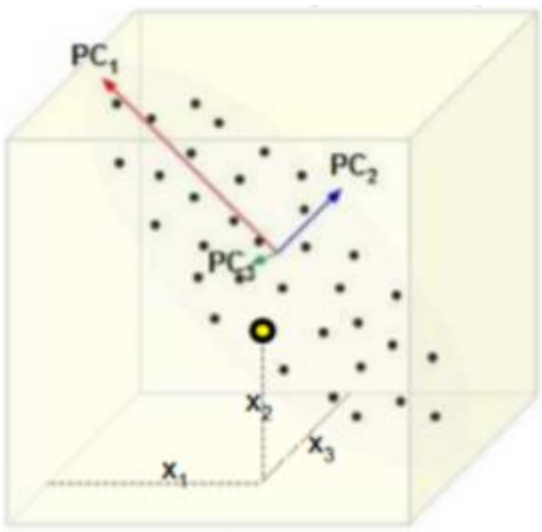
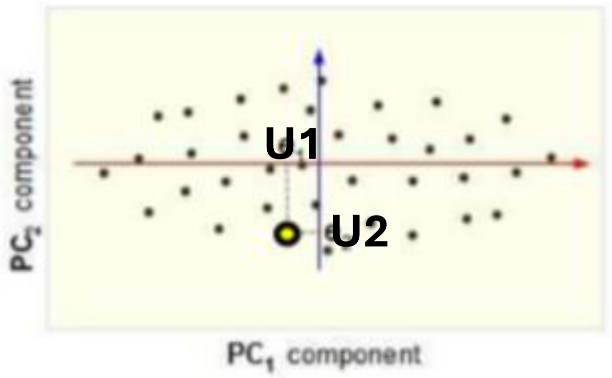
Principal Component Analysis (PCA)

Coordinates in the New System

- Assume that **after centering** the data points, the yellow point in the figure has coordinate $X_1 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ (left figure).
- The new dimension is denoted by $Y_1 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ (right figure).
- Assume that the first two principal components are vectors U_1 and U_2 (red and blue arrows).
- To find y_1 we need to project vector X_1 onto U_1. (What's the dimension of U_1?)
- $y_1 = P_{U_1}^{X_1} = U_1.X_1 = U_1^T X_1$
- Similarly: $y_2 = P_{U_2}^{X_1} = U_2.X_1 = U_2^T X_1$
- Let $P = \begin{bmatrix} \uparrow & \uparrow \\ U_1 & U_2 \\ \downarrow & \downarrow \end{bmatrix}$ where P is a 3×2 matrix.
- $\color{red}{\boldsymbol{Y_1}} = \begin{bmatrix} U_1^T X_1 \\ U_2^T X_1 \end{bmatrix} = \begin{bmatrix} \leftarrow & U_1^T & \rightarrow \\ \leftarrow & U_2^T & \rightarrow \end{bmatrix} X_1 = \color{red}{\boldsymbol{P^T X_1}}$





Projection of All Data Points in Matrix Form

- Assume that all data points are in the **mean deviation** form:
- $X = [X_1 \quad \ldots \quad X_N]$ where X is a $d \times N$ matrix. (d is the dimension of the original data).
- Assume that P denotes the first m principal components:

$$P = [U_1 \quad \ldots \quad U_m]$$

- The new set of coordinates (also called **scores**) can be found from:

$$\boldsymbol{Y = P^T X}$$

- What's the dimension of each element in the above expression?
  - X is $\mathrm{d} \times N$
  - Y is $m \times N$
  - P is $\mathrm{d} \times m$; P^T is $\mathrm{m} \times \mathrm{d}$

PCA: A Linear Combination

- Revisiting the expression for new coordinates, let $U_1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $U_2 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$
- Recall that $y_1 = P_{U_1}^{X_1} = U_1 . X_1 = U_1^T X_1$ and $y_2 = P_{U_2}^{X_1} = U_2 . X_1 = U_2^T X_1$
- We can write $y_1 = a_1 x_1 + a_2 x_2 + a_3 x_3$ and $y_2 = c_1 x_1 + c_2 x_2 + c_3 x_3$.
- Therefore, we can view the new dimension y as a linear combination of original coordinates where the weights are given by elements in the eigenvectors U_1 and U_2.

Example

- Given the following data set, find PCA and transform the data points. Find the total variance (sum of variances in each dimension) in the old and new data sets.

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \mathbf{X}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

Solution in Python

```python
X = np.array([[1,4,7,8],[2,2,8,4],[1,13,1,5]])
X
```

```
array([[ 1,  4,  7,  8],
       [ 2,  2,  8,  4],
       [ 1, 13,  1,  5]])
```

```python
m = X.mean(axis=1)
print(m)
B = X - m.reshape(3,1)
B
```

```
[5. 4. 5.]
```

```
array([[-4., -1.,  2.,  3.],
       [-2., -2.,  4.,  0.],
       [-4.,  8., -4.,  0.]])
```

```python
Sx = B@B.T/(X.shape[1]-1)
Sx
```

```
array([[10.,  6.,  0.],
       [ 6.,  8., -8.],
       [ 0., -8., 32.]])
```

```python
from numpy import linalg
l,P = linalg.eig(Sx)
```

```python
Y = P.T@B
Y
```

```
array([[-0.15412474,  0.65266291,  1.20729192, -1.7058301 ],
       [ 5.25133768,  0.0191478 , -2.81268299, -2.45780249],
       [-2.89822328,  8.28092172, -5.16054848, -0.22214996]])
```