Lecture 9: Total Least Square and Principal Component Analysis

Road Map of the Statistics Part

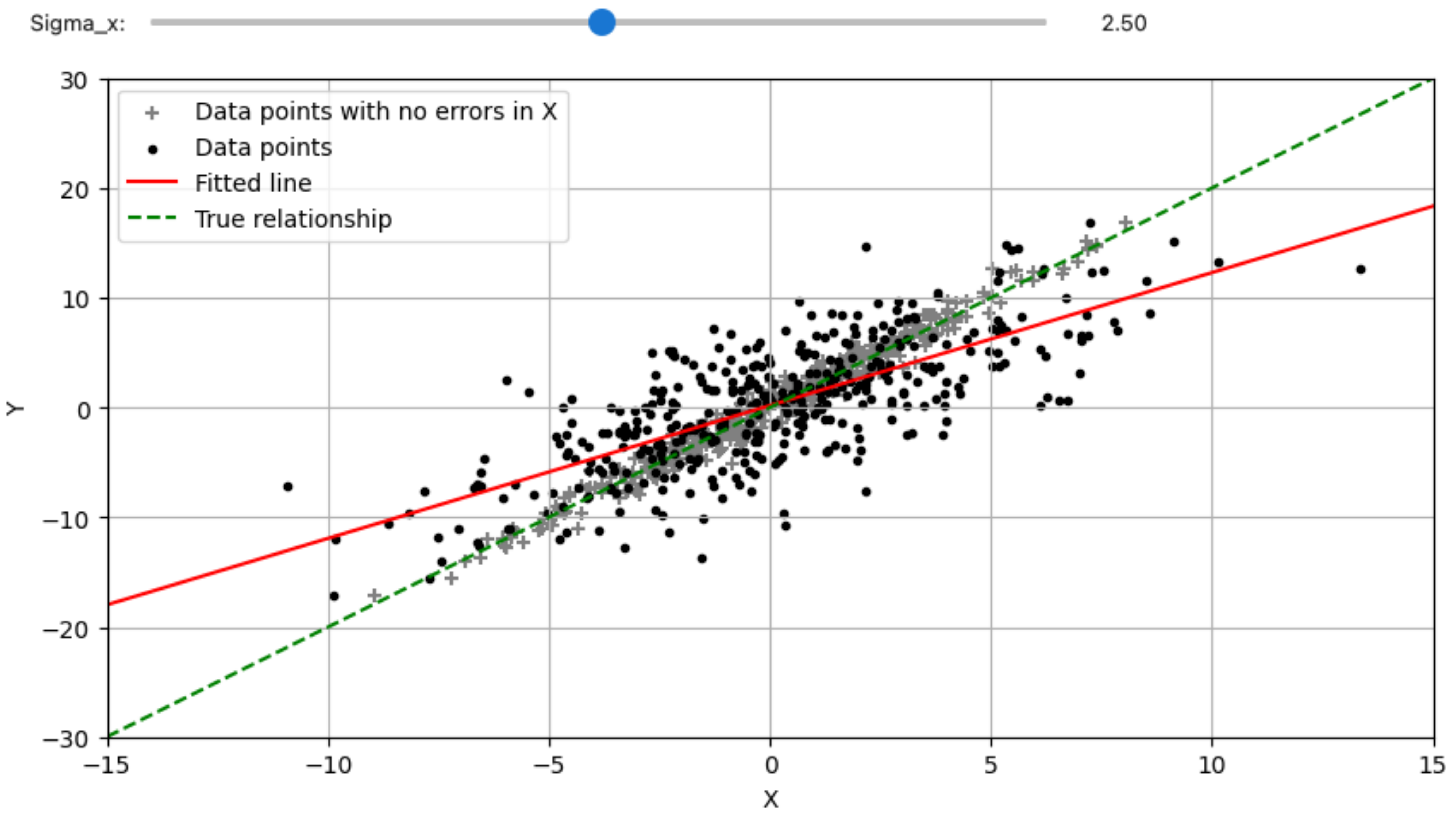
	Lecture 5	Lecture 6	Lecture 7	Lecture 8
Quantification Technique	Mean, variance, skewness, & kurtosis	Pearson's Correlation (Linear relationship)	Linear regression (OLS)	Model Selection
Uncertainty & Significance	Gaussian distribution Chi-2 distribution	r, p = scipy.stats.pearsonr(x, y)	results.summary( )	Training error vs. prediction error
Assumptions	Data is Gaussian or follows specific types of distribution Independent Sampling	Data is Gaussian Independent Sampling	x is noise free Error is Gaussian Independent Sampling Equal err variance	
Test assumptions	K-S test		Auto-correlation (Effective Sample Size)	
Treatment		Bootstrapping	Block Bootstrapping	

1. **Total least square** for mitigating regression dilution

2. Going towards higher dimensions - **Principal Component Analysis**

Errors in Predictor and Regression Dilution

Regression dilution: Underestimate of OLS slope when x contain errors.

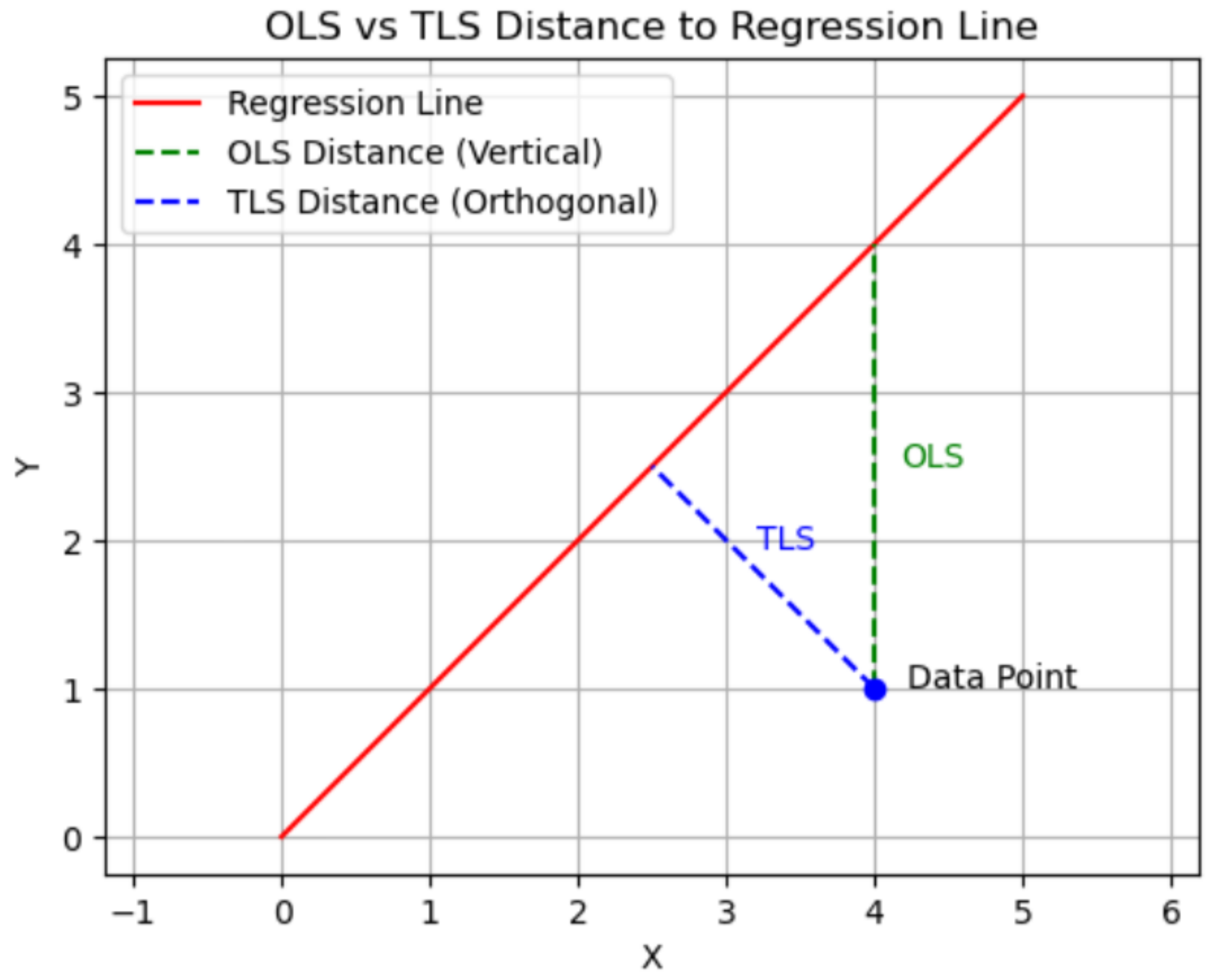


Accounting for regression dilution: Total Least Square

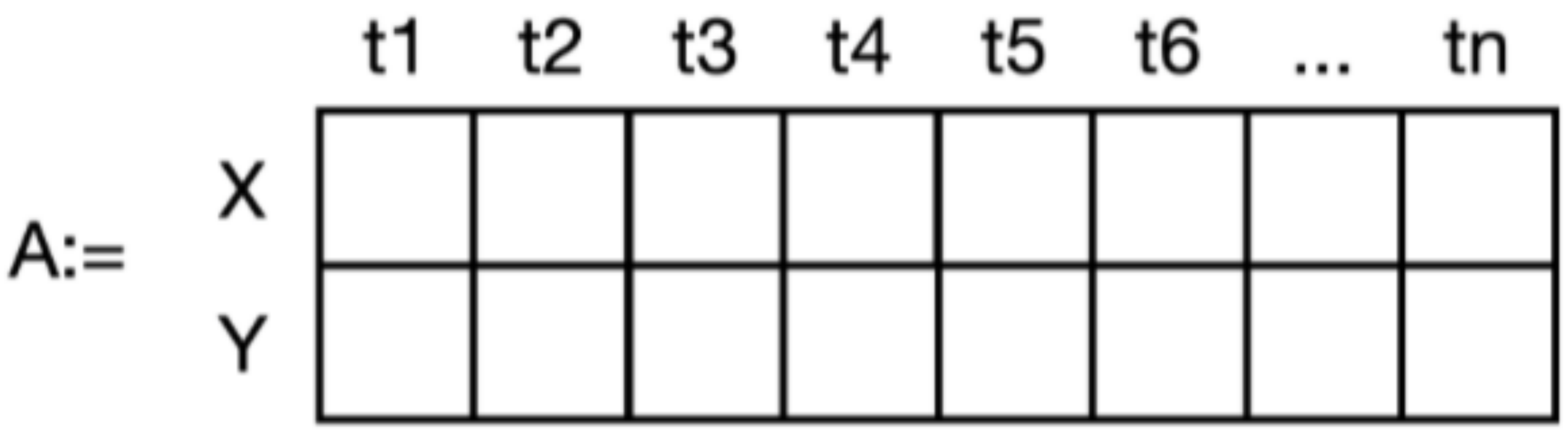
When **X and Y have the same unit.**

Total Least Square mitigates regression dilution by minimising the distance to the fitted line.

