Sample statistics and linear regression

NEU 466M

Spring 2020

Mean

$$\{x_1, \cdots, x_N\}$$ N samples of variable x

$$\langle x \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} x_i \quad \text{sample mean}$$

$\texttt{mean}(\texttt{x})$

other notation: $\bar{x}$

Binned version of mean

$$\{x_1, \cdots, x_N\}$$ N samples of variable x

$$\{c_1, \cdots c_B\}, B \text{ bins}$$

$$\{n_1, \cdots n_B\} \text{ counts per bin}$$

$$\langle x \rangle \equiv \frac{1}{N} \sum_{i=1}^{B} n_i c_i \qquad \text{sample mean}$$

Variance

$$\{x_1, \cdots, x_N\}$$

$$\langle (x - \langle x \rangle)^2 \rangle \equiv \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2$$

sample variance

a measure of the "scatter"/spread of the data around its mean value

homework: show that $\langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$

Standard deviation

$$\{x_1, \cdots, x_N\}$$

$$\sqrt{\langle(x - \langle x\rangle)^2} \quad \text{standard deviation}$$

Covariance

$$\{x_1, \cdots, x_N\}\{y_1, \cdots, y_N\}$$

N samples each of variables x, y

$$C(x, y) \equiv \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle)(y_i - \langle y \rangle)$$

sample covariance

($C(x, x)$ is simply sample variance of x)

Covariance: what does it measure?

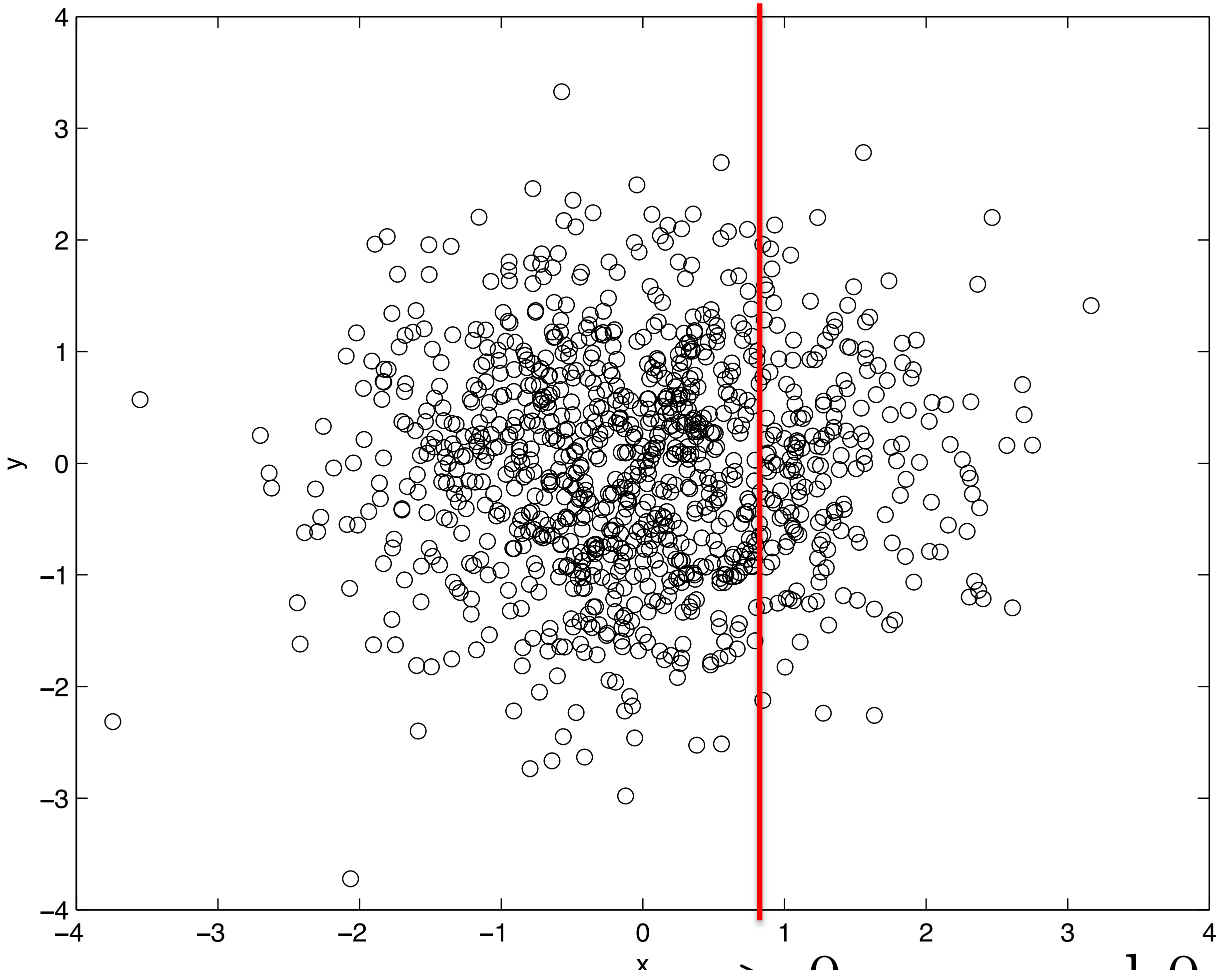
$$C(x,y) \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \langle x \rangle)(y_i - \langle y \rangle)$$

