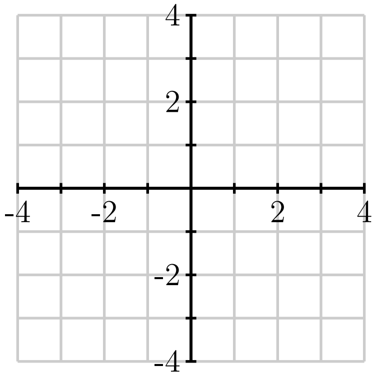
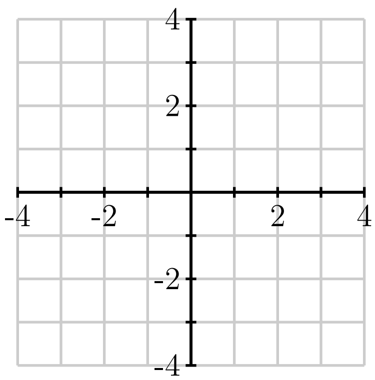
Mathematics 327
Variance and covariance

Let's begin with three data points $\mathbf{x}_1 = (1,1)$, $\mathbf{x}_2 = (2,1)$, and $\mathbf{x}_3 = (3,4)$ where each point $\mathbf{x}_j = (x_j, y_j)$.

Find the mean $\overline{\mathbf{x}}$. Then plot the points and their mean.



Now construct the de-meaned data where $\widetilde{\mathbf{x}}_j = \mathbf{x}_j - \overline{\mathbf{x}}$ and plot these data points.

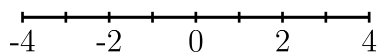


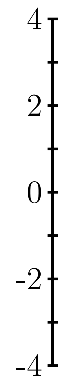
The *total variance* of a data set is the average of the length squared of the de-meaned data; that is, the total variance is

$$V = \frac{1}{N}\sum_j |\widetilde{\mathbf{x}}|^2 .$$

Find the total variance of our data set with three points. Which point makes the greatest contribution to the total variance?

Now project the data onto the x- and y-axes.





We can define variances

$$V_{xx} = \frac{1}{N}\sum_j \tilde{x}_j^2, \qquad V_{yy} = \frac{1}{N}\sum_j \tilde{y}_j^2.$$

Find both variances V_{xx} and V_{yy}. Determine which is larger and explain why this is so.

What is the relationship between V_{xx}, V_{yy} and V? Explain why this relationship holds.

Define the covariance $V_{xy} = \frac{1}{N}\sum_j \tilde{x}_j\tilde{y}_j$ and evaluate it for our data set.

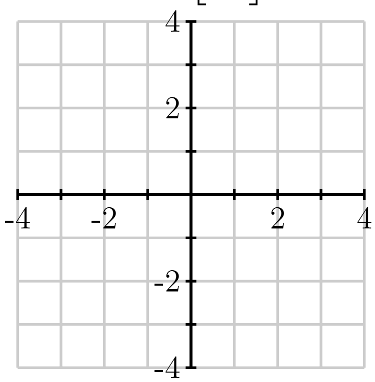
We can form the data matrix A and the covariance matrix C as

$$A = \begin{bmatrix} \widetilde{\mathbf{x}}_1 & \widetilde{\mathbf{x}}_2 & \dots & \widetilde{\mathbf{x}}_N \end{bmatrix}, \qquad C = \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix}$$

Write these matrices for our specific example.

Now verify that $C = \frac{1}{N}AA^T$ and explain why this relationship should hold.

Just as we found the variance after projecting along the x- and y-axes, we can do this for any other direction. For example, project the de-meaned data points onto the line defined by $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and find the variance $V_{\mathbf{v}_1}$.



Also, project the de-meaned data onto the line defined by $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and find the variance $V_{\mathbf{v}_2}$.

What is the relationship between $V_{\mathbf{v}_1}$, $V_{\mathbf{v}_2}$, and the total variance V, and explain why this relationship holds.

Suppose that $\mathbf{u}$ is a unit vector, which we think of as defining a direction in the plane. Call $\widetilde{\mathbf{x}}_{\mathbf{u}}$ the projection of $\widetilde{\mathbf{x}}$ onto the line defined by $\mathbf{u}$. We define the variance in the $\mathbf{u}$ direction to be

$$V_{\mathbf{u}} = \frac{1}{N}\sum_j |(\widetilde{\mathbf{x}}_j)_{\mathbf{u}}|^2 .$$

Explain why $\widetilde{\mathbf{x}}_{\mathbf{u}} = (\widetilde{\mathbf{x}} \cdot \mathbf{u})\mathbf{u}$ and hence

$$V_{\mathbf{u}} = \frac{1}{N} \sum_j (\widetilde{\mathbf{x}}_j \cdot \mathbf{u})^2 = \frac{1}{N} \sum_j (\widetilde{\mathbf{x}}_j^T \mathbf{u})^2.$$

Explain why

$$V_{\mathbf{u}} = \frac{1}{N} \left|A^T \mathbf{u}\right|^2 = \frac{1}{N}(A\mathbf{u}) \cdot (A\mathbf{u}) = \frac{1}{N}(A\mathbf{u})^T (A\mathbf{u}) = \mathbf{u}^T \left(\frac{1}{N} A^T A\right) \mathbf{u} = \mathbf{u}^T C \mathbf{u}.$$

In other words, $V_{\mathbf{u}}$ is given by the quadratic form defined by the covariance matrix C. Find the maximum variance $V_{\mathbf{u}}$ and its direction.

Find the minimum variance $V_{\mathbf{u}}$ and its direction.