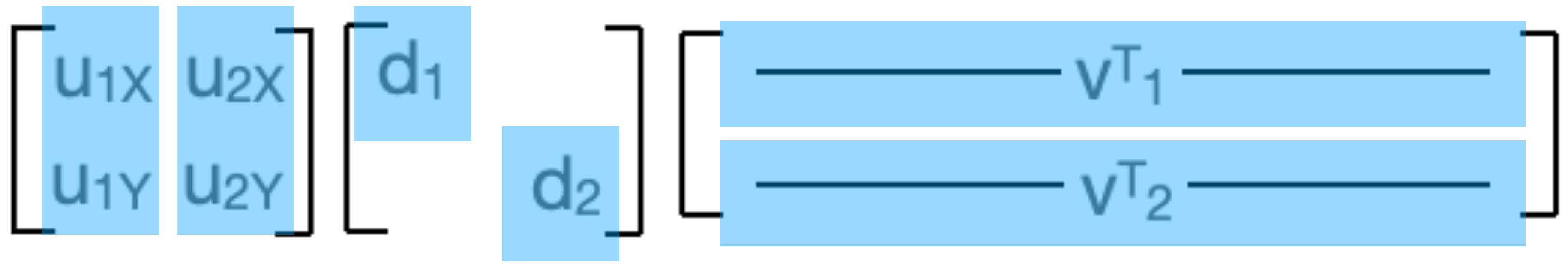
There is no package for calculating TLS directly in python, but we can use the SVD function to calculate TLS alternatively.



Compare with Ordinary Least Squares

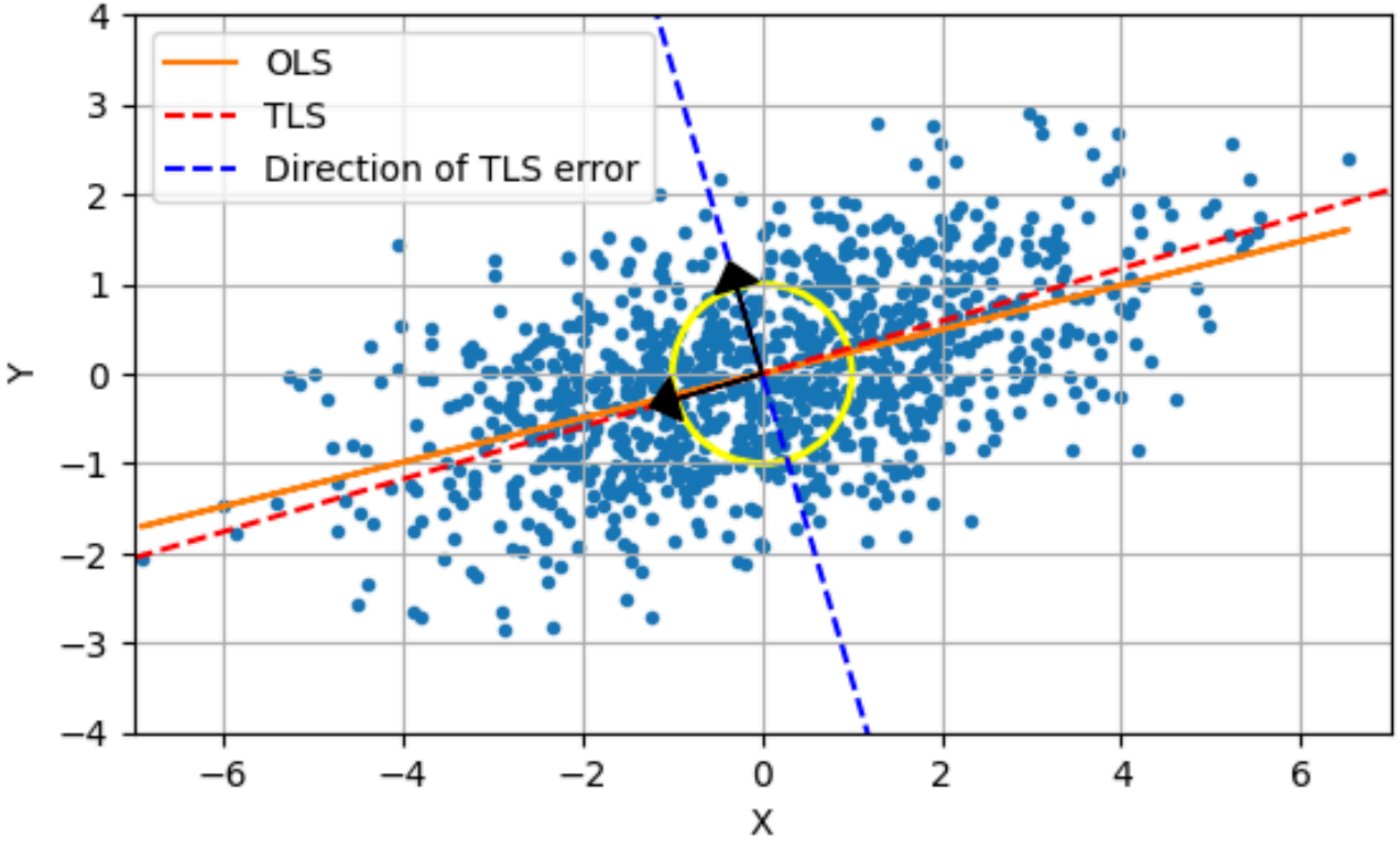



```python
import numpy as np
U, D, VT = np.linalg.svd(A)
```



