# Aljabar Matriks:
# Ukuran Jarak

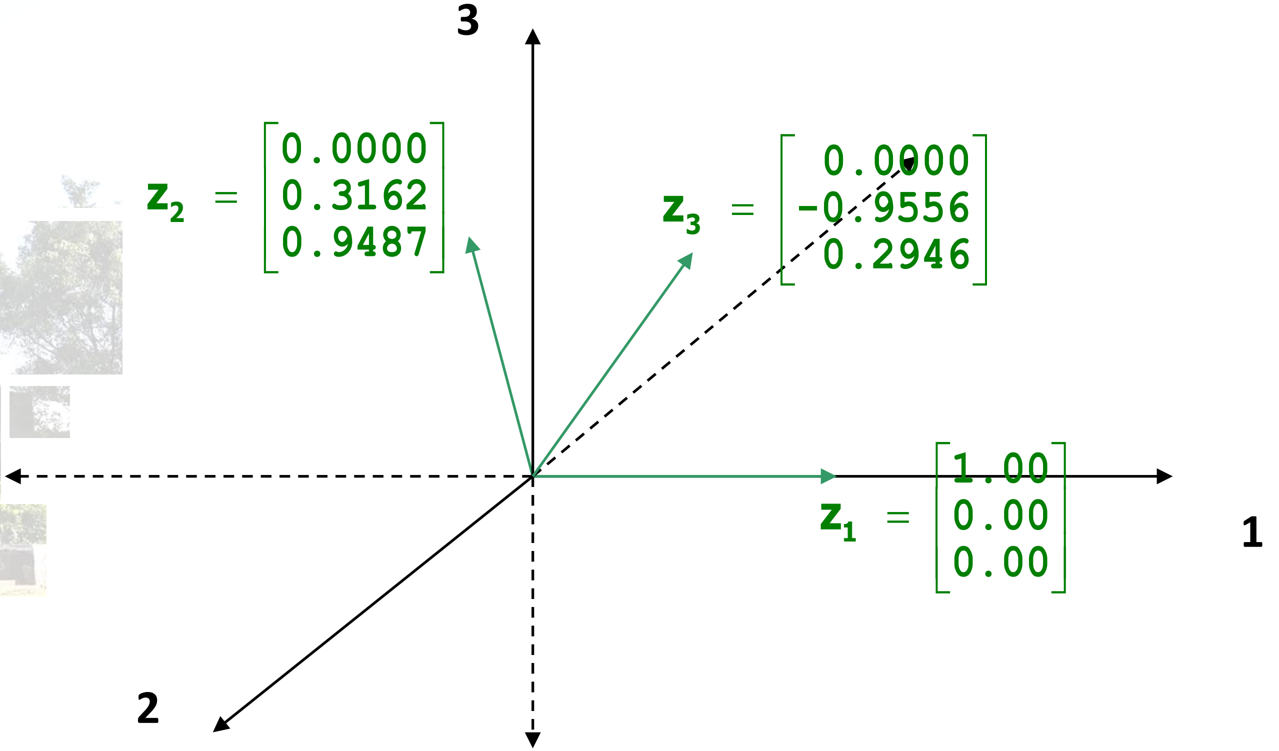
# Measuring Distance

Euclidean (straight line) distance – The *Euclidean* distance between two points **x** and **y** (whose coordinates are represented by the elements of the corresponding vectors) in p-space is given by

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

# Measuring Distance

For a previous example



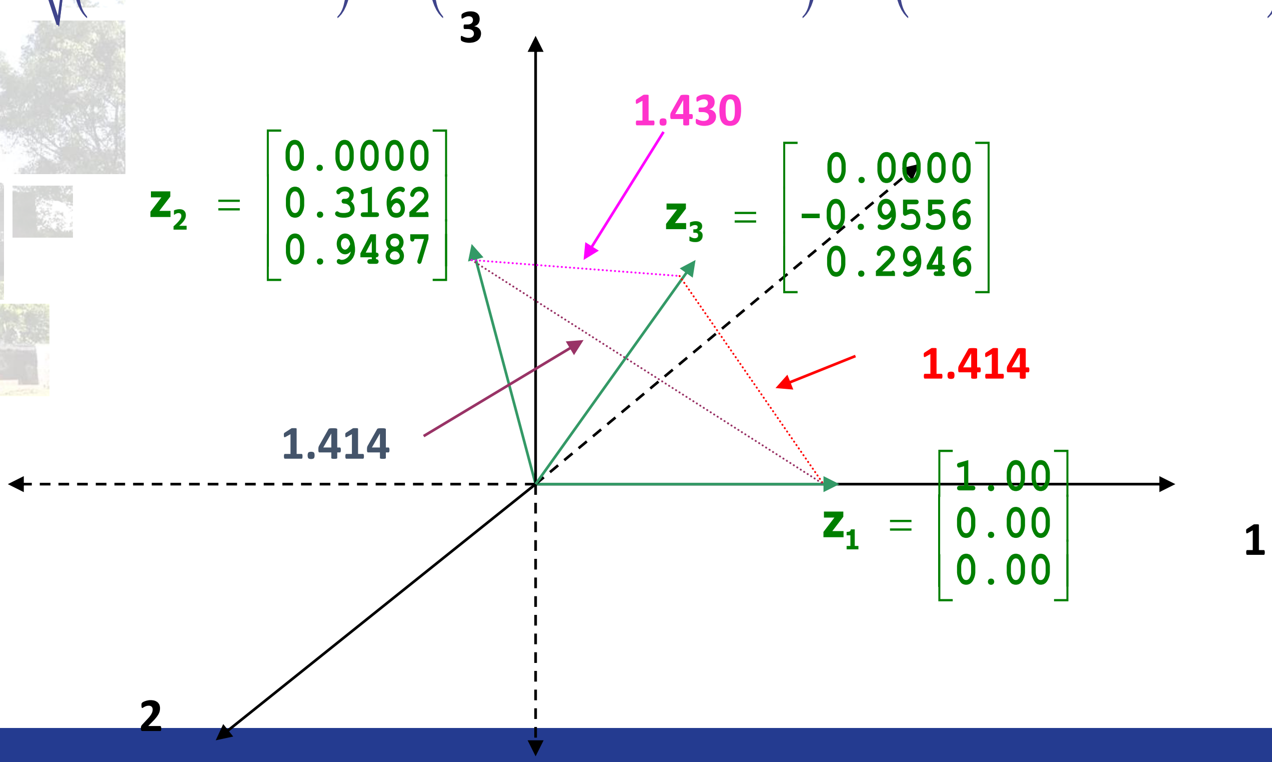

the Euclidean (straight line) distances are

# Measuring Distance

$$d(\mathbf{z}_1,\mathbf{z}_2) = \sqrt{(1.00-0.00)^2 + (0.00-0.3162)^2 + (0.00-0.9487)^2} = 1.414$$

$$d(\mathbf{z}_1,\mathbf{z}_3) = \sqrt{(1.00-0.00)^2 + (0.00+0.9556)^2 + (0.00-0.2946)^2} = 1.414$$

$$d(\mathbf{z}_2,\mathbf{z}_3) = \sqrt{(0.00-0.00)^2 + (0.3162+0.9556)^2 + (0.9487-0.2946)^2} = 1.430$$



# Measuring Distance

Notice that the lengths of the vectors are their distances from the origin:

$$d(\mathbf{0},\mathbf{P}) = \sqrt{(x_1 - 0)^2 + \cdots + (x_p - 0)^2}$$
$$= \sqrt{x_1^2 + \cdots + x_p^2}$$

This is yet another place where the Pythagorean Theorem rears its head!

