PCA and OLS have different goals.

Goals of PCA—

argmin_V (X - X*V*V_transpose)^2

subject to:

Principal component vectors in V, pairwise orthogonal.

where:

X is the matrix that needs to be reduced in dimension,

V is the transition matrix of PCA,

V_transpose is the transpose matrix of the PCA transition matrix.

Usually, when we use PCA, we want to use the transition matrix V to reduce the high-dimensional raw data X to the low-dimensional data X_=X*V. In the process of reducing the dimension, we want to ensure that the information in X is preserved as much as possible, or in other words we want the variance of X_ to be as large as possible. We achieve this goal by maximizing VAR(X*V), in the formula The V is the variable we want to solve, and V stores the information on how we transform X.

The result X_ after dimensionality reduction, on the one hand, retains the information in X; on the other hand, the dimension is much lower than X. It can be seen that the goal of PCA is not to form a certain mapping relationship, but to compress information while retaining information, so PCA is a kind of unsupervised learning method. We can usually use PCA to extract principal components in the data for analysis, or reduce high-dimensional data to low-dimensional data to improve computational efficiency, Another scenario where PCA is used is data visualization.

Goals of OLS—

argmin_B(X*B - Y)^2

where:

X is the matrix that needs to be mapped,

B is the regression coefficient in linear regression using the OLS method.

Usually, when we use OLS to solve linear regression, we want the predicted value of the model Y_hat=X*B to be as close to the target variable Y as possible. When solving the regression problem, we hope that the estimated value Y_hat is as close to the real value Y as possible, that is, to make the residual between Y_hay and Y as small as possible. It can be seen that the role of OLS is to find the mapping relationship between X and Y. X is an independent variable and Y is a dependent variable, so using OLS to solve linear regression is a supervised learning algorithm. We can usually use OLS to perform linear regression to explore the linear relationship between independent and dependent variables.

Explain how PCA can be used to come up with a line that "best describes the data". Define what "best" means here.

We can use PCA to reduce the high-dimensional data X to 1 dimension to obtain a line that preserves the information in X as much as possible.
For example, if V is the first principal component vector of X and shape(V)=(D,), we can obtain the most informative line by X*V.
When we use the word "best", we mean that PCA tries to ensure that the information in the original data is preserved, or that "best" is equivalent to "maximize", and we maximize VAR(X*V).

How does this line compare with the one that results from ordinary least-squares linear regression and when would you choose one versus the other?

The results generated by PCA are completely different from those generated by OLS, because their goals were different from the beginning. The goal of PCA is to preserve the information in the original data X, and the goal of OLS is to find the mapping relationship between the independent variable X and the target variable Y.
In the unsupervised learning scenario, I will choose the PCA method, which can reduce the dimensionality of the data, help me better understand the data, help me visualize the data, and analyze the statistical properties of the data. In a supervised learning scenario, I would choose

the OLS method, which can help me find linear relationships between variables and help me understand how each independent variable affects the dependent variable.

Conditions of use of PCA:

1. Make sure the scale of the variables are consistent. If inconsistent, variables with larger units will have a greater impact on the results.
2. Make sure to center each variable. Centering makes the math easier.

Conditions of use of OLS:

1. The linear regression model is linear in parameters(B).
2. There is no multicollinearity in X.
3. The conditional mean of the error term is zero.
4. There is homoscedasticity and no autocorrelation for error term.

Draw or plot an example for D =2 to facilitate your discussion.

PCA example—

see code part

OLS example—

see code part