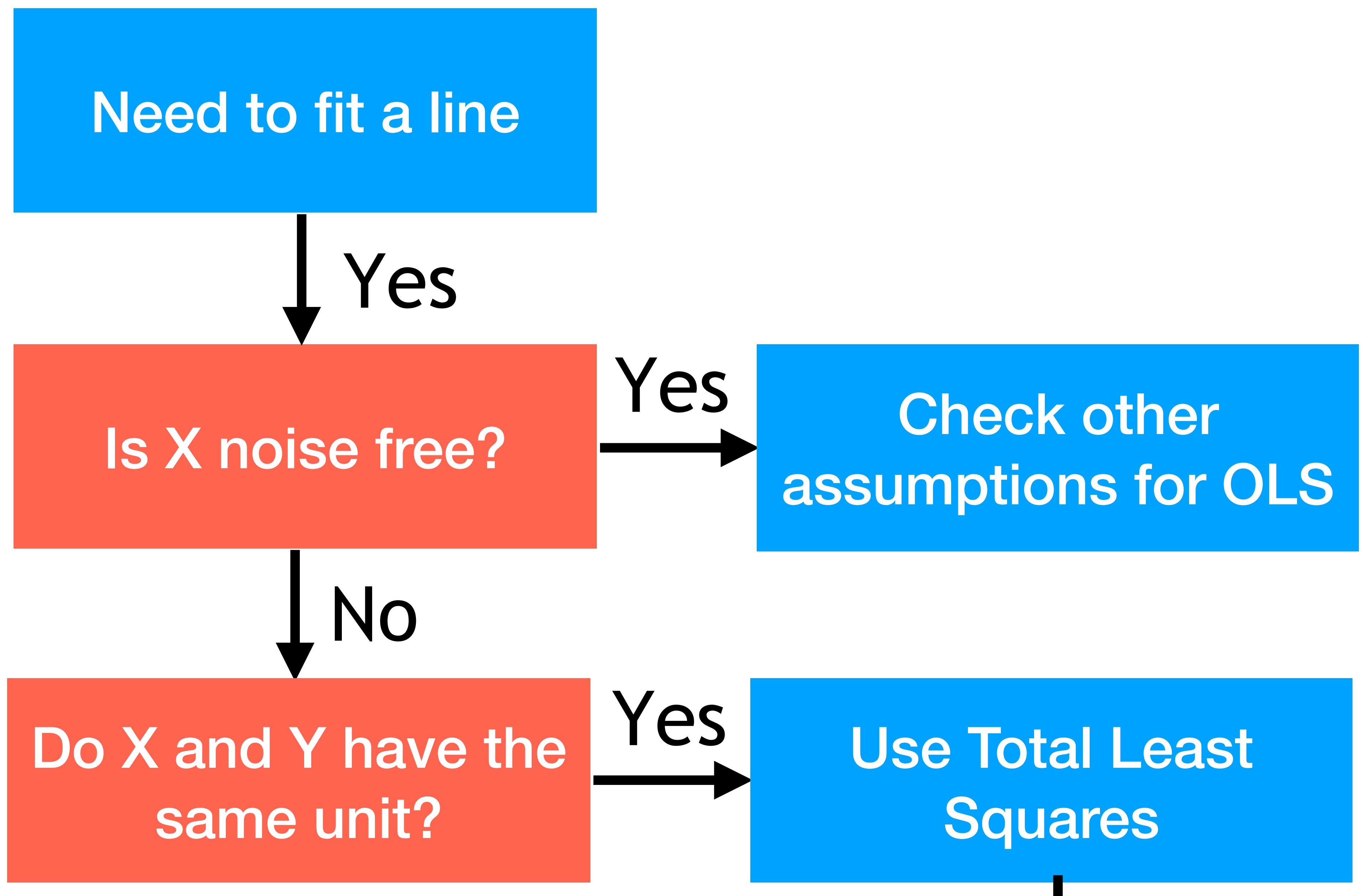
direction of the new coordinate

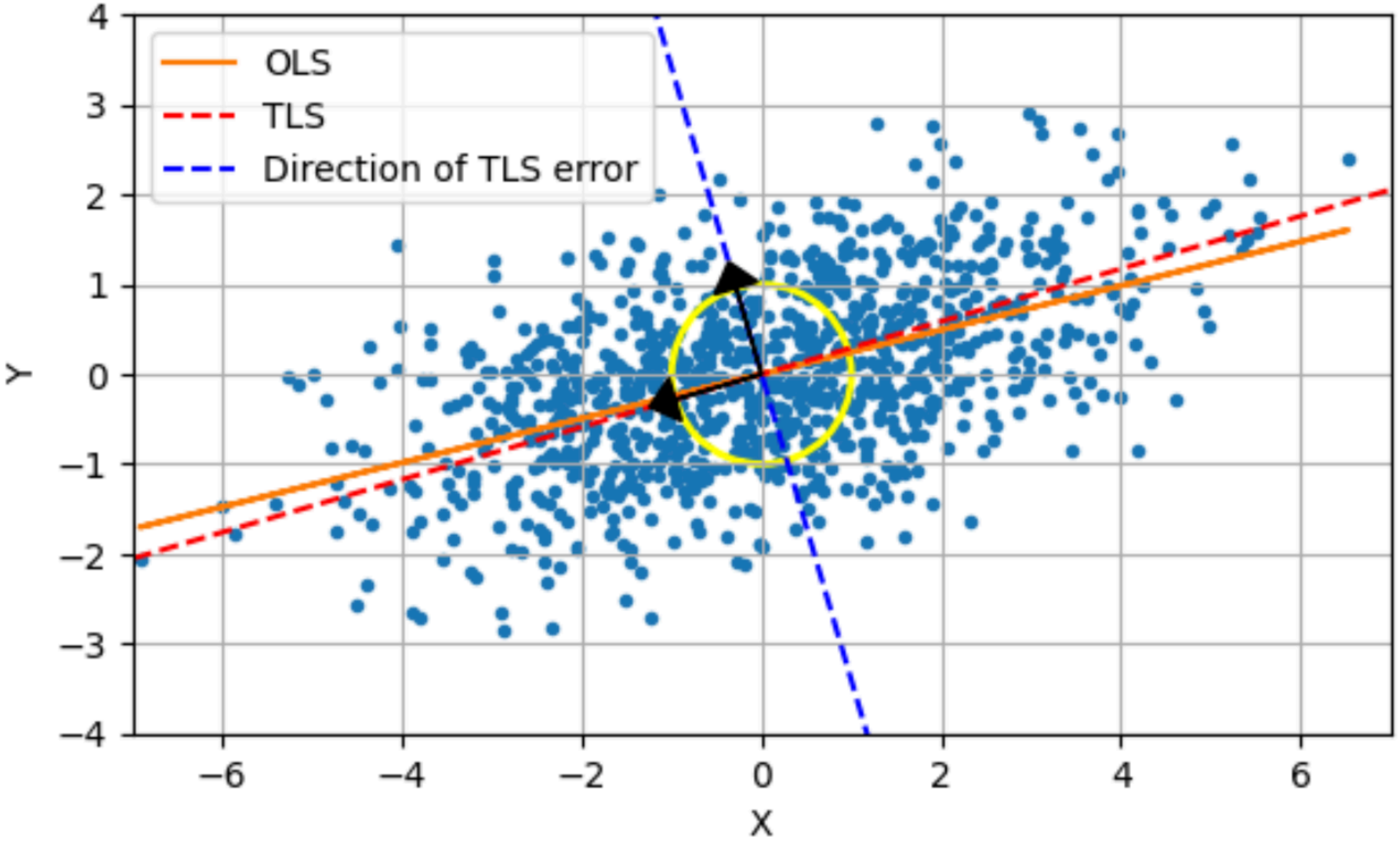
slope = u_{1Y}/u_{1X}

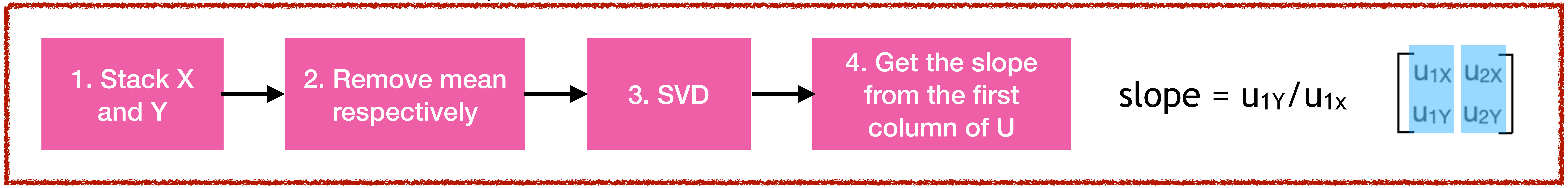


Regression dilution is mitigated!