# Measuring Distance

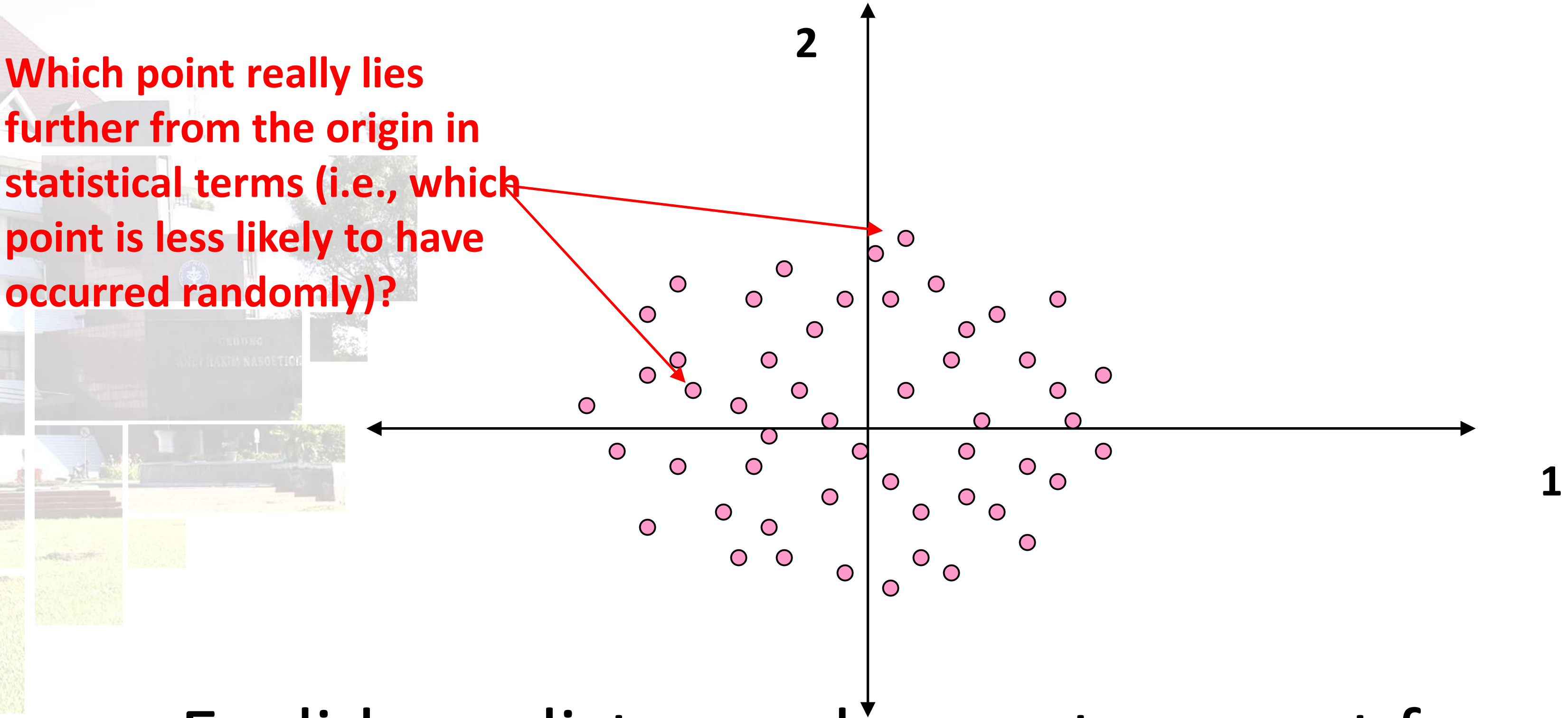
Problem – What if the coordinates of a point **x** (i.e., the elements of vector **x**) are random variables with differing variances?

Suppose

- we have n pairs of measurements on two variables $X_1$ and $X_2$, each having a mean of zero
- $X_1$ is more variable than $X_2$
- $X_1$ and $X_2$ vary independently

# Measuring Distance

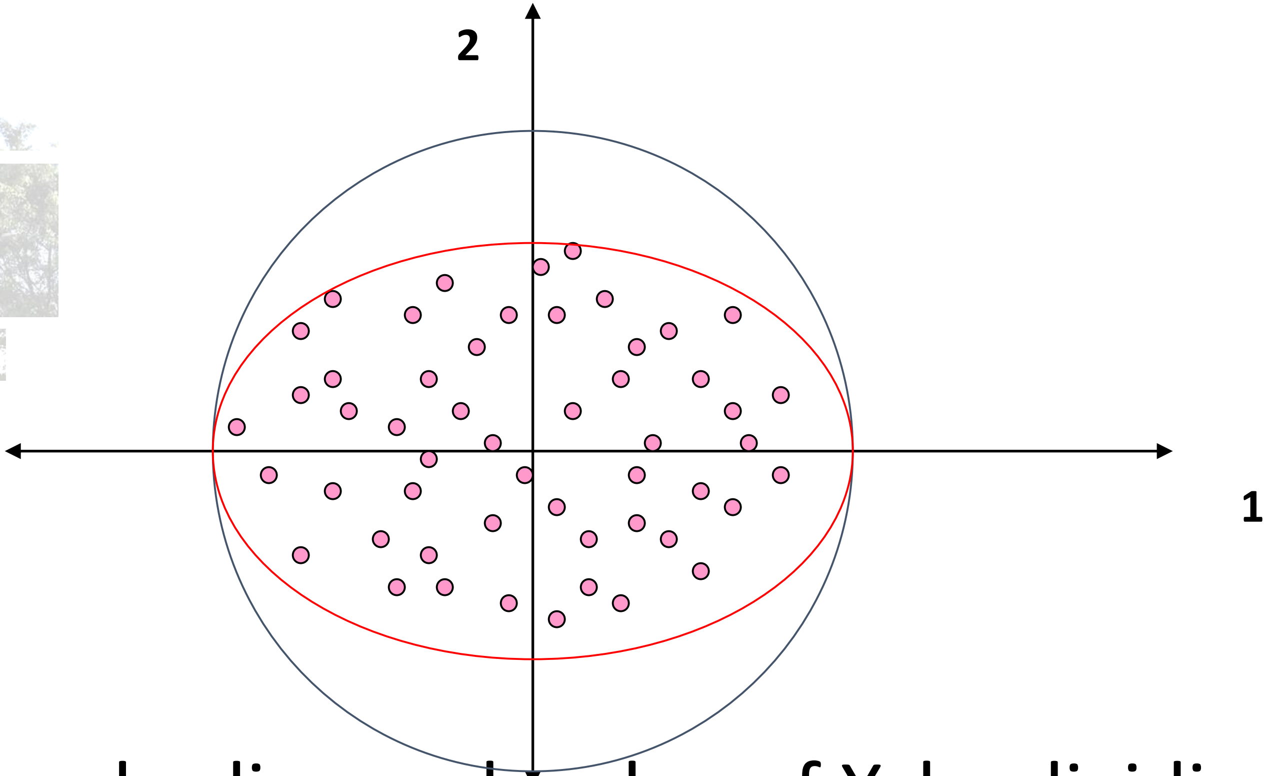
A scatter diagram of these data might look like this:




