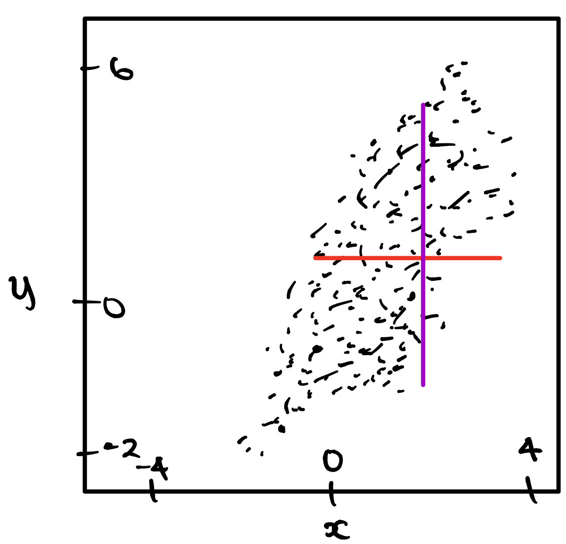
In 1-dimension the variance can be calculated by,

$$\mathrm{Var}\,X = \frac{1}{N}\sum^{N}(x-\bar{x})^2$$

In higher dimensions the squaring operation is undefined.
In addition variance can occur with respect to many different domains $x, y, z, \ldots$, etc.

The variance of data with respect to only one domain is ill suited to describing the relationship between two different variables/domains.

Covariance – describes the relationship between two variables.



As an example the variances are drawn on the left.

The variances of this dataset can be replicated by many other datasets – sometimes even with the same means.

This makes it impossible to explain any correlation between two variables using only the spread/variance of the data.

The extension of variance to covariance, is a tool that can be used to describe correlation.

$$\mathrm{Cov}[x,y] = E[(x-\mu_x)(y-\mu_y)]$$

$\mu_x = E[x]$

$\mu_y = E[y]$

Also can be written.

$\mathrm{var}\, X = \mathrm{cov}\, X = E[(X-E[X])(X-E[X])^T]$.

In two dimensions there are four measurements of interest.

Var x

Var Y

Cov x, Y

Cov Y, x

These can be used to form a Covariance Matrix

$$\left[\begin{array}{c|c} \mathrm{var}X & \mathrm{cov}X,Y \\ \hline \mathrm{cov}Y,X & \mathrm{var}Y \end{array}\right]$$

- If Cov $x, Y > 0$, then Y increases with x.
- If Cov $x, Y < 0$, then Y decreases with x.
- If Cov $x, Y = 0$, then there is no correlation.

The Covariance matrix is always a symmetric positive definite matrix. Variances on diagonal and the cross-covariance / covariances on the off diagonals.

$$\left[\begin{array}{c|c|c} \mathrm{var}X & \mathrm{cov}X,Y & \mathrm{cov}X,Z \\ \hline \mathrm{cov}Y,X & \mathrm{var}Y & \mathrm{cov}Y,Z \\ \hline \mathrm{cov}Z,X & \mathrm{cov}Z,Y & \mathrm{var}Z \end{array}\right]$$

The covariance matrix increases proportionate to the number of dimensions.

The covariance of high dimensional dataset can be described.

$$D = \{x_1, \dots, x_n\}, \quad x_i \in \mathbb{R}^D$$

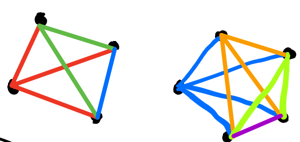
$$\mathrm{Cov}(V) = \frac{1}{n}\begin{bmatrix} x_1-\bar{x} & x_2-\bar{x} \\ y_1-\bar{y} & y_2-\bar{y} \end{bmatrix}\begin{bmatrix} x_1-\bar{x} & y_1-\bar{y} \\ x_2-\bar{x} & y_2-\bar{y} \end{bmatrix}$$

$\mathrm{var}[D]$ or $\mathrm{Cov}(D)$ $= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$, computes the covariance matrix $(D \times D)$ where μ is the mean value.

The covariance of two distributed real valued random variables

$$\begin{aligned}
\mathrm{cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\
&= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

x	1	2	3	4
y	5	6	7	8



diff, maybe avg

$$\frac{1}{4^2}\Bigg[\frac{(1-1)(5-5)}{2} + \frac{(1-2)(5-6)}{2} + \frac{(1-3)(5-7)}{2} + \frac{(1-4)(5-8)}{2}$$
$$\frac{(2-1)(6-5)}{2} + \frac{(2-2)(6-6)}{2} + \frac{(2-3)(6-7)}{2} + \frac{(2-4)(6-8)}{2}$$
$$\frac{(3-1)(7-5)}{2} \quad - \quad - \quad \dots$$
$$\frac{(4-1)(8-5)}{2} \quad - \quad \dots \quad - \quad \frac{(4-4)(8-8)}{2}$$

← Always 0.

$$\stackrel{?}{=}$$

$$\frac{1}{n}\Bigg[\left(1-\tfrac{1}{n}\Sigma x\right)\left(5-\tfrac{1}{n}\Sigma y\right) + \left(2-\tfrac{1}{n}\Sigma x\right)\left(6-\tfrac{1}{n}\Sigma y\right) + \left(3-\tfrac{1}{n}\Sigma x\right)\left(7-\tfrac{1}{n}\Sigma y\right) + \left(4-\tfrac{1}{n}\Sigma x\right)\left(8-\tfrac{1}{n}\Sigma y\right)\Bigg]$$

Why is $\mathrm{cov}(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)(y_i-y_j)\frac{1}{2}$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\boxed{j>i}}^{n}(x_i-x_j)(y_i-y_j) \; ?$$

$(x_1-\bar{x})(y_1-\bar{y})$ $\quad x_1y_1 - \bar{x}y_1 - x_1\bar{y} + \bar{x}\bar{y}$

$(x_2-\bar{x})(y_2-\bar{y})$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n}(x_iy_i - x_iy_j - x_jy_i + x_jy_j)$$

$$x_1y_1 \quad x_1y_2 \quad x_1y_3$$
$$x_2y_1 \quad x_2y_2 \quad x_2y_3$$
$$x_3y_1 \quad x_3y_2 \quad x_3y_3$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j>i}^{n}x_iy_i + \sum_{i=1}^{n}\sum_{j>i}^{n}(x_jy_j - x_iy_j - x_jy_i)\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n}x_iy_i + \sum_{i=1}^{n}\sum_{j>i}x_iy_i\right.$$

$$x_1y_1 - x_1y_1, \quad x_1y_1 - x_1y_2 \quad x_1y_1 - x_1y_3$$

$$x_2\cancel{y_2} - x_2y_1$$

$$x_3y_3 - x_3y_1$$

$$x_1(y_1 - y_1) \quad x_1(y_1 - y_2) \quad x_1(y_1 - y_3)$$

$$x_2(y_2 - y_1) \quad x_2(y_2 - y_3)$$

$$x_3(y_3 - y_1) \quad x_3(y_3 - y_2)$$

$$x_1(y_1(3) - y_1 - y_2 - y_3)$$
$$x_3(y_3(3) - y_1 - y_2 - y_3)$$

In geometry, an affine transformation is a vector manipulation which preserves points, straight lines and planes.

- does not preserve angles between lines
- does <u>not</u> preserve distance between points
- does preserve the ratio of distances between points.