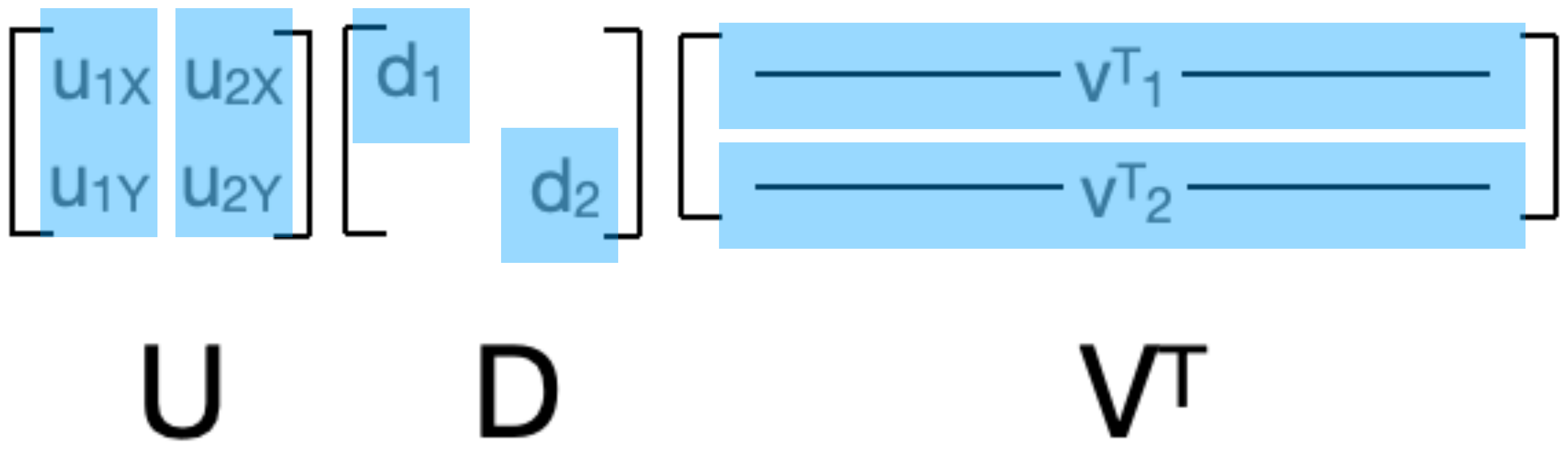
Summary for TLS using SVD



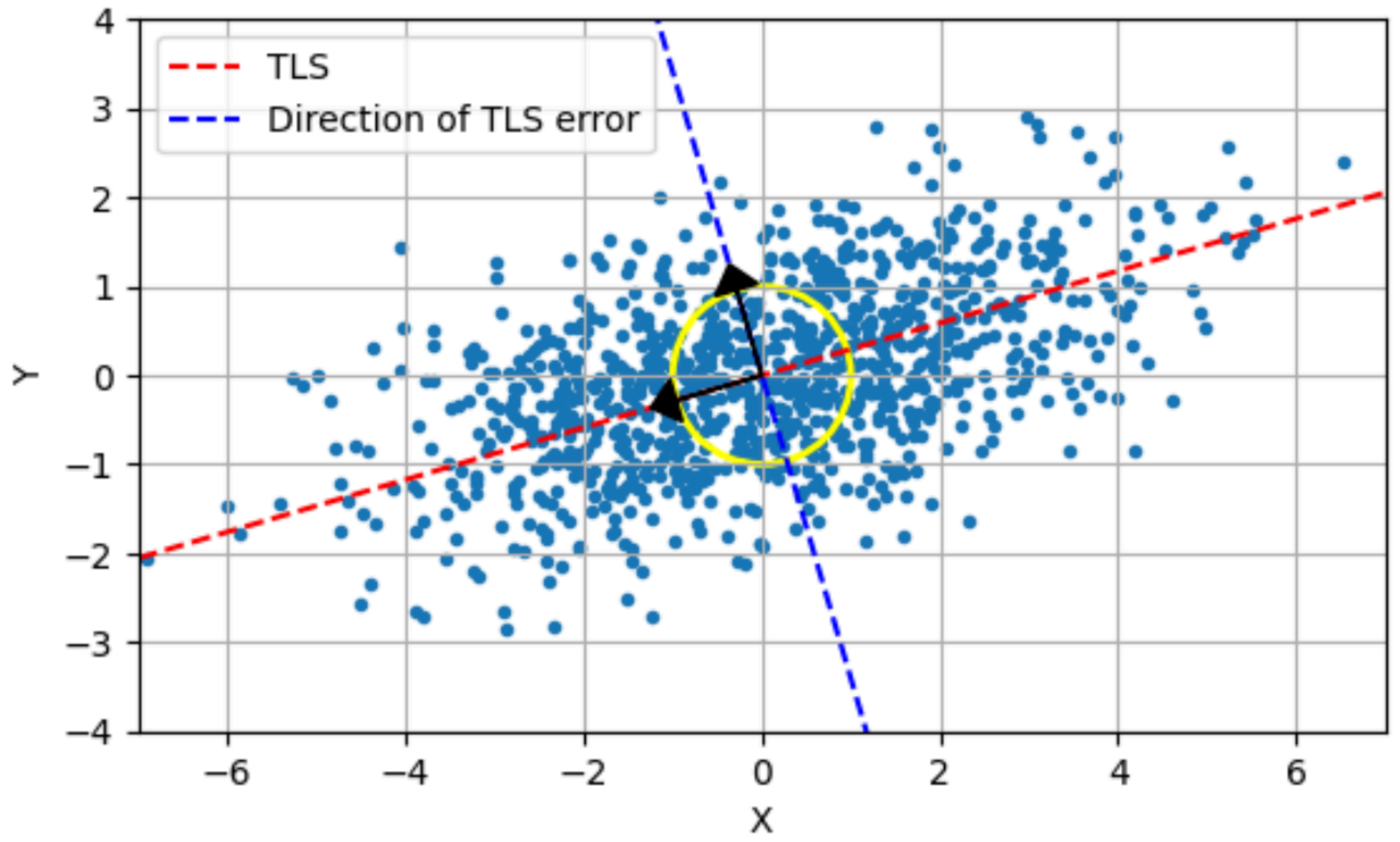




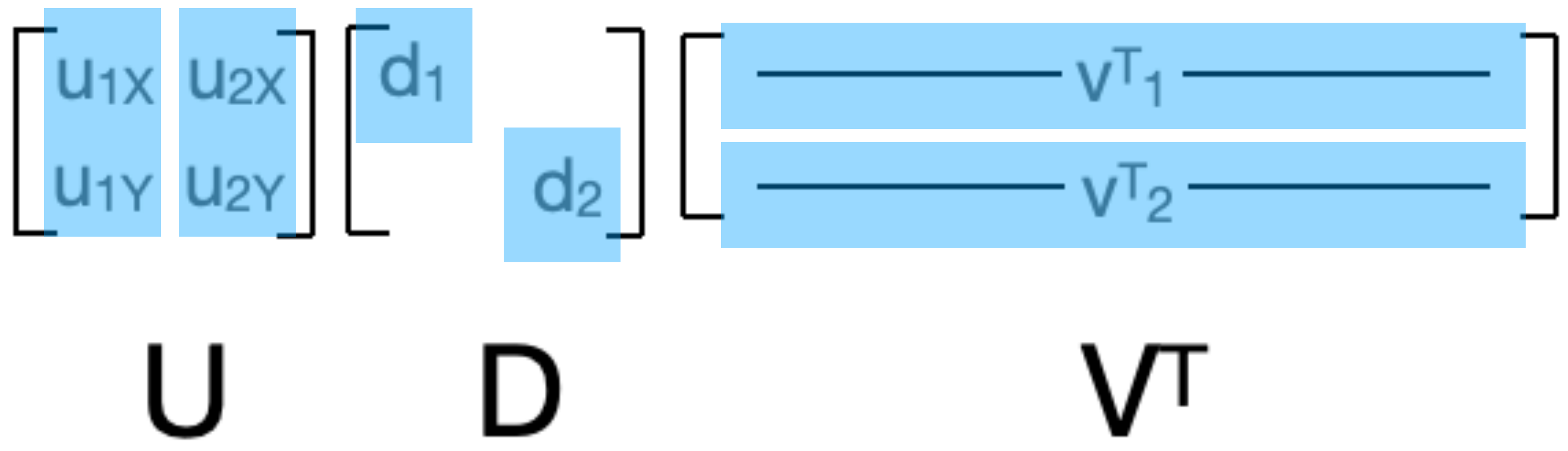
Understanding D and V



direction of the
new coordinate



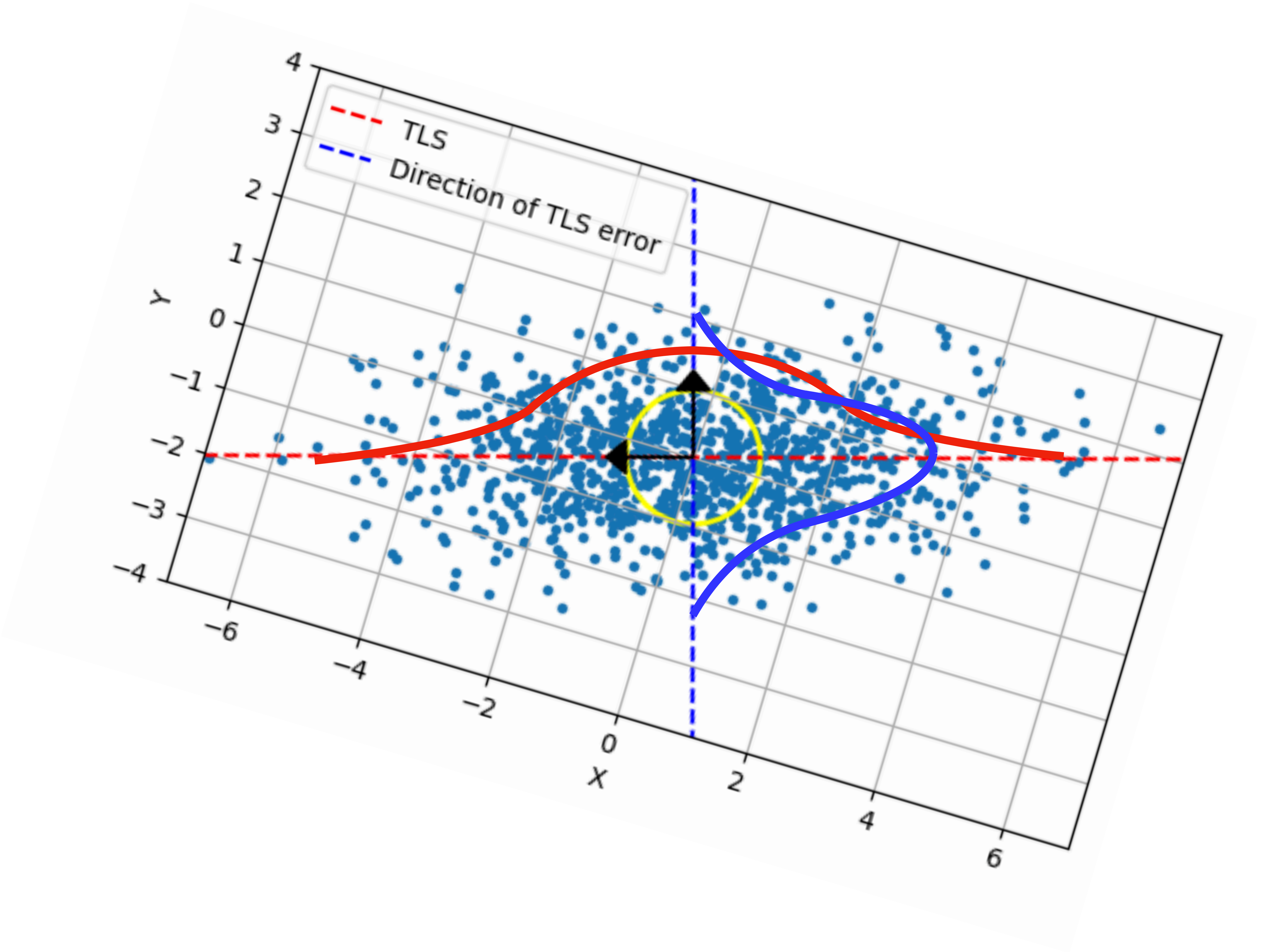
Understanding D and V



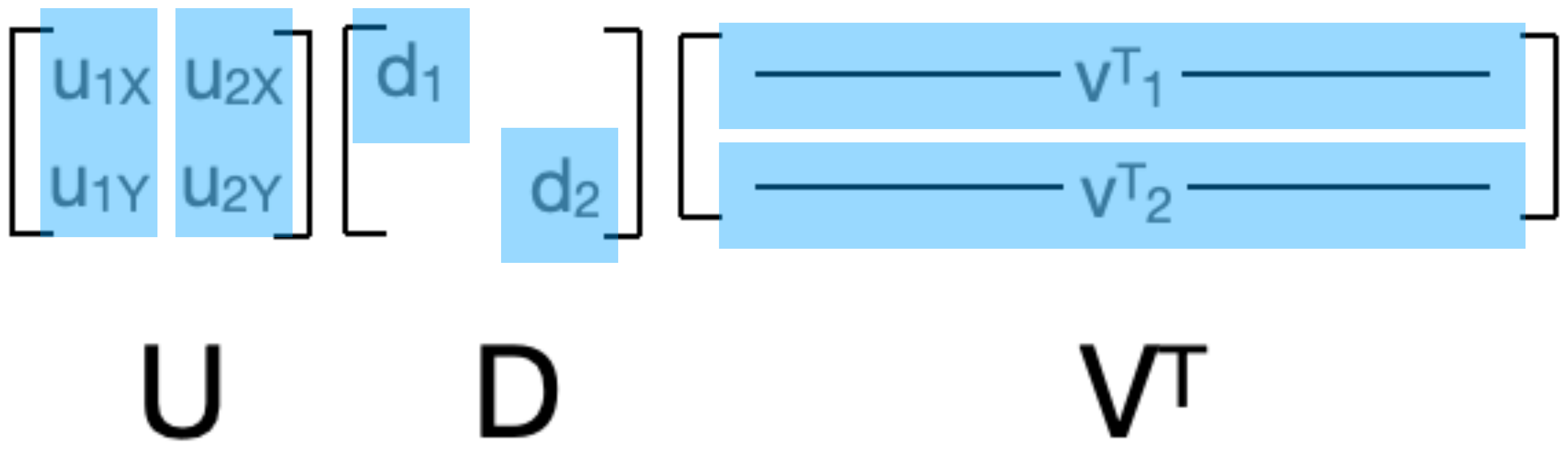
direction of the new coordinate

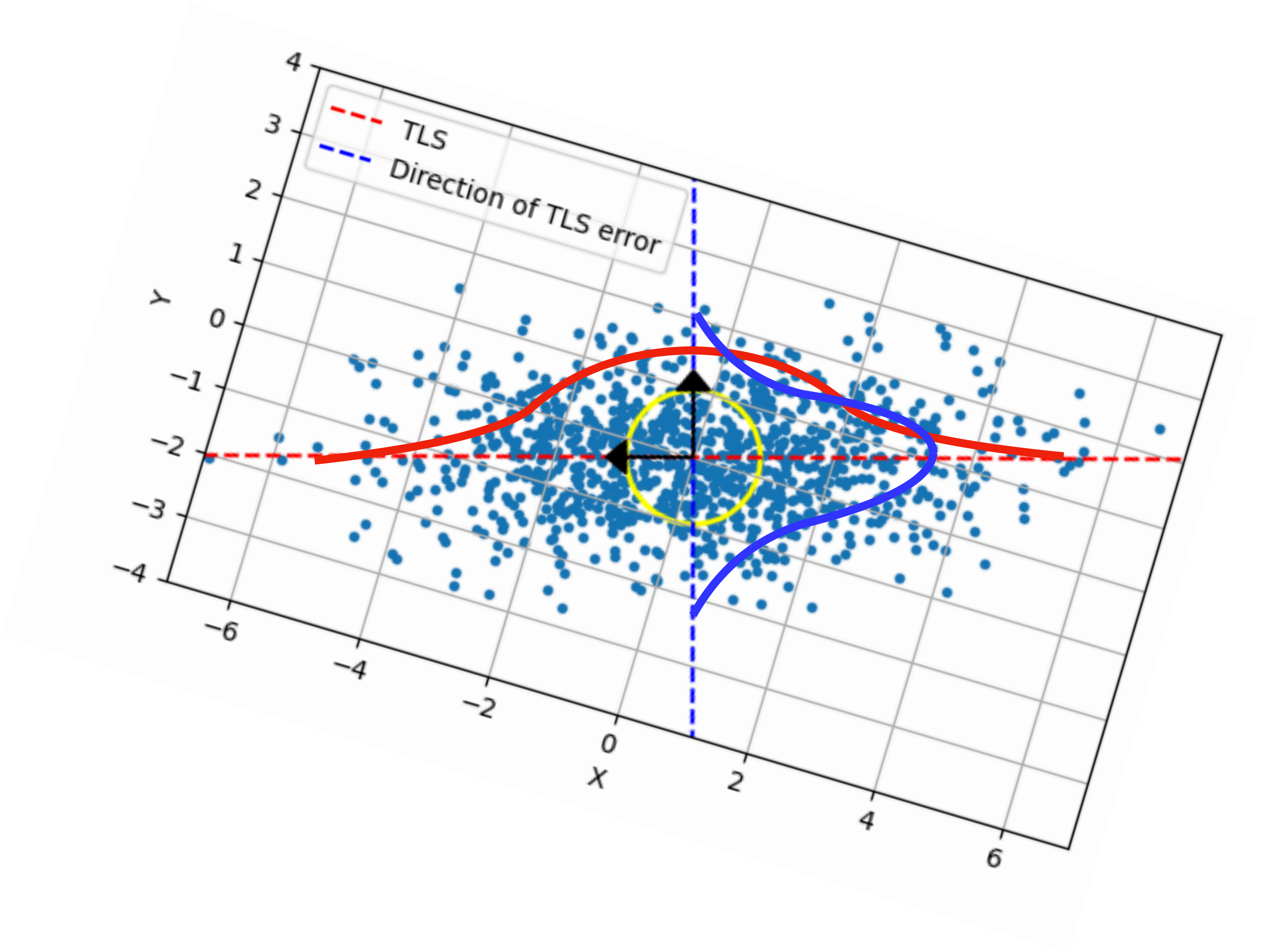