Euclidean distance does not account for differences in variation of $X_1$ and $X_2$!

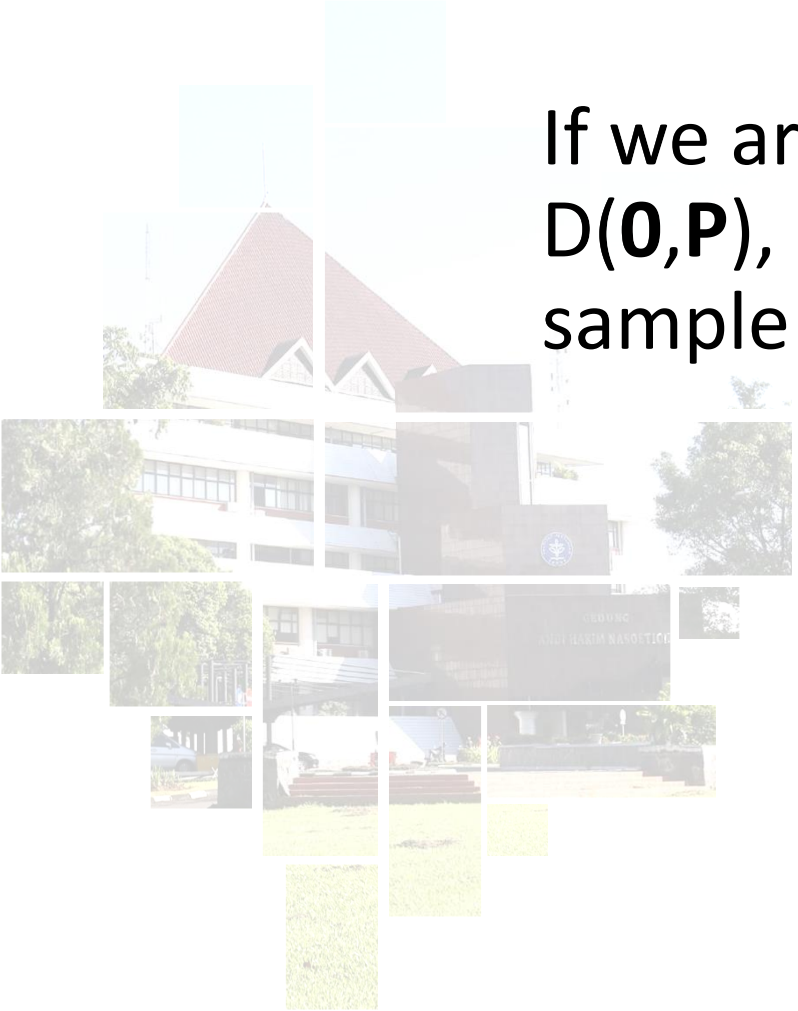
# Measuring Distance

How do we take the relative dispersions on the two axes into consideration?




We standardize each value of $X_i$ by dividing by its standard deviation.

# Measuring Distance

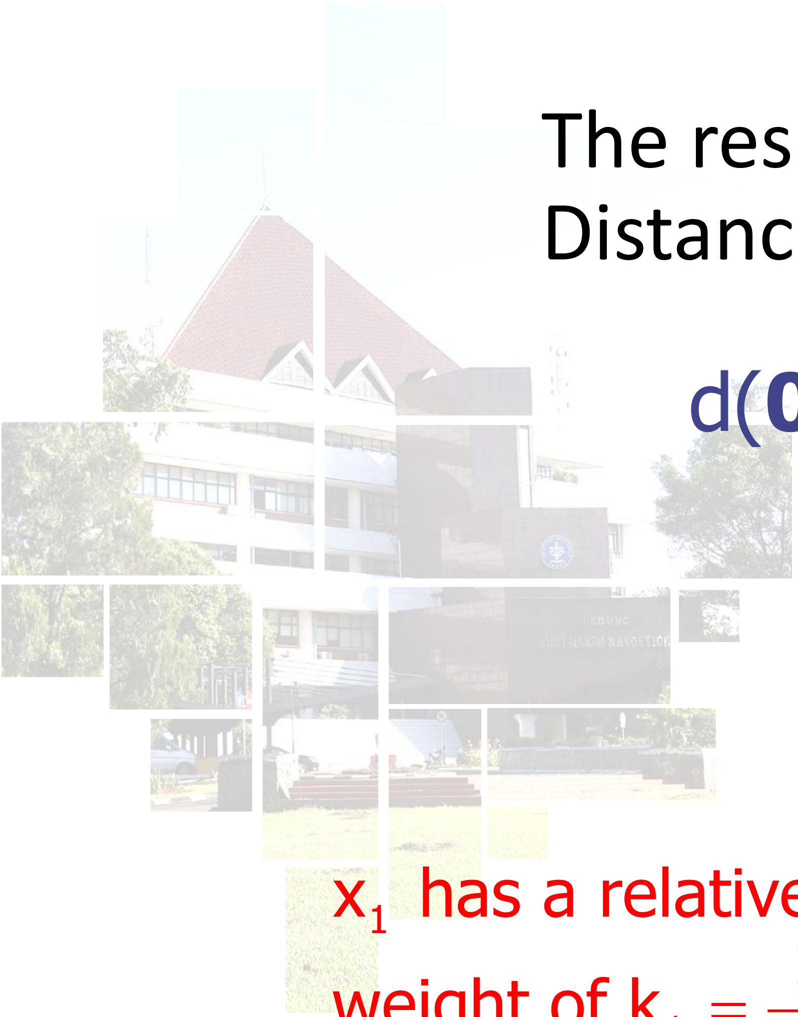


If we are looking at distances from the origin D(**0**,**P**), we could divide coordinate i by its sample standard deviation $\sqrt{s_{ii}}$:

$$x_i^* = \frac{x_i}{\sqrt{s_{ii}}}$$

IPB University
Bogor Indonesia

# Measuring Distance



The resulting measure is called Statistical Distance or Mahalanobis Distance:

$$d(\mathbf{0},\mathbf{P}) = \sqrt{\left(x_1^*\right)^2 + \cdots + \left(x_p^*\right)^2}$$

$$= \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \cdots + \left(\frac{x_p}{\sqrt{s_{pp}}}\right)^2}$$

