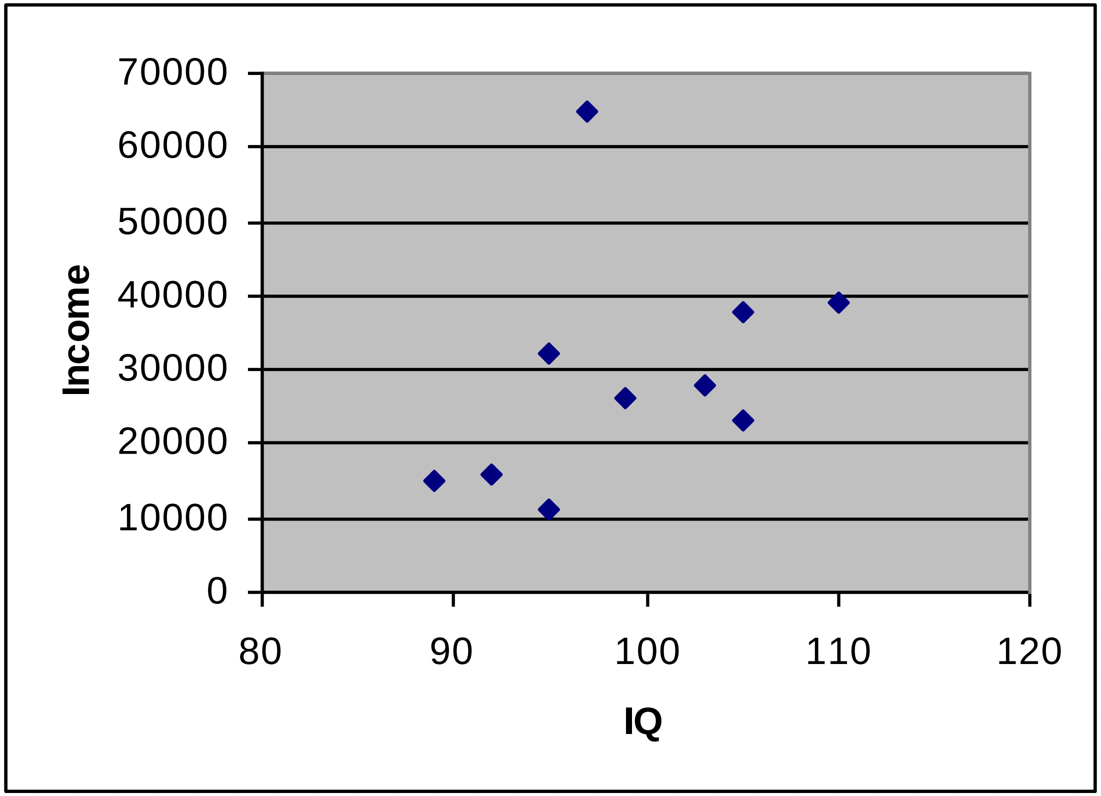
Example

ID	Education	Income	Education: Comparison to the mean	Income: Comparison to the mean	Cross product	
1	8	15000	-4.4	-14300	62920	
2	12	23000	-0.4	-6300	2520	
3	11	16000	-1.4	-13300	18620	
4	18	65000	5.6	35700	199920	
5	16	39000	3.6	9700	34920	
6	8	11000	-4.4	-18300	80520	
7	14	28000	1.6	-1300	-2080	
8	12	38000	-0.4	8700	-3480	
9	16	32000	3.6	2700	9720	
10	9	26000	-3.4	-3300	11220	
Mean	12.4	29300		Sum	414800	
				Covariance	46088.89	

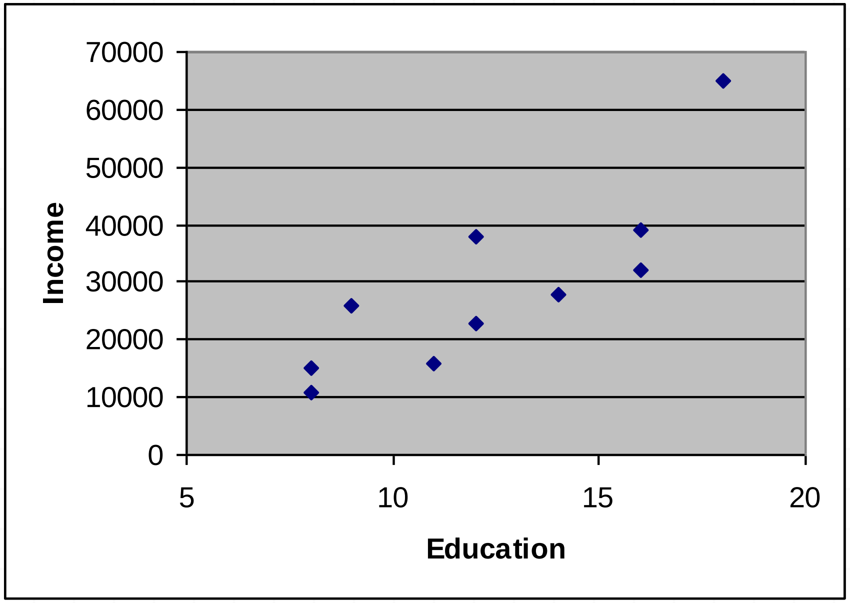
How about IQ and Income?