- If x, y both deviate from their means together (both up then both down) then terms in sum are positive, C(x,y) > 0.
- If x,y deviate from their means independent of each other, then terms in the sum are randomly positive and negative, C(x,y) ~=0.
- If x,y deviate from their means in opposite directions, then terms in sum are negative, C(x,y) < 0.

Literally, covariance is a measure of co-variation.

Covariance example I

x, y independent



$\texttt{x} = \texttt{randn}(1000, 1)$

$\texttt{y} = \texttt{randn}(1000, 1)$

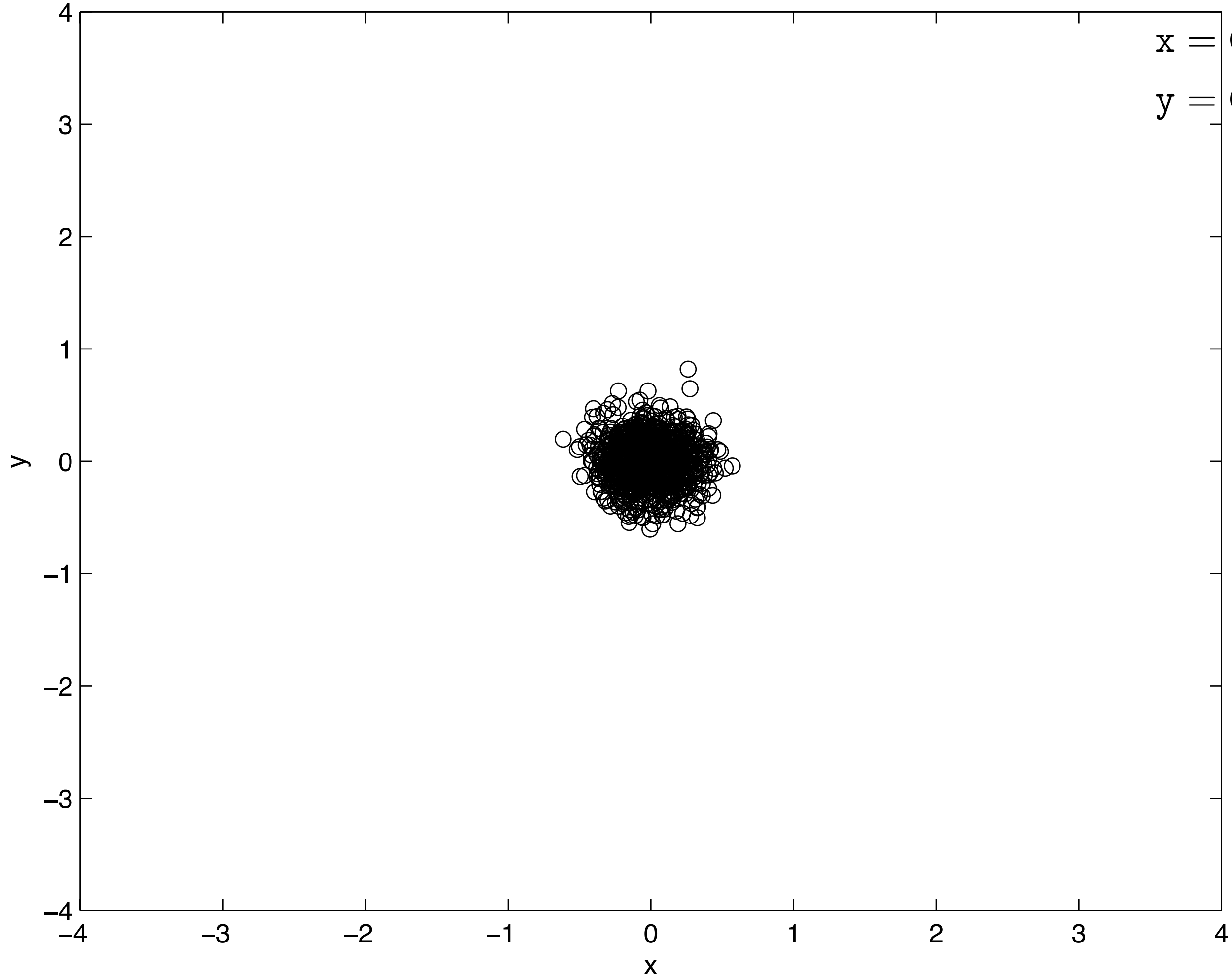
$C(x, y) = 0.009;$

$C(x, x) = 1.069$

$x > 0, y$ around 0 without bias

Covariance example II

x, y independent



$\mathtt{x} = 0.2 * \mathtt{randn}(1000, 1)$

$\mathtt{y} = 0.2 * \mathtt{randn}(1000, 1)$

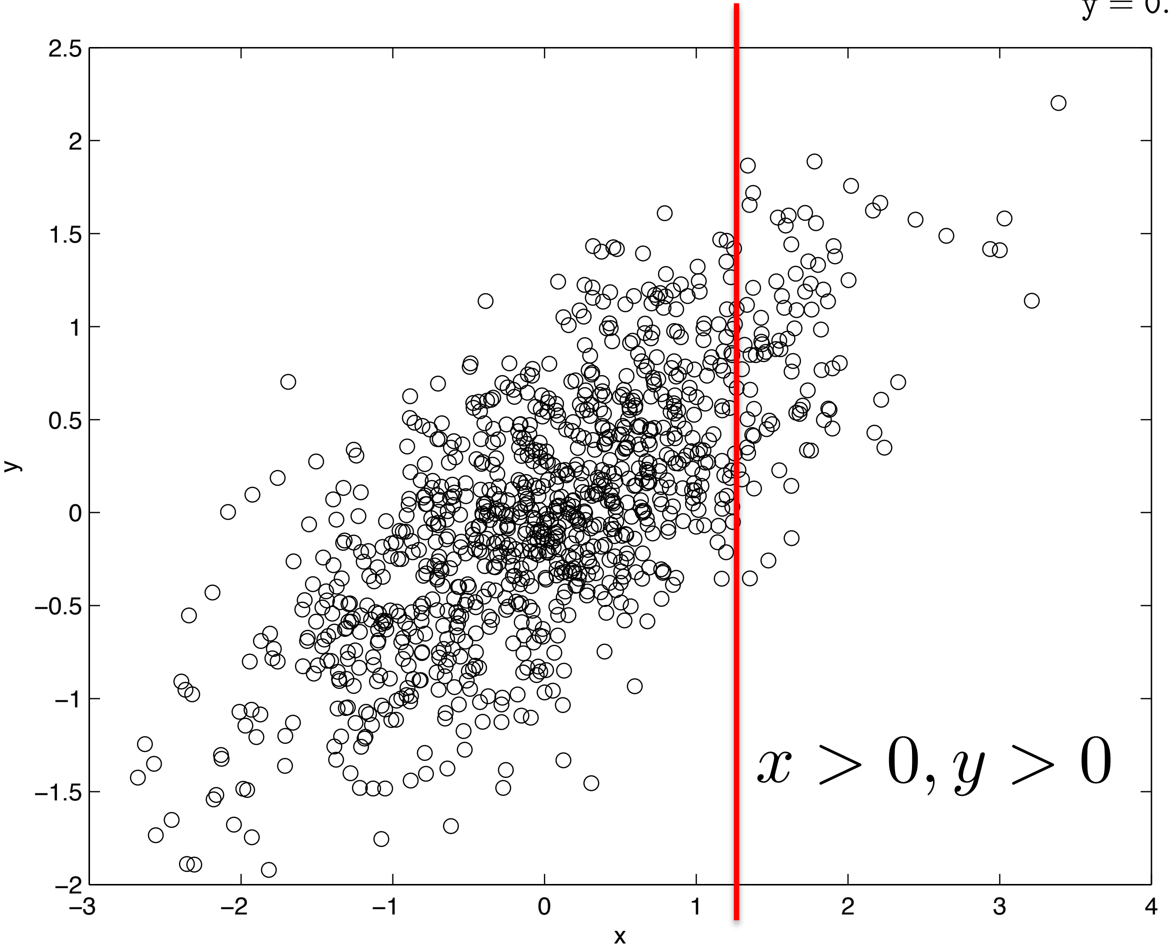
$$C(x, y) = 0.001;\ C(x, x) = 0.0407$$

Covariance example III

x, y not independent

$\mathtt{x} = \mathtt{randn}(1000, 1)$

$\mathtt{y} = 0.5 * \mathtt{x} + 0.5 * \mathtt{randn}(1000, 1)$



$$C(x, x) = 0.907;\ C(x, y) = 0.464;\ C(y, y) = 0.469$$

Alternative notation

- Mean: $\langle x \rangle,\ \bar{x},\ \mu_x,\ E(x)$

- Variance: $\langle x^2 \rangle - \langle x \rangle^2,\ \overline{x^2} - \bar{x}^2,\ \sigma_x^2,\ var(x),\ C(x,x)$

- Covariance:

$$\langle xy \rangle - \langle x \rangle \langle y \rangle,\ \overline{xy} - \bar{x}\bar{y},\ \sigma_{xy}^2,\ cov(x),\ C(x,y)$$

- Standard deviation

$$\sqrt{\langle x^2 \rangle - \langle x \rangle^2},\ \sqrt{\overline{x^2} - \bar{x}^2},\ \sigma_x,\ std(x)$$