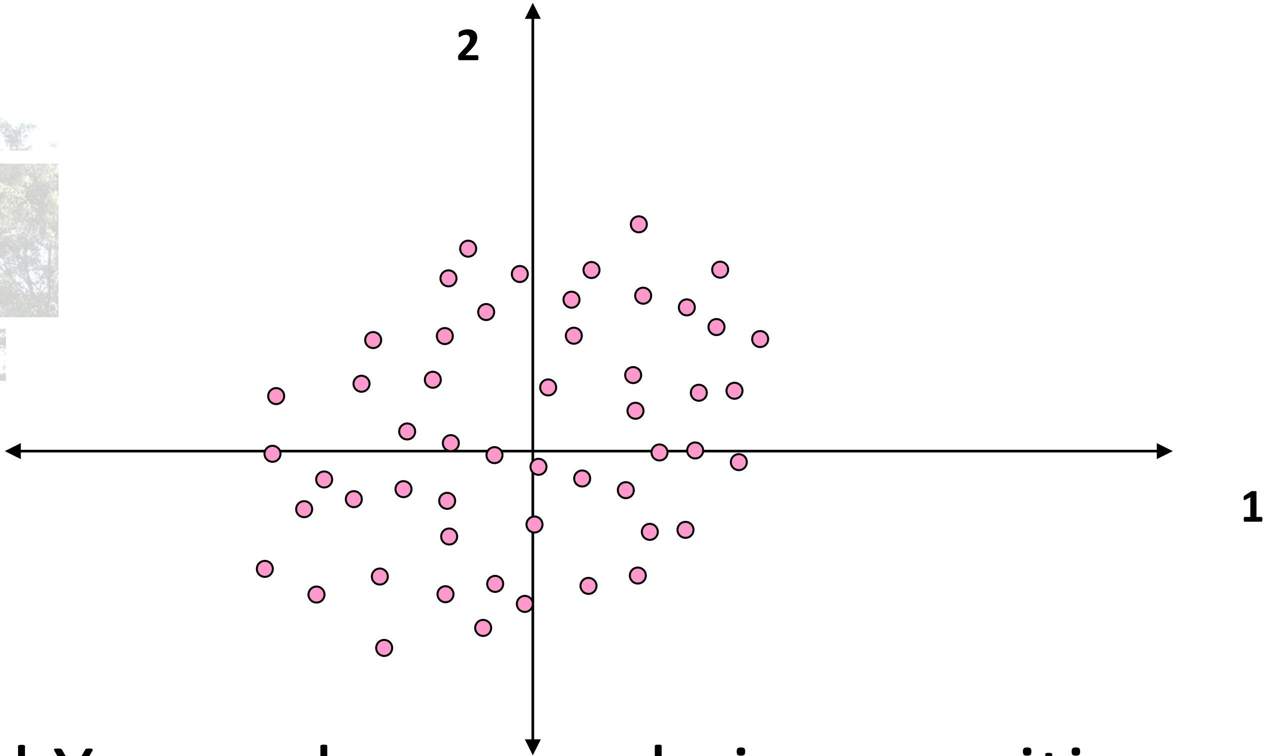
$$= \sqrt{\frac{x_1^2}{s_{11}} + \cdots + \frac{x_p^2}{s_{pp}}}$$

$x_1$ has a relative weight of $k_1 = \dfrac{1}{s_{11}}$

$x_p$ has a relative weight of $k_p = \dfrac{1}{s_{pp}}$

# Measuring Distance

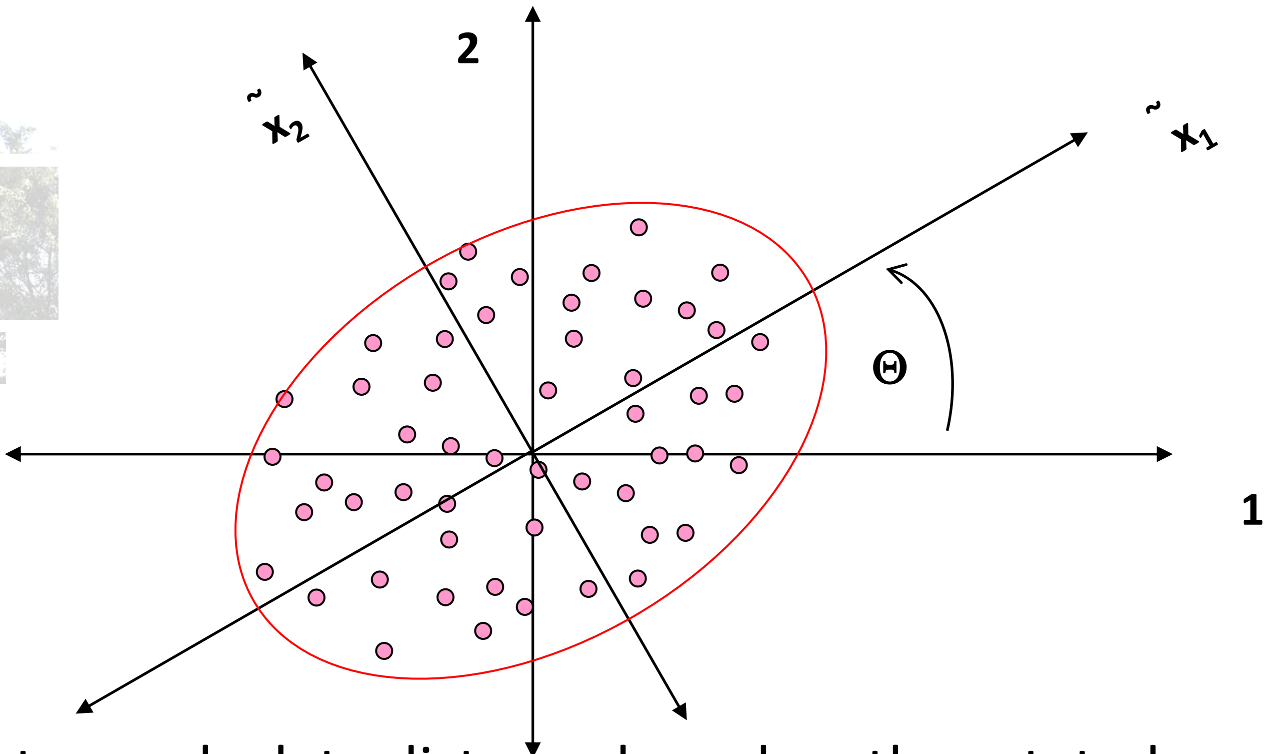
What if the scatter diagram of these data looked like this:




$X_1$ and $X_2$ now have an obvious positive correlation!