standard deviation

standardised location in the new coordinate



Important properties of SVD





(1) Individual columns of **U** are **orthogonal**.

The new directions are perpendicular to each other.

(2) Individual columns of **V** are **orthogonal**.

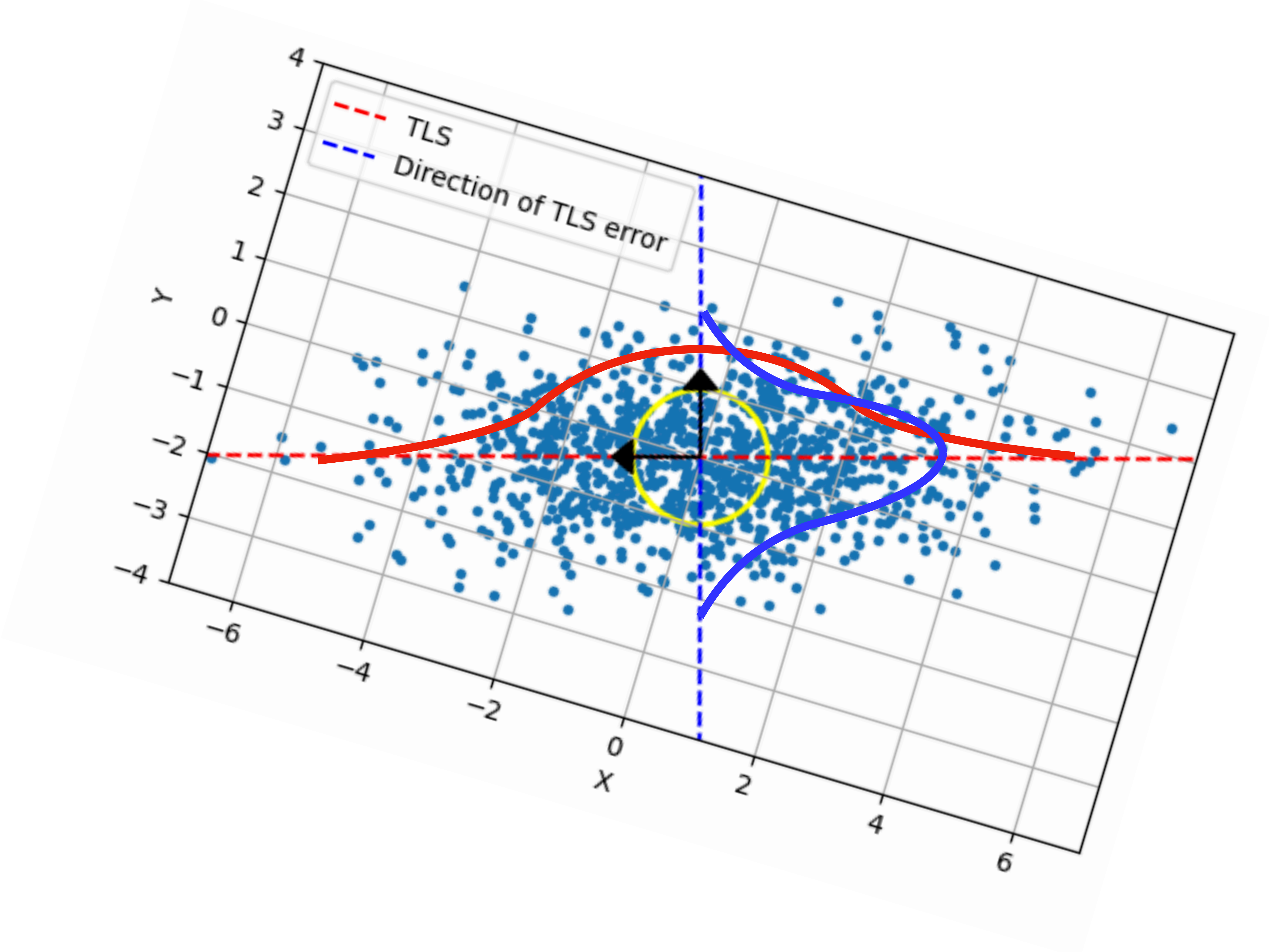
Pearson's correlations of locations in the new coordinate is zero.

(3) D_i is ranked in a **descending order**.