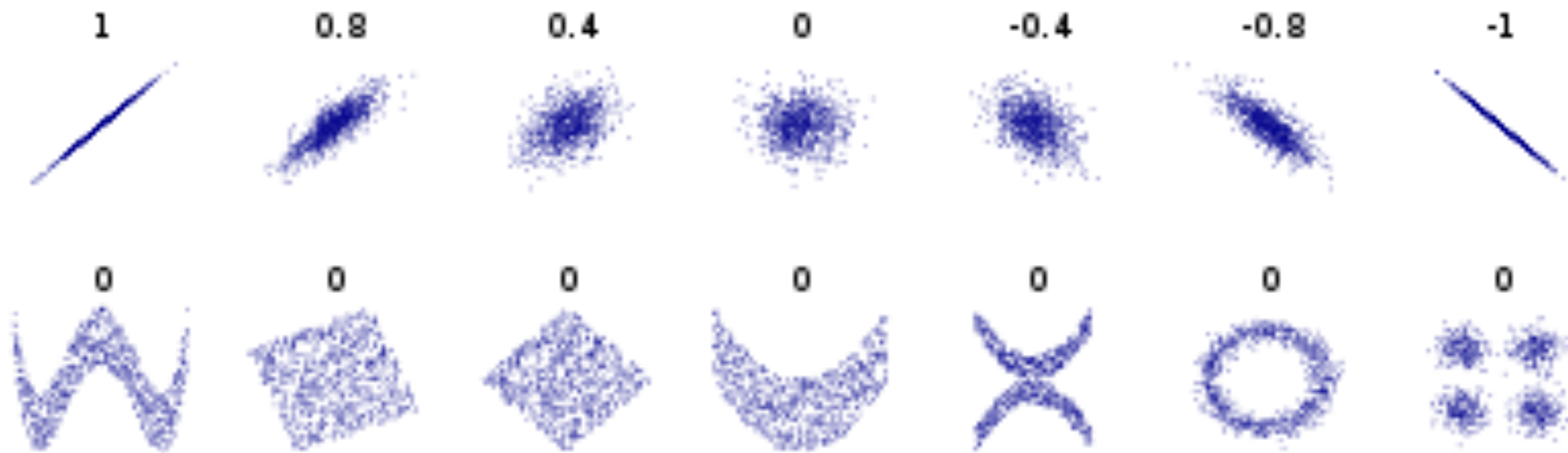
Pearson's correlation coefficient

$$\rho(x,y) = \frac{\Big\langle \big(x - \langle x \rangle\big)\big(y - \langle y \rangle\big) \Big\rangle}{\sqrt{\Big\langle \big(x - \langle x \rangle\big)^2 \Big\rangle \Big\langle \big(x - \langle x \rangle\big)^2 \Big\rangle}}$$

$$\rho(x,y) = \frac{C(x,y)}{\sigma_x \sigma_y}$$

shorter-form notation

Pearson's correlation coefficient and covariance only measure *linear dependency*



from: https://en.wikipedia.org/wiki/Correlation_and_dependence

Robust statistics?

- Mean, variance are easy to compute, widely used/useful.
- But not robust: sensitive to outliers.
- More robust alternative to mean: median.

Application

LINEAR REGRESSION IN TERMS OF SAMPLE STATISTICS

Regression: curve-fitting

Scalar explanatory variable (X) and response variable (Y); N samples

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

$$\tilde{y}(x) = w_0 + w_1 x + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$\text{free parameters: } (w_0, w_1, \cdots, w_M)$$

Linear least-squares regression

$$E = \frac{1}{2}\sum_{n=1}^{N}[\tilde{y}(x_n; \mathbf{w}) - y_n]^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}[\sum_{j=0}^{M} w^j x_n^j - y_n]^2$$

M=1 for linear regression

$$= \frac{1}{2}\sum_{n=1}^{N}[w^0 + w^1 x_n - y_n]^2$$

To solve for best w^0, w^1:

$$\frac{dE}{dw^0} = 0, \quad \frac{dE}{dw^1} = 0$$

Linear least-squares regression

$$E = \frac{1}{2} \sum_{n=1}^{N} [w^0 + w^1 x_n - y_n]^2$$

$$\frac{dE}{dw^0} = \sum_{n=1}^{N} [w^0 + w^1 x_n - y_n]$$

$$= N w^0 + N w^1 \langle x \rangle - N \langle y \rangle = 0$$

$$w^0 + w^1 \langle x \rangle - \langle y \rangle = 0 \qquad (1)$$

Linear least-squares regression

$$E = \frac{1}{2} \sum_{n=1}^{N} [w^0 + w^1 x_n - y_n]^2$$

$$\frac{dE}{dw^1} = \sum_{n=1}^{N} [w^0 + w^1 x_n - y_n] x_n$$

$$= N w^0 \langle x \rangle + N w^1 \langle x^2 \rangle - N \langle xy \rangle = 0$$

$$w^0 \langle x \rangle + w^1 \langle x^2 \rangle - \langle xy \rangle = 0 \quad (2)$$

Linear least-squares regression

$$w^1 = \frac{C(x,y)}{C(x,x)} \quad \text{slope}$$

$$w^0 = \langle y \rangle - w^1 \langle x \rangle \quad y - \text{intercept}$$

In homework: check matlab's polyfit with this optimal expression for linear-least squares fitting.