# Measuring Distance

We can plot a rotated coordinate system on axes $\tilde{x}_1$ and $\tilde{x}_2$:




This suggests that we calculate distance based on the rotated axes $\tilde{x}_1$ and $\tilde{x}_2$.

# Measuring Distance

The relation between the original coordinates $(x_1, x_2)$ and the rotated coordinates $(\tilde{x}_1, \tilde{x}_2)$ is provided by:

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$

$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

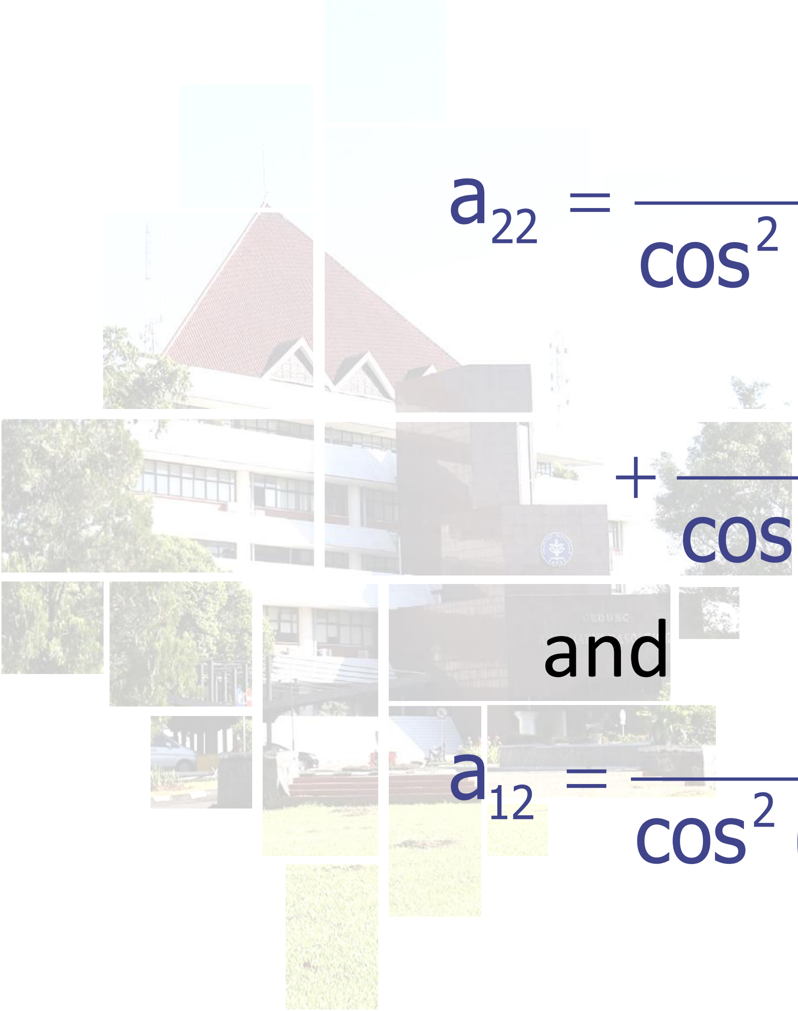
# Measuring Distance

Now we can write the distance from $P = (\tilde{x}_1, \tilde{x}_2)$ to the origin in terms of the original coordinates $x_1$ and $x_2$ of P as

$$d(\mathbf{0},\mathbf{P}) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

where

$$a_{11} = \frac{\cos^2(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} + \frac{\sin^2(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$



# Measuring Distance

$$a_{22} = \frac{\sin^2(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} + \frac{\cos^2(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$

and

$$a_{12} = \frac{\cos(\theta)\sin(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} - \frac{\sin(\theta)\cos(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$

# Measuring Distance

Note that the distance from P = $(x_1, x_2)$ to the origin for uncorrelated coordinates $x_1$ and $x_2$ is

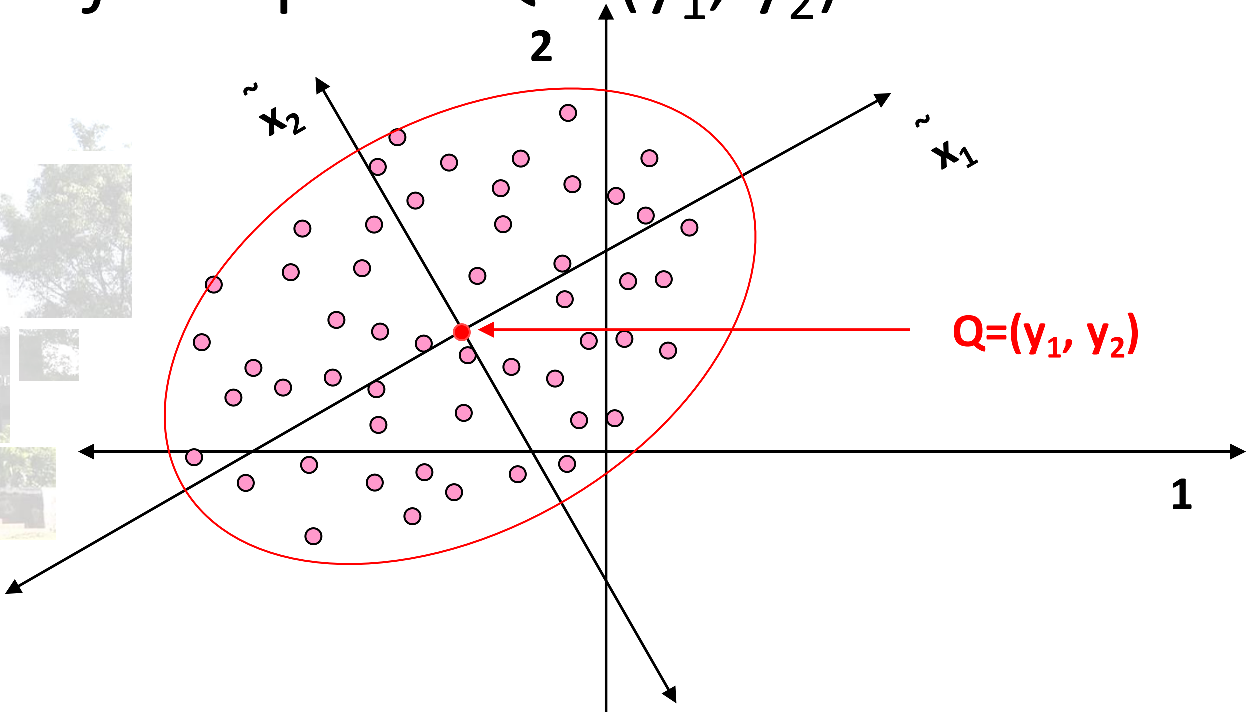
$$d(\mathbf{0},\mathbf{P}) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