Orthogonal = Perpendicular = Pearson's Correlation is zero

SVD is an effective tool to find major modes of variations for exploring data

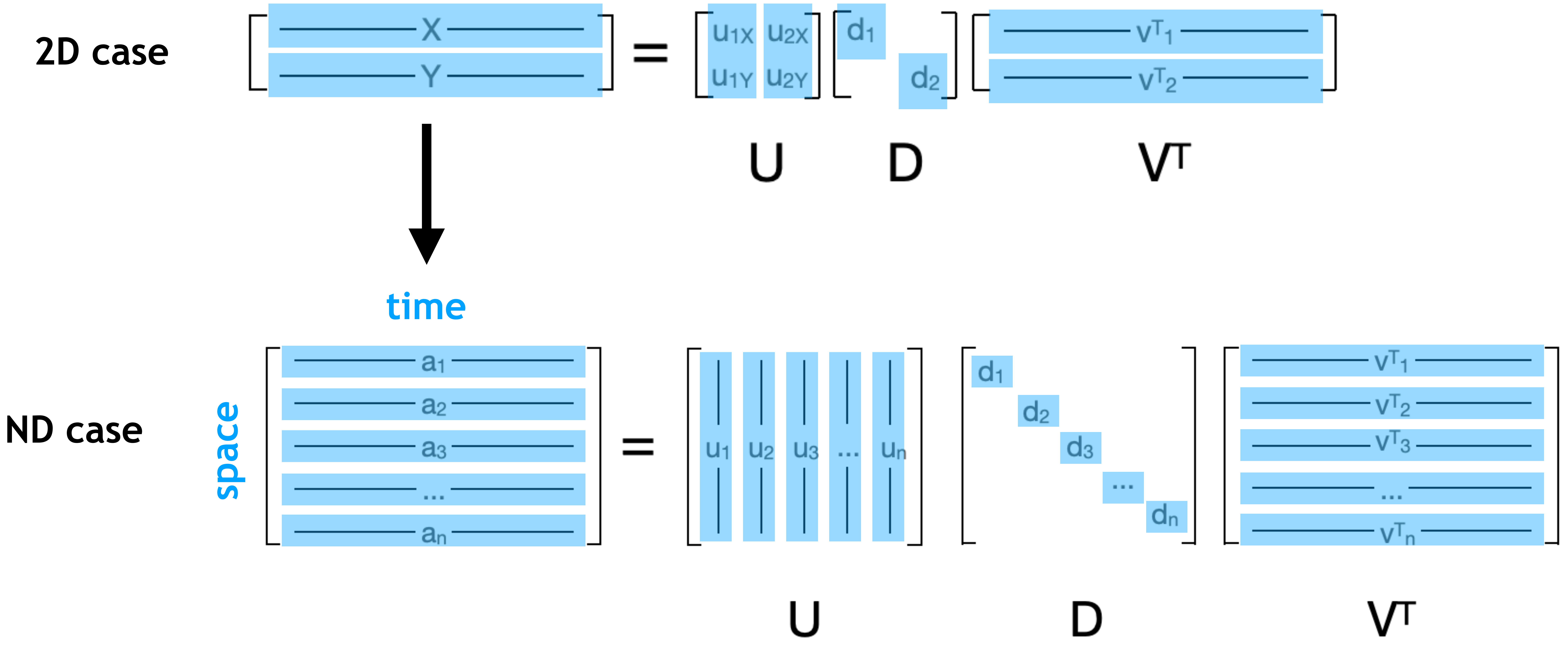
When SVD is applied to high dimensional data, it is often called **Principal Component Analysis (PCA)** or **Empirical Orthogonal Functions (EOF).**



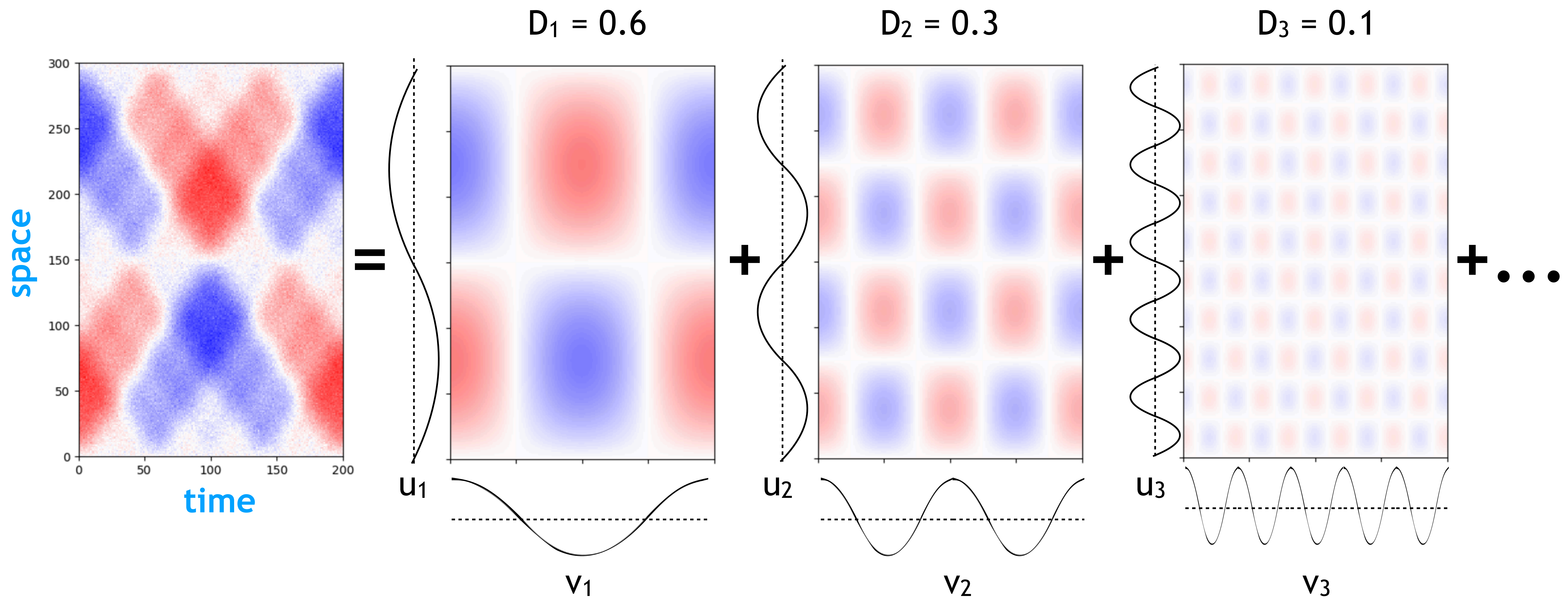
In the TOP3 methods used methods to reveal modes of variability in ocean, earth, and climate data!

