Variance, Covariance, and Correlation

Terminology review

- Samples and Features
- Tidy Data Matrix

k columns

$$X = \begin{array}{c} N \\ \text{rows} \\ \text{(sample)} \end{array} \begin{pmatrix} \vdots & & \\ \vdots & & \end{pmatrix}$$

Mean

The sample mean of a feature is

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \mu_X = \frac{1}{N} \sum x_i$$

Variance

Definition: The (sample) variance of the feature measurements $x_1, \ldots, x_n$ is

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu_X)^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \mu_X^2$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad E = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \sigma_X^2 = \frac{1}{N} \| X - \mu_X E \|^2$$

Covariance

Definition: If $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ are two feature vectors then the (sample) covariance is

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)$$

$$X = \begin{array}{c} \\ N \end{array} \begin{array}{cc} X & Y \\ \left(\begin{array}{cc} \vdots & \vdots \\ \vdots & \vdots \end{array}\right) & \end{array}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\frac{1}{N}(X - \mu_X E) \cdot (Y - \mu_Y E)$$

Correlation

Definition: Given feature vectors X and Y, the (sample) correlation coefficient r_{XY} is

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_{XX}\sigma_{YY}} = \frac{(X-\mu_X E)\cdot(Y-\mu_Y E)}{\|(X-\mu_X E)\|\,\|(Y-\mu_Y E)\|} \quad \theta$$

$$a\cdot b = \|a\|\|b\| \cos\theta$$

$$\sigma_{XY} = \frac{1}{N}\cdot(X-\mu_X)\cdot(Y-\mu_Y)$$

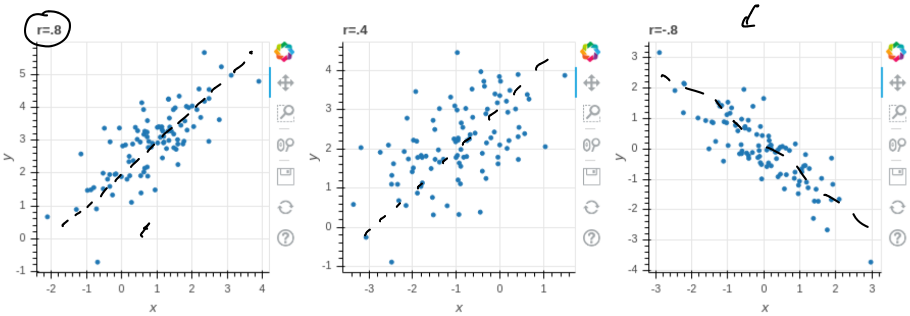
$$\sigma_{XX}^2 = \frac{1}{N}\|(X-\mu_X E)\|^2 \qquad \sigma_{YY}^2 = \frac{1}{N}\|(Y-\mu_Y E)\|^2$$

$X-\mu_X$

$r_{xy} = \cos\theta$

$Y-\mu_Y$

y_1 … y_N



$x_1 - - - - x_n$

$x\uparrow\ y\uparrow$ $\qquad$ $X\uparrow\ y\downarrow$

The covariance matrix

Definition: Let X be an $N \times k$ data matrix, and let X_0 be its centered version. The (sample) covariance matrix is the $k \times k$ symmetric matrix

$$\boxed{D_0 = \frac{1}{N} X_0^\intercal X_0.}$$

Covariance matrix

X_0 = centered data matrix
columns of X_0 have mean zero

X_0 = X, but subtract the average of each column from that column.

each column is the measurements of a feature

$$X_0 = \begin{pmatrix} x_{11} & x_{12} & & x_{1k} \\ x_{21} & x_{22} & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & & x_{Nk} \end{pmatrix}$$

d_{ii} diagonal entries of D_0

$$d_{11} = \frac{1}{N} \sum_{j=1}^{N} x_{j1}^2$$

$$d_{ii} = \frac{1}{N} \sum_{j=1}^{N} x_{ji}^2 = \sigma_{ii}^2$$

$X_0^T X_0$

$$X_0^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ \vdots & & & \\ x_{1k} & \cdots & & x_{Nk} \end{pmatrix}$$

$k \times N$

$d_{ij} = (i^{th}$ column of $X_0) \cdot \frac{1}{N} (j^{th}$ column of $X_0)$

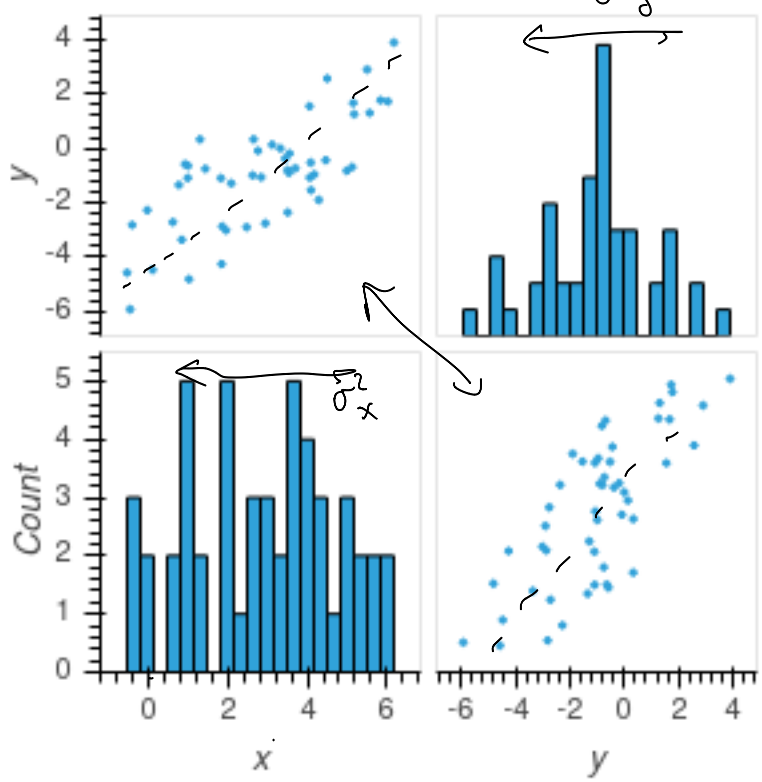
$= \frac{1}{N} \sigma_{ij}^2$

$$X_0^T X_0 = \begin{pmatrix} \sum_{j=1}^{N} x_{j1}^2 & \cdot & \\ & & \\ & & \end{pmatrix}$$

covariances

$$D_0 = \begin{pmatrix} \text{variances} & \cdot & \\ & \ddots & \\ & & \ddots \end{pmatrix}$$

Visualizing the covariance matrix



density grid plot.

50 points in $\mathbb{R}^2$

$$X_0 = \begin{array}{c} \begin{matrix} x & y \end{matrix} \\ \left(\begin{matrix} \quad & \quad \end{matrix} \right) \end{array}$$

50 samples

D_0 is a 2×2 matrix

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$