for weights

$$a_{ij} = \frac{1}{s_{ij}}$$

# Measuring Distance

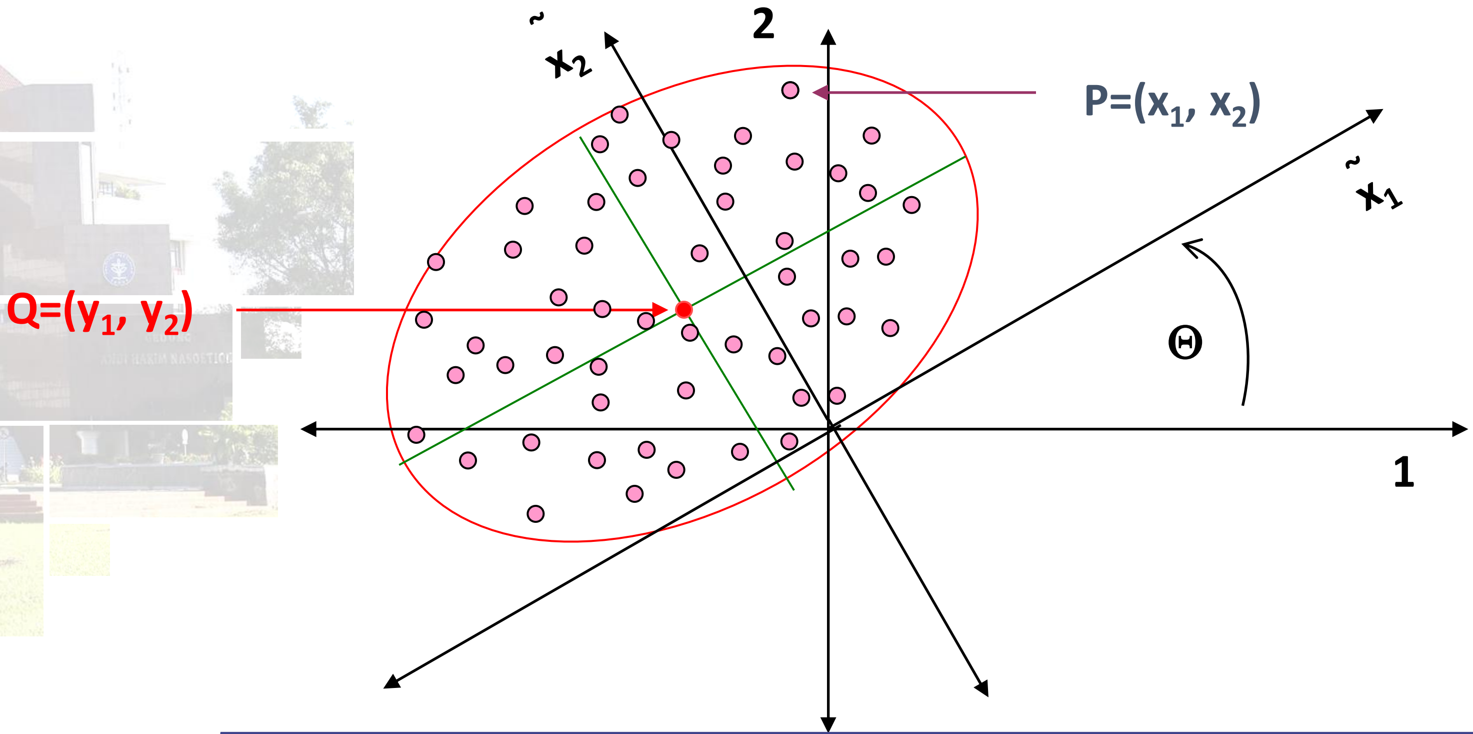
What if we wish to measure distance from some *fixed* point $Q = (y_1, y_2)$?




In this diagram, $Q = (y_1, y_2) = (\bar{x}_1, \bar{x}_2)$ is called the centroid of the data.