Linear least-squares regression

$$w^1 = \frac{C(x,y)}{C(x,x)} \quad \text{slope}$$

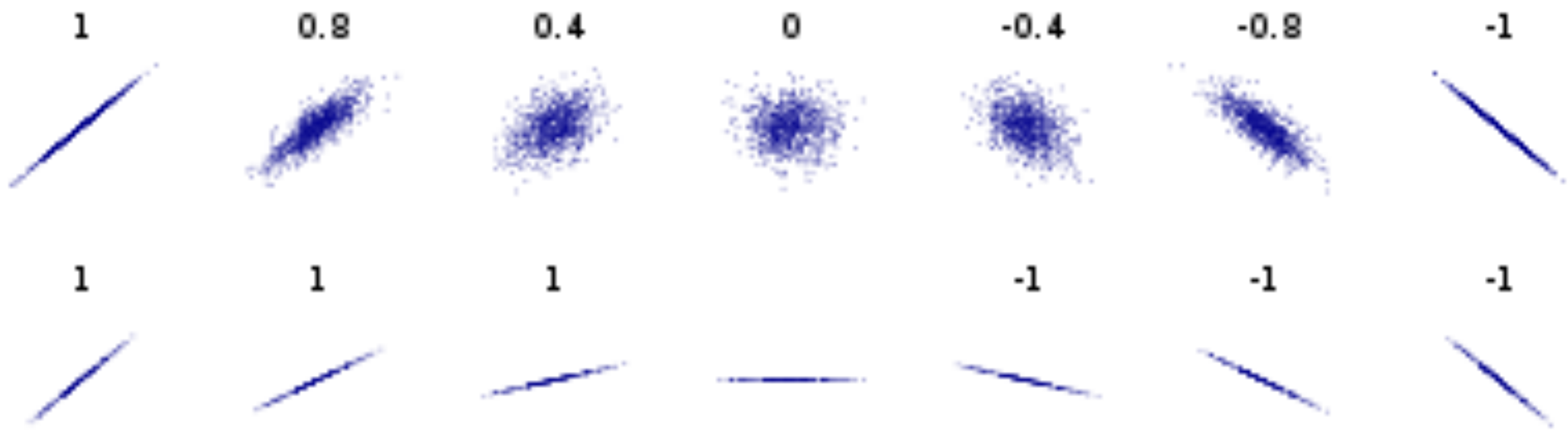
$$w^0 = \langle y \rangle - w^1 \langle x \rangle \quad y - \text{intercept}$$

Contrast with w^1: Pearson's correlation $\rho(x,y) = \frac{C(x,y)}{\sigma_x \sigma_y}$

Different normalizations:

- Different correlation coefficient for same slope but different amounts of x,y-scatter.
- Same correlation for different slopes and different x,y scatter.
- Correlation: more strongly penalizes y-scatter, more weakly penalizes x-scatter.

Slope versus Pearson's correlation coefficient



same slope
different ρ

different
slope, same ρ

from: https://en.wikipedia.org/wiki/Correlation_and_dependence

Application

BACK TO SAMPLE STATISTICS: MULTIVARIATE

Multiple variables: covariance matrix

$$\{x_{\alpha 1}, \cdots, x_{\alpha N}\}$$ N samples of the αth variable x_α

K different variables x_α, labeled by α, β = {1,…,K}:

$$C_{\alpha\beta} \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_{\alpha i} - \langle x_\alpha \rangle)(x_{\beta i} - \langle x_\beta \rangle)$$