Organise High-dimensional Data

```
U, D, VT = np.linalg.svd(A)
```



An example of PCA/EOF using Synthetic Data

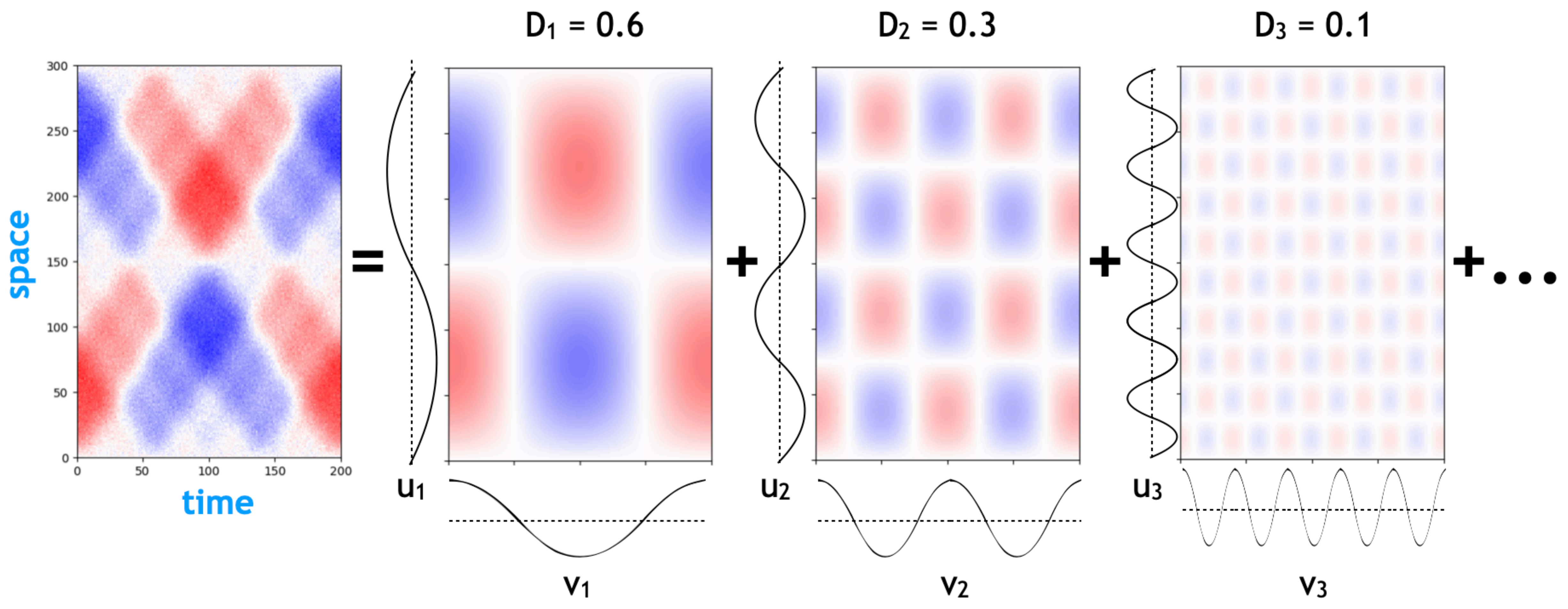


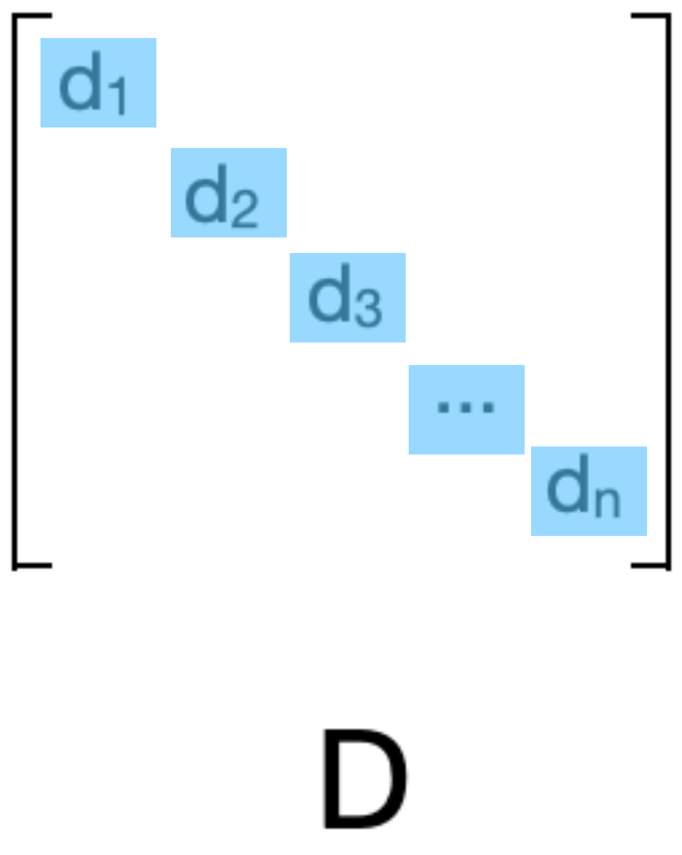
(1) Time series (v) must have zero correlations to each other.

(2) Modes are ranked in a descending order.

(3) Both U, V, and D are empirical from data rather than pre-defined.

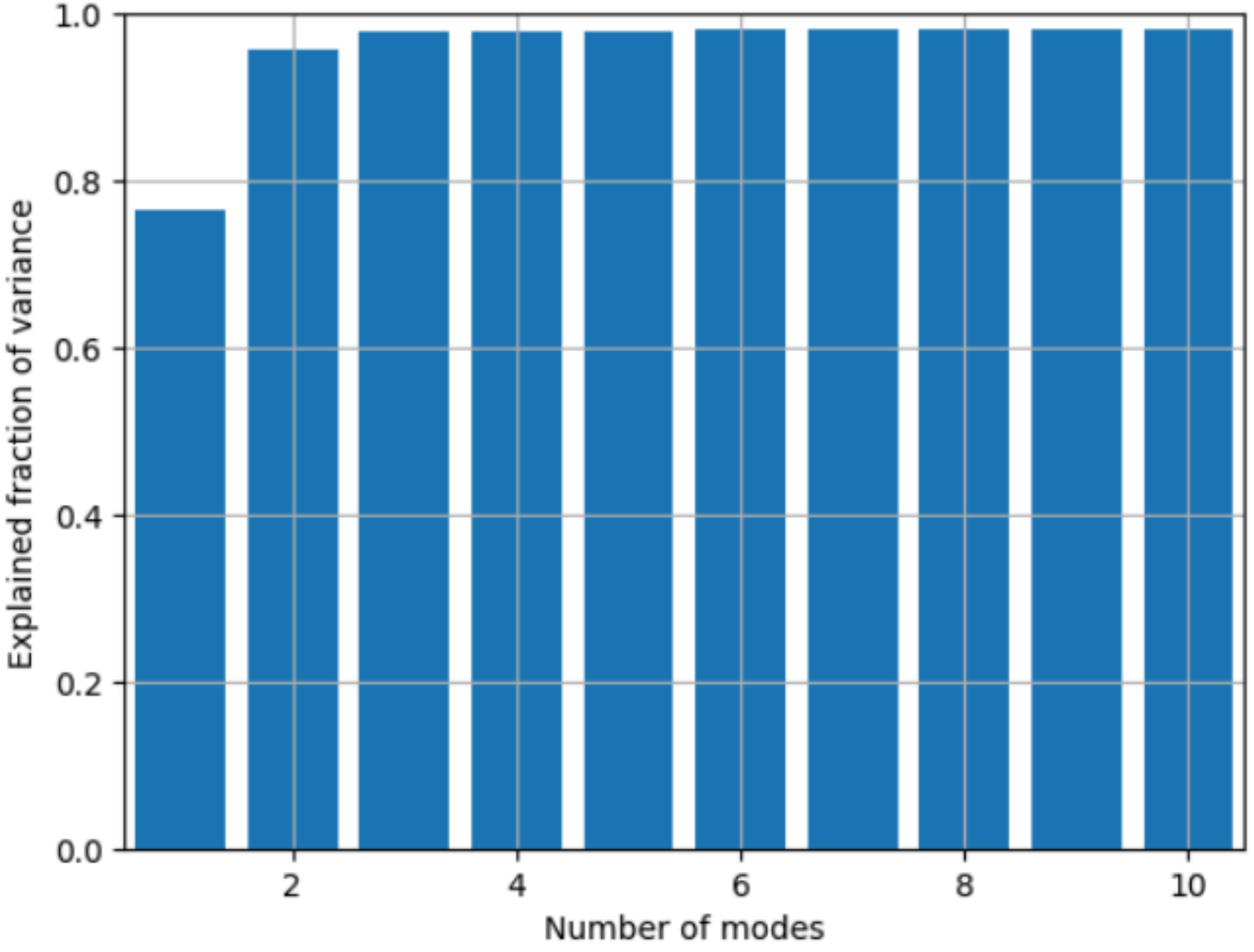
Explained Variance



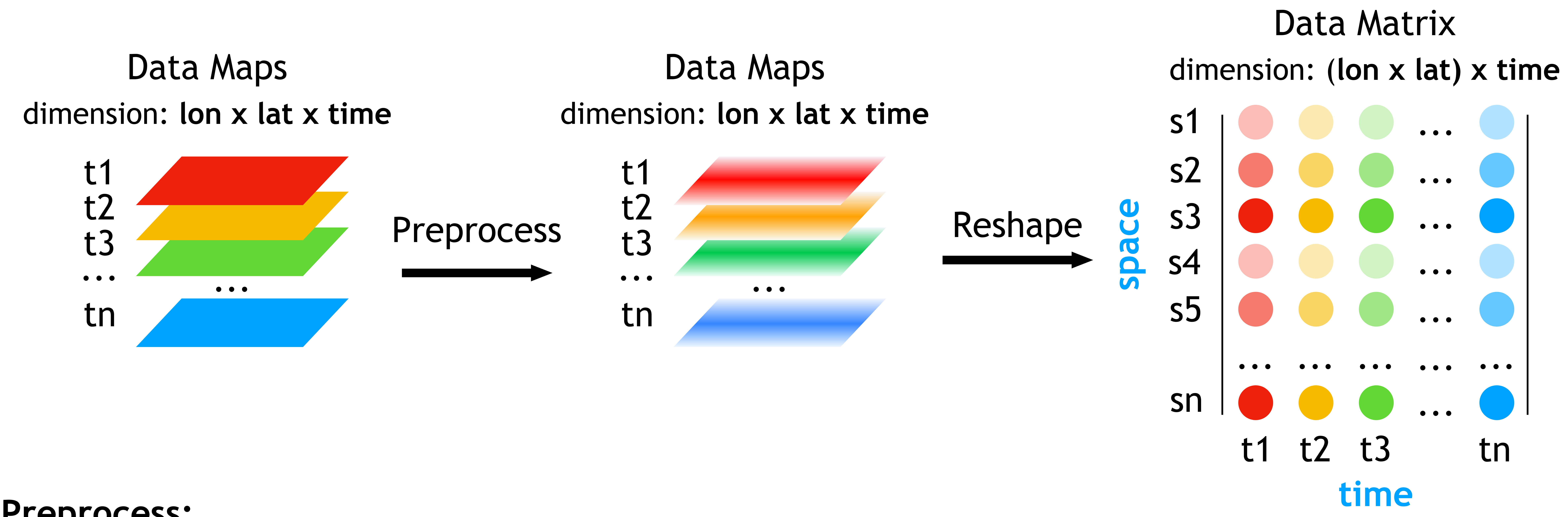


$$FC(s) = \frac{\sum_{i=1}^{s} d_i^2}{\sum_{i=1}^{n} d_i^2}$$

Usually, the cut off is 50% - 90% depending on your problem.