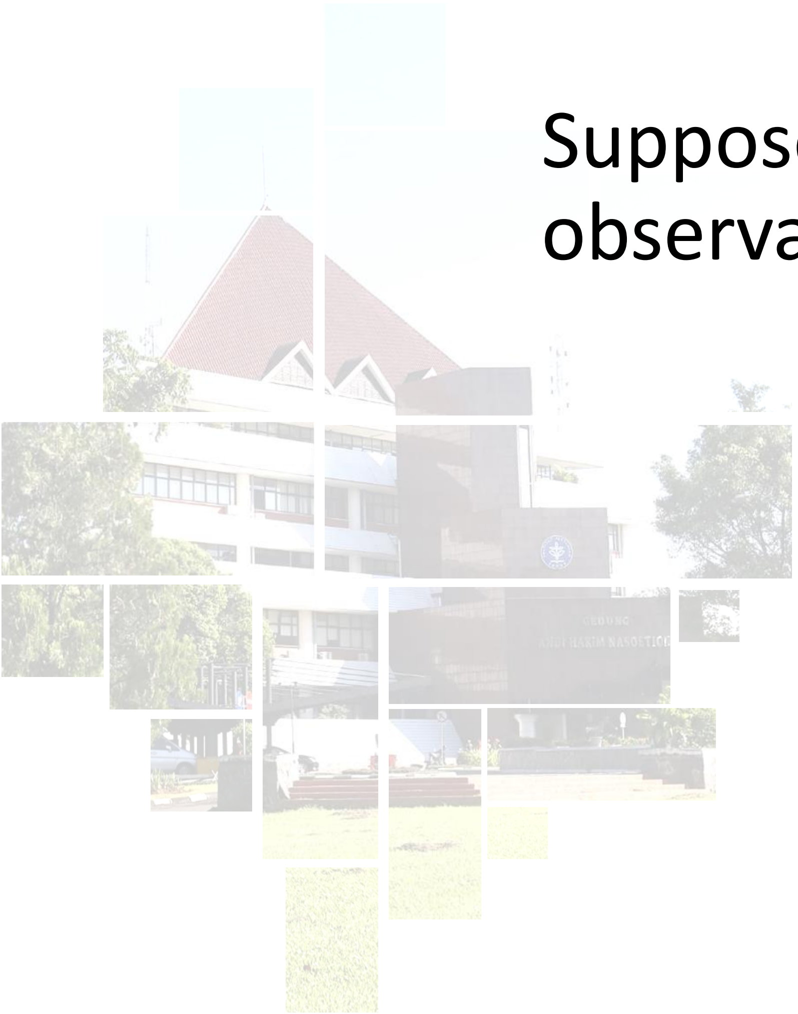
# Measuring Distance

The distance from any point p to some fixed point Q = $(y_1, y_2)$ is




$$d(\mathbf{P},\mathbf{Q}) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

# Measuring Distance

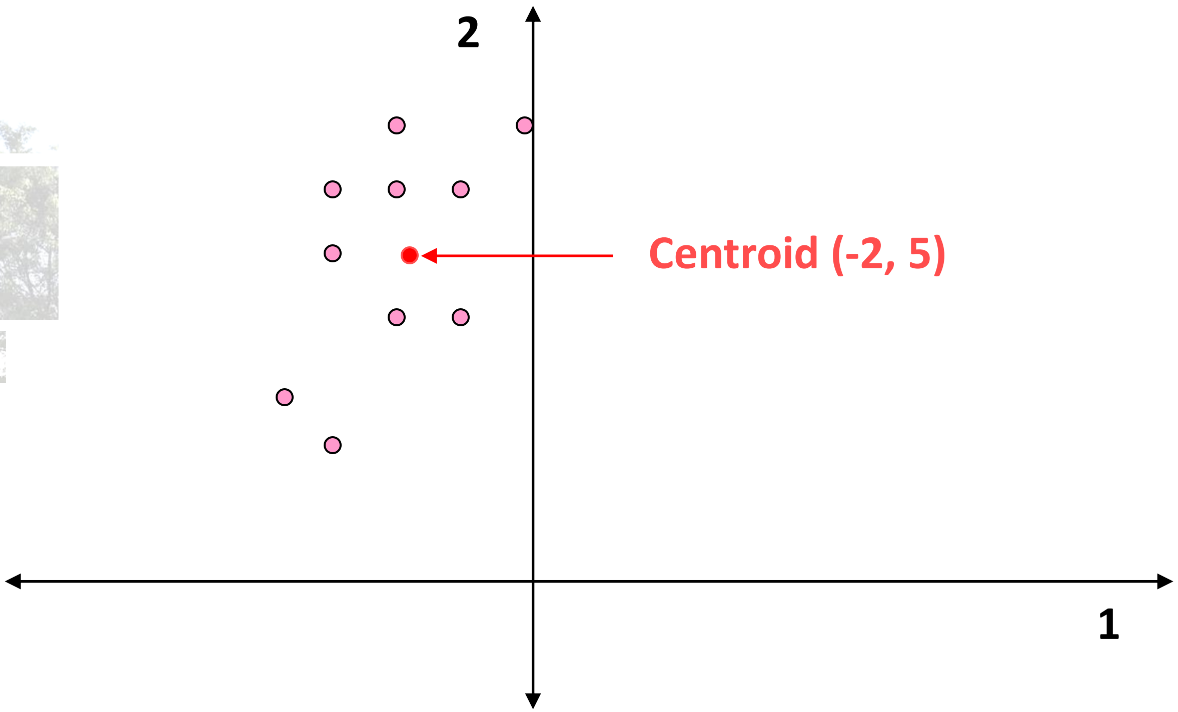


Suppose we have the following ten bivariate observations (coordinate sets of $(x_1, x_2)$):

| Obs # | $x_1$ | $x_2$ |
|---|---|---|
| 1 | -3 | 3 |
| 2 | -2 | 6 |
| 3 | -1 | 4 |
| 4 | -2 | 4 |
| 5 | -3 | 6 |
| 6 | -1 | 6 |
| 7 | -3 | 2 |
| 8 | -3 | 5 |
| 9 | 0 | 7 |
| 10 | -2 | 7 |
| $x_i$ | -2.0 | 5.0 |

# Measuring Distance

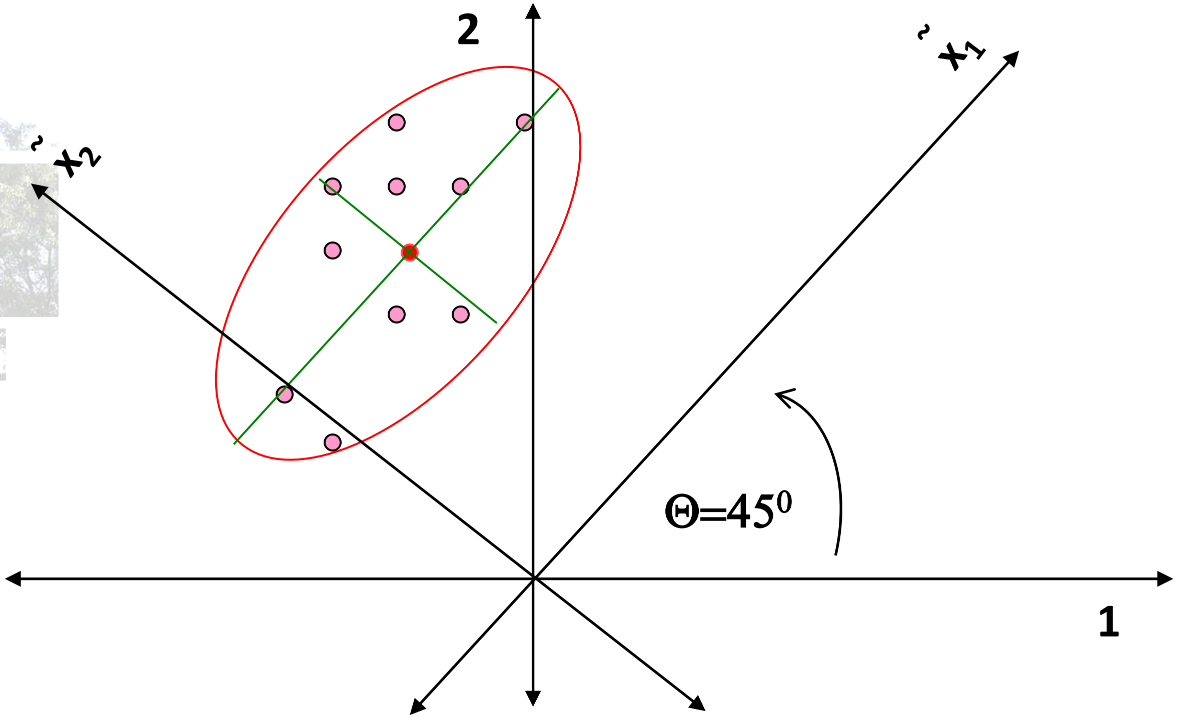
The plot of these points would look like this:




The data suggest a positive correlation between $x_1$, and $x_2$.