$$= cov(x_\alpha, x_\beta)$$

$K \times K$ dim since K variables

sample covariance matrix

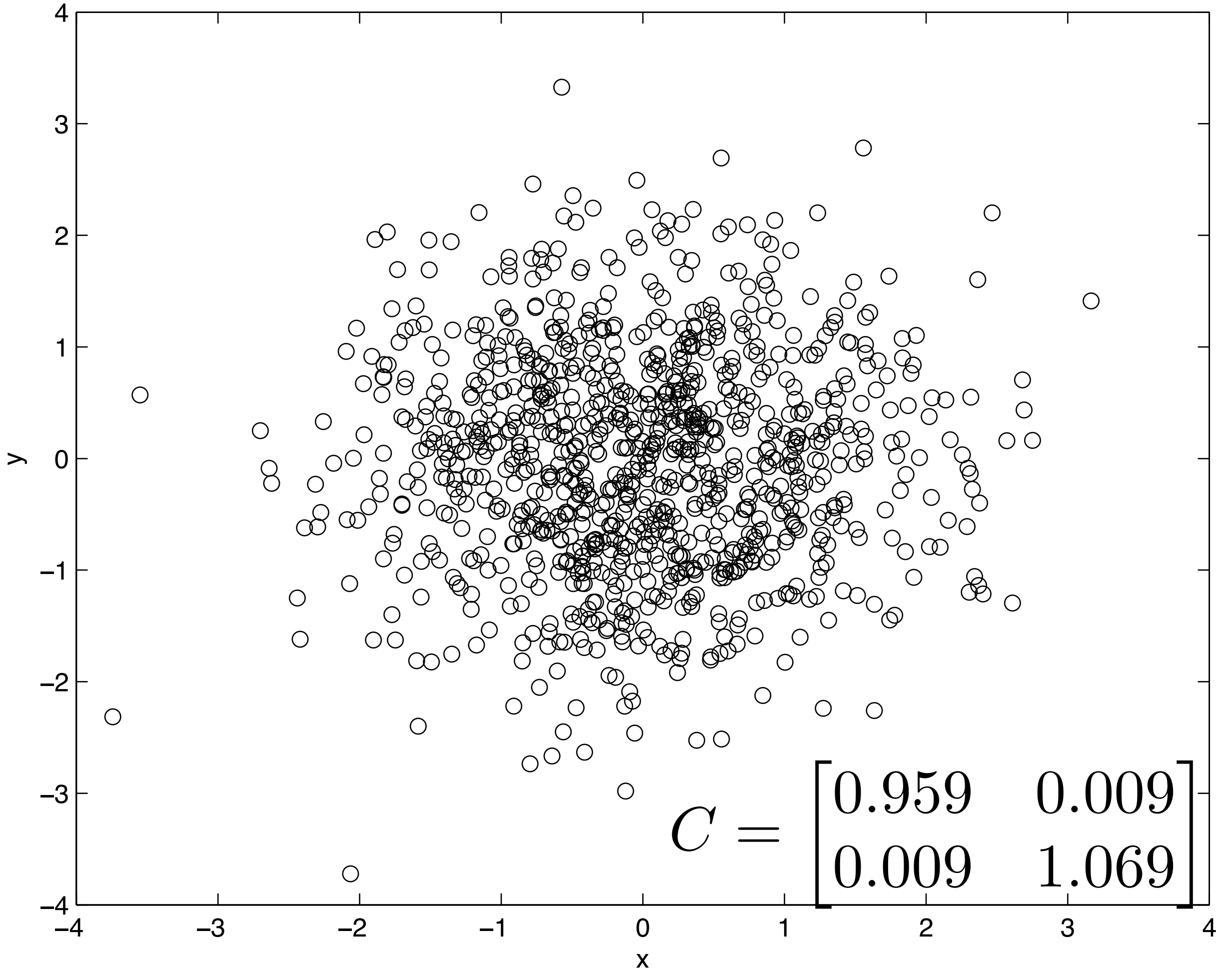
Covariance matrix

- (α,β) element is covariance between x_α, x_β.
- Diagonal of covariance matrix is variance of each variable: $var(x_\alpha)$ or $C(x_\alpha, x_\alpha)$.
- K^2 entries total, but only half of off-diagonal terms are independent because of symmetry ($C(x_\beta, x_\alpha) = C(x_\alpha, x_\beta)$).
- Thus only $(K^2-K)/2 + K = K(K+1)/2$ independent terms.

Q's: How do do linear regression in multivariate case? Will it involve covariance matrix?