Practical Notes on PCA/EOF analysis for ocean, earth, and climate data

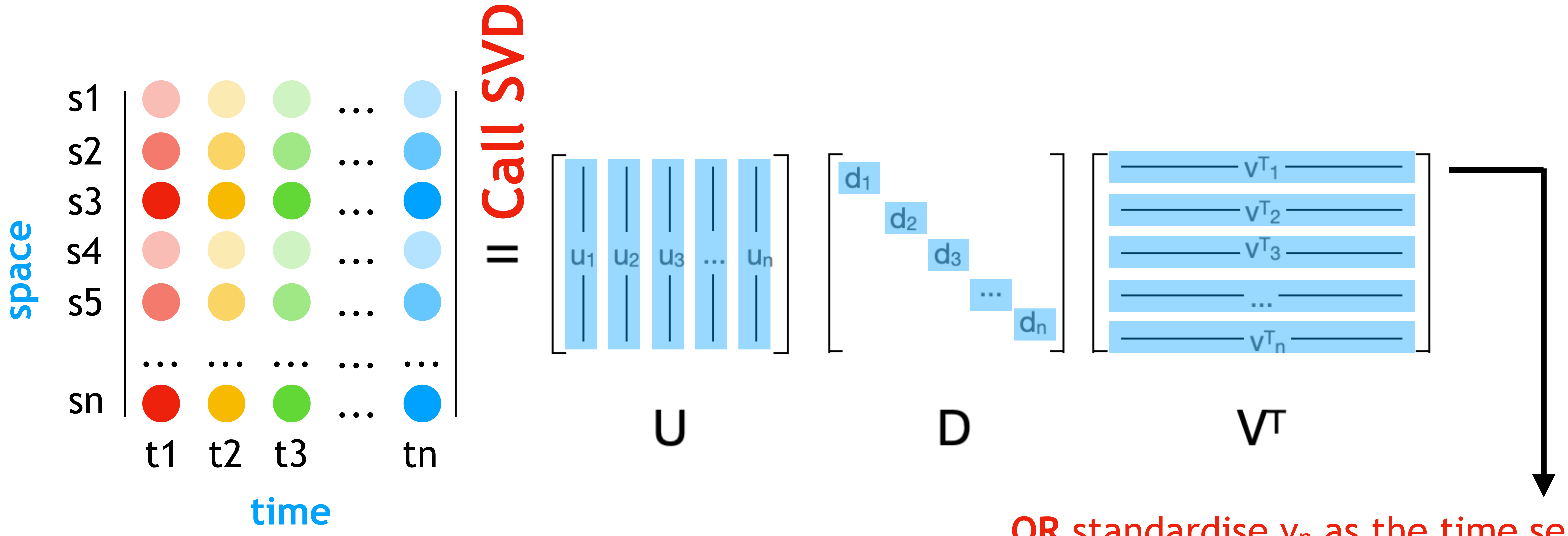


Preprocess:

(1) We often **remove seasonal cycle** (and sometimes long-term trends) before EOF analysis.

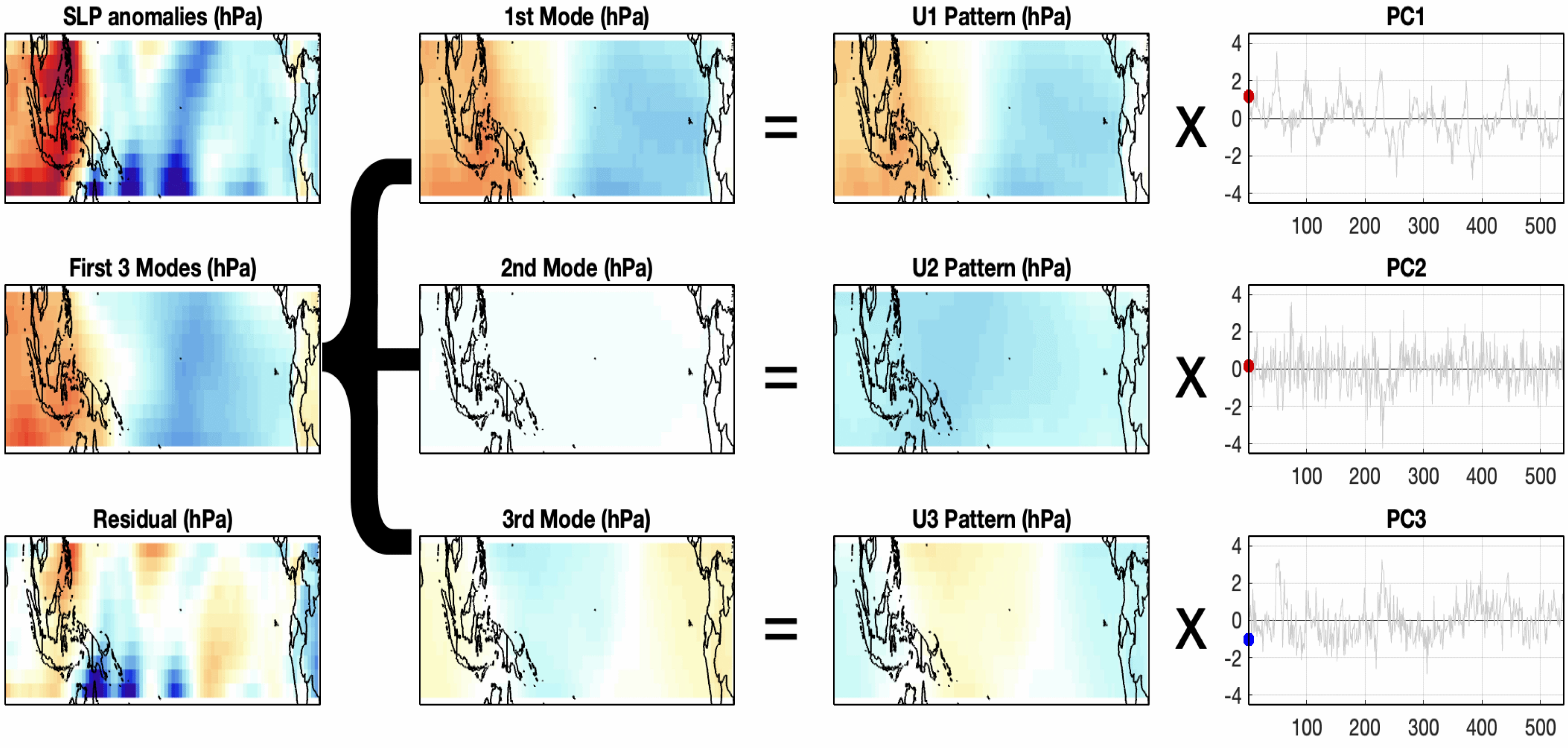
(2) Grid boxes needs to be weighted by the square root of cosine latitude to account for the fact that high latitude grid boxes have smaller area.

Practical Notes on PCA/EOF analysis for ocean, earth, and climate data

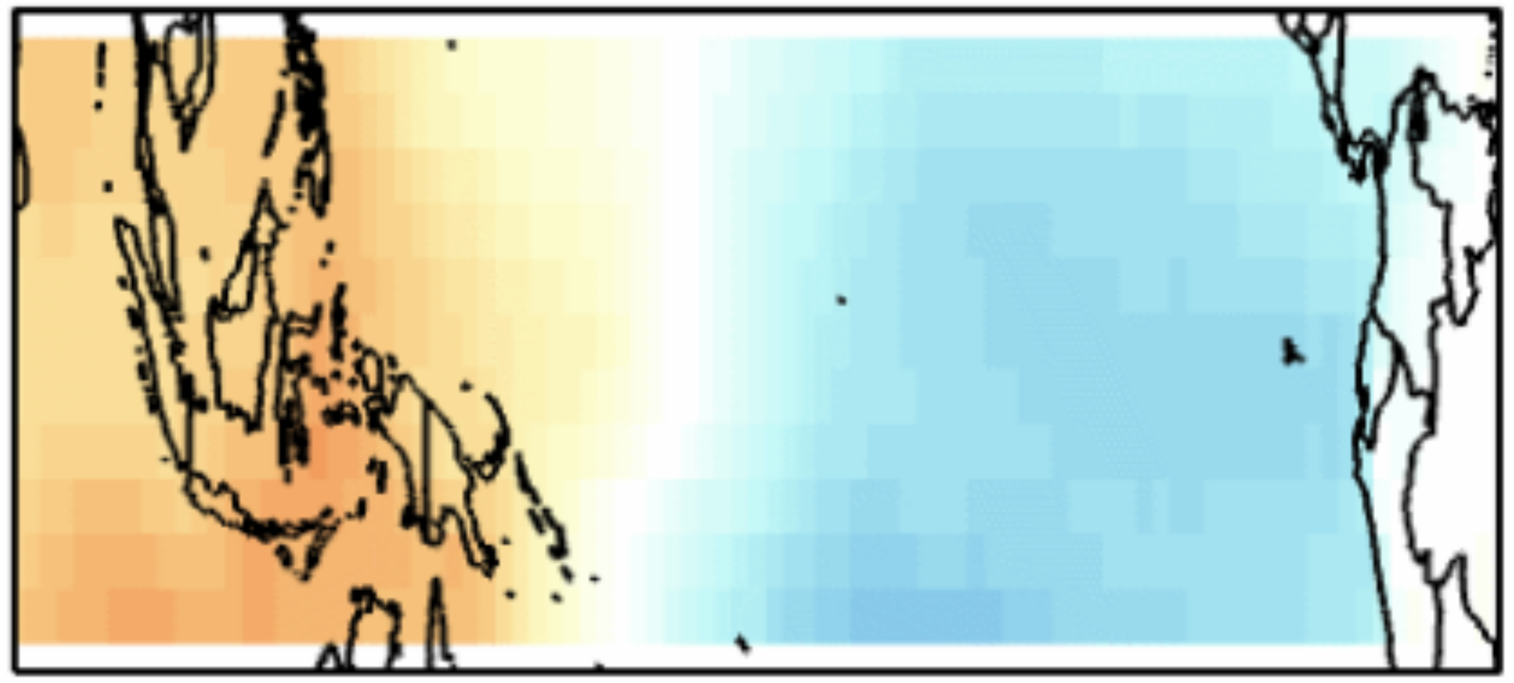


OR standardise v_n as the time series, and perform a linear regression of data against the standardised v_n at each location, and plot the regression slope as a map to get the pattern.

An example using Sea-Level Pressure (SLP) over the Equatorial Pacific



Using Caropy to plot maps




```
import cartopy.crs as ccrs

fig = plt.figure(figsize=(10, 5))
ax = fig.add_subplot(1, 1, 1, projection=ccrs.PlateCarree(...));
ax.set_extent([65, 295, -30, 30], crs=ccrs.PlateCarree());

slp_contour = ax.pcolor(..., transform=ccrs.PlateCarree(),...)
cbar = plt.colorbar(slp_contour, ...);
cbar.set_label('Sea Level Pressure [--]');

ax.coastlines();
ax.add_feature(cfeature.BORDERS, linestyle=':');
```

Road Map of the Statistics Part