# Measuring Distance

The inscribing ellipse (and major and minor axes) look like this:



# Measuring Distance

The rotational weights are:

$$a_{11} = \frac{\cos^2\left(45^0\right)}{\cos^2\left(45^0\right)\left(1.11\right) + 2\sin\left(45^0\right)\cos\left(45^0\right)\left(0.8\right) + \sin^2\left(45^0\right)\left(2.89\right)}$$

$$+ \frac{\sin^2\left(\theta\right)}{\cos^2\left(45^0\right)\left(2.89\right) - 2\sin\left(45^0\right)\cos\left(45^0\right)\left(0.8\right) + \sin^2\left(45^0\right)\left(1.11\right)}$$

$$= 0.6429$$

# Measuring Distance

and:

$$a_{22} = \frac{\sin^2\left(45^0\right)}{\cos^2\left(45^0\right)(1.11) + 2\sin\left(45^0\right)\cos\left(45^0\right)(0.8) + \sin^2\left(45^0\right)(2.89)} + \frac{\cos^2\left(\theta\right)}{\cos^2\left(45^0\right)(2.89) - 2\sin\left(45^0\right)\cos\left(45^0\right)(0.8) + \sin^2\left(45^0\right)(1.11)}$$

$$= 0.6429$$

# Measuring Distance

and:

$$a_{12} = \frac{\sin(45^0)\cos(45^0)}{\cos^2(45^0)(1.11) + 2\sin(45^0)\cos(45^0)(0.8) + \sin^2(45^0)(2.89)}$$

$$= \frac{\cos(\theta)\sin(45^0)}{\cos^2(45^0)(2.89) - 2\sin(45^0)\cos(45^0)(0.8) + \sin^2(45^0)(1.11)}$$

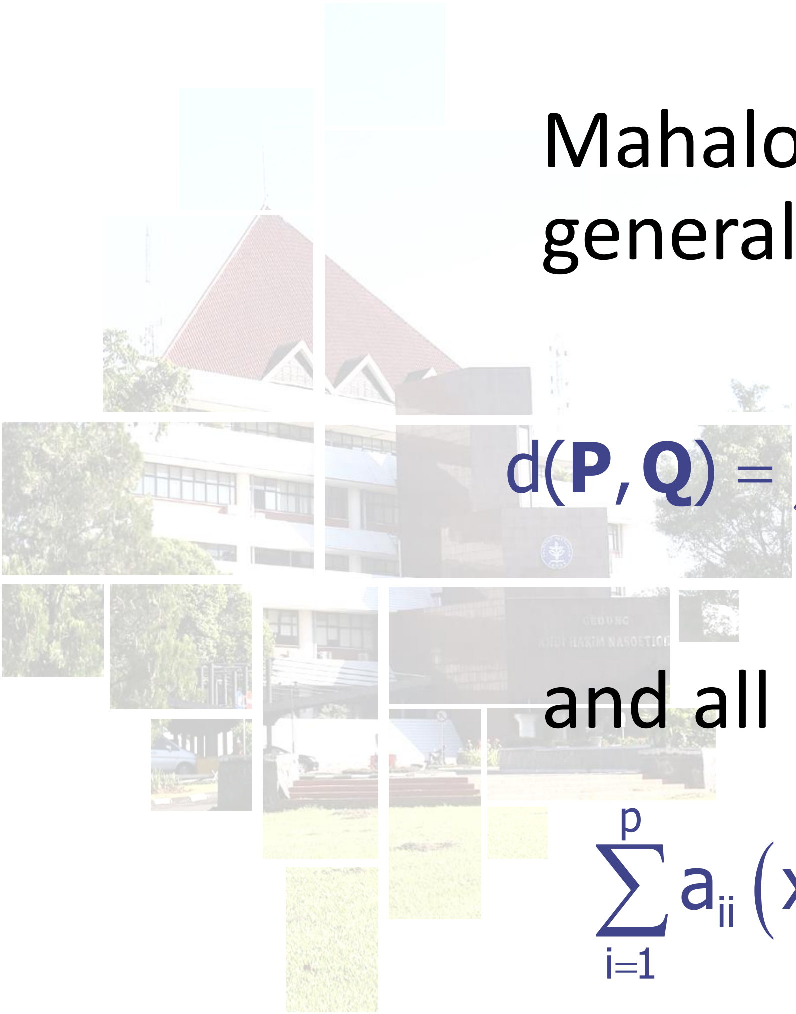
$$= -0.3571$$

# Measuring Distance

So the distances of the observed points from their centroid **Q** = (-2.0, 5.0) are:

| Obs # | $x_1$ | $x_2$ | $\tilde{x}_1$ | $\tilde{x}_2$ | Euclidean D(P,Q) | Mahalanobis D(P,Q) |
|---|---|---|---|---|---|---|
| 1 | -3 | 3 | 0.0000 | 4.2426 | 2.2361 | 1.3363 |
| 2 | -2 | 6 | 2.8284 | 5.6569 | 1.0000 | 0.8018 |
| 3 | -1 | 4 | 2.1213 | 3.5355 | 1.4142 | 1.4142 |
| 4 | -2 | 4 | 1.4142 | 4.2426 | 1.0000 | 0.8018 |
| 5 | -3 | 6 | 2.1213 | 6.3640 | 1.4142 | 1.4142 |
| 6 | -1 | 6 | 3.5355 | 4.9497 | 1.4142 | 0.7559 |
| 7 | -3 | 2 | -0.7071 | 3.5355 | 3.1623 | 2.0702 |
| 8 | -3 | 5 | 1.4142 | 5.6569 | 1.0000 | 0.8018 |
| 9 | 0 | 7 | 4.9497 | 4.9497 | 2.8284 | 1.5119 |
| 10 | -2 | 7 | 3.5355 | 6.3640 | 2.0000 | 1.6036 |
| $x_i$ | -2.0 | 5.0 | | | | |

# Measuring Distance



Mahalonobis distance can easily be generalized to p dimensions:

$$d(\mathbf{P},\mathbf{Q}) = \sqrt{\sum_{i=1}^{p} a_{ii}\left(x_i - y_i\right)^2 + 2\sum_{i=1}^{j-1}\sum_{j=2}^{p} a_{ij}\left(x_i - y_i\right)\left(x_j - y_j\right)}$$

and all points satisfying

$$\sum_{i=1}^{p} a_{ii}\left(x_i - y_i\right)^2 + 2\sum_{i=1}^{j-1}\sum_{j=2}^{p} a_{ij}\left(x_i - y_i\right)\left(x_j - y_j\right) = c^2$$

form a hyperellipsoid with centroid **Q**.

# IPB University

— Bogor Indonesia —

Inspiring Innovation with Integrity

in Agriculture, Ocean and Biosciences for a Sustainable World