Covariance example I

x, y independent



$$C = \begin{bmatrix} 0.959 & 0.009 \\ 0.009 & 1.069 \end{bmatrix}$$

$\mathtt{x} = \mathtt{randn}(1000, 1)$

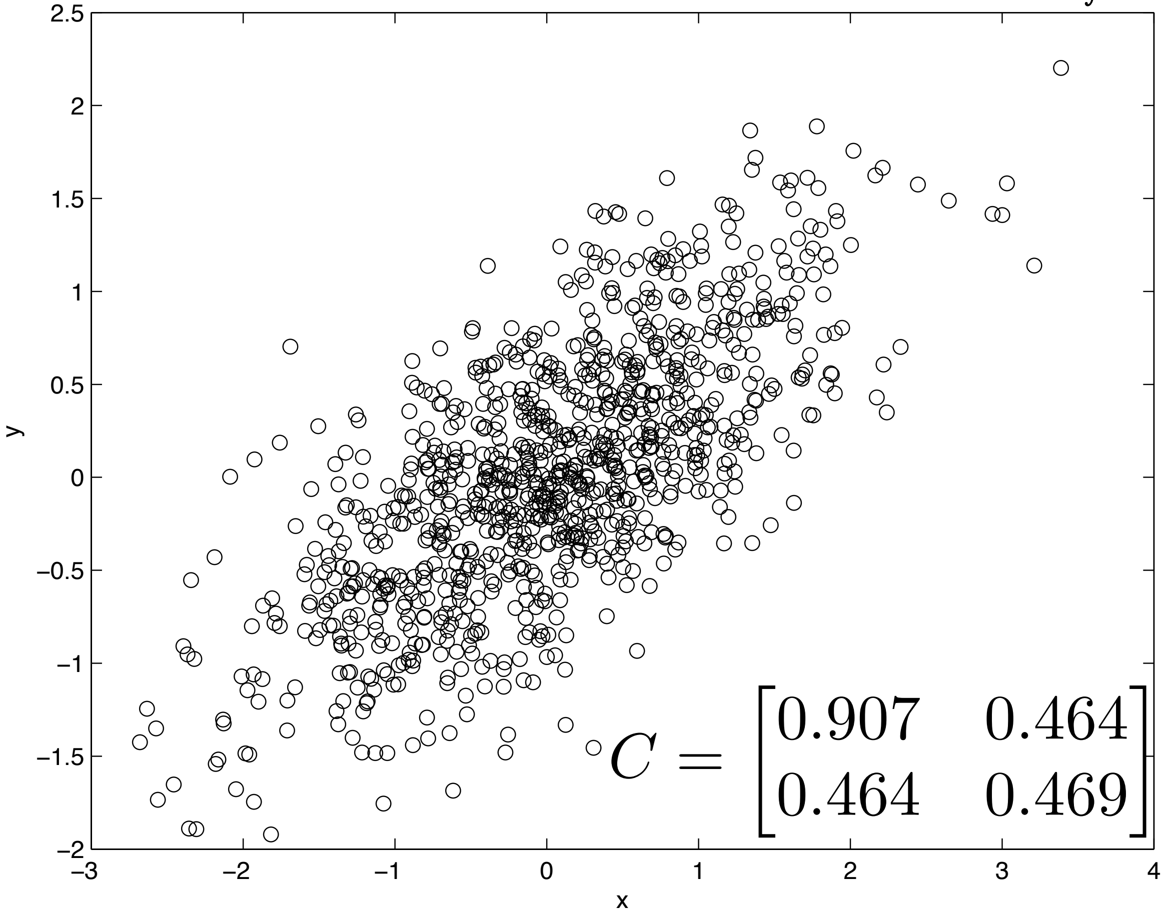
$\mathtt{y} = \mathtt{randn}(1000, 1)$

Covariance example III

x, y not independent

$\texttt{x} = \texttt{randn}(1000, 1)$

$\texttt{y} = 0.5 * \texttt{x} + 0.5 * \texttt{randn}(1000, 1)$



$$C = \begin{bmatrix} 0.907 & 0.464 \\ 0.464 & 0.469 \end{bmatrix}$$

Summary

- Defined sample mean and variance of a variable
- Defined covariance between a pair of variables
- Solved optimal (least-squares) linear regression between two variables in terms of mean, covariance
- Covariance matrix: covariance between all K(K+1)/2 unique pairs of K variables