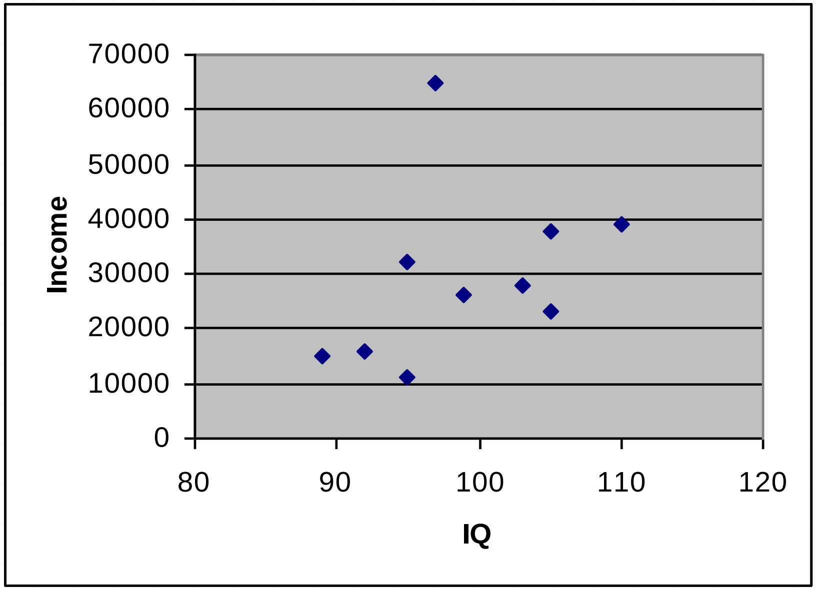
ID	IQ	Income				
1	89	15000				
2	105	23000				
3	92	16000				
4	97	65000				
5	110	39000				
6	95	11000				
7	103	28000				
8	105	38000				
9	95	32000				
10	99	26000				

Inspect the association



Which association is stronger?





Compute covariance (again)

ID	IQ	Income	IQ: Comparison to the mean	Income: Comparison to the mean	Cross product
1	89	15000	-9.4	-14300	134420
2	83	23000	-15.4	-6300	97020
3	92	16000	-6.4	-13300	85120
4	101	65000	2.6	35700	92820
5	110	39000	11.6	9700	112520
6	95	11000	-3.4	-18300	62220
7	103	28000	4.6	-1300	-5980
8	105	38000	6.6	8700	57420
9	107	32000	8.6	2700	23220
10	99	26000	0.6	-3300	-1980
Mean	98.4	29300		Sum	656800
				Covariance	72977.78

Comparison

- Covariance between education and income is: 46,088.89
- Covariance between IQ and income is: 72,977.78

$$s_{x,y}^{2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Comparison

- Unit of measurement of covariance between education and income is:
  - Years of education dollars
- Unit of measurement of covariance between IQ and income is:
- IQ points dollars

- QUESTION: When we need something in universal units, how do we measure it? (think t- test, think effect size)

How do we compare

- Try to express both quantities in terms of a single unit of analysis
- →CORRELATION
  - Divide covariance by the standard deviation of each variable

Education and Income

ID	Education	Income	Education: Comparison to the mean	Income: Comparison to the mean	Cross product
1	8	15000	-4.4	-14300	62920
2	12	23000	-0.4	-6300	2520
3	11	16000	-1.4	-13300	18620
4	18	65000	5.6	35700	199920
5	16	39000	3.6	9700	34920
6	8	11000	-4.4	-18300	80520
7	14	28000	1.6	-1300	-2080
8	12	38000	-0.4	8700	-3480
9	16	32000	3.6	2700	9720
10	9	26000	-3.4	-3300	11220
Mean	12.4	29300		Sum	414800
SD	3.533962208	15705.979		Covariance	46088.89
				Correlation	0.830366

IQ and Income

ID	IQ	Income	IQ: Comparison to the mean	Income: Comparison to the mean	Cross product
1	89	15000	-9.4	-14300	134420
2	83	23000	-15.4	-6300	97020
3	92	16000	-6.4	-13300	85120
4	101	65000	2.6	35700	92820
5	110	39000	11.6	9700	112520
6	95	11000	-3.4	-18300	62220
7	103	28000	4.6	-1300	-5980
8	105	38000	6.6	8700	57420
9	107	32000	8.6	2700	23220
10	99	26000	0.6	-3300	-1980
Mean	98.4	29300		Sum	656800
SD	8.553102101	15705.979		Covariance	72977.78
				Correlation	0.543253

Covariance Matrix

$$Cov\ (\Sigma) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_m) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_m, x_1) & cov(x_m, x_2) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

$$Cov\ (\Sigma) = \frac{1}{n}(X - \bar{X})(X - \bar{X})^T; where\ X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Covariance Matrix

$$Cov\,(\Sigma) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_m) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_m, x_1) & cov(x_m, x_2) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

- Diagonal elements are variances, i.e. $\mathrm{Cov}(x, x) = var(x)$.
- Covariance Matrix is symmetric.
- It is a positive semi-definite matrix.

Covariance Matrix

- Covariance is a real symmetric positive semi-definite matrix.
  - All eigenvalues must be real
  - Eigenvectors corresponding to different eigenvalues are orthogonal
  - All eigenvalues are greater than or equal to zero
  - Covariance matrix can be diagonalized,

    i.e. $\mathbf{Cov = PDP^T}$