

What is Partial Least Squares and how does it relate to data analysis?

Authored by
stats writer

April 22, 2024

RECOMMENDED CITATION

stats writer (2024). *What is Partial Least Squares and how does it relate to data analysis?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=138028>

Partial Least Squares (PLS) is a statistical method used for data analysis, specifically in situations where there are multiple variables and the relationship between them is complex. It is a multivariate technique that combines aspects of principal component analysis and regression analysis to identify patterns and relationships within the data. PLS is particularly useful for handling datasets with high dimensionality, collinearity, and non-normality. It is often used in fields such as chemometrics, marketing, and neuroscience to extract meaningful information from large and complex datasets. Overall, PLS is a powerful tool for analyzing data and identifying important factors that contribute to a particular outcome or phenomenon.

An Introduction to Partial Least Squares

One of the most common problems that you'll encounter in machine learning is multicollinearity. This occurs when two or more predictor variables in a dataset are highly correlated.

When this occurs, a model may be able to fit a training dataset well but it may perform poorly on a new dataset it has never seen because it overfits the training set.

One way to get around the problem of multicollinearity is to use principal components regression, which calculates M linear combinations (known as "principal components") of the original p predictor variables and then uses the method of least squares to fit a linear regression model using the principal components as predictors.

The drawback of principal components regression (PCR) is that it does not consider the response variable when calculating the principal components.

Instead, it only considers the magnitude of the variance among the predictor variables captured by the principal components. Because of this, it's possible that in some cases the principal components with the largest variances aren't actually able to predict the response variable well.

A technique that is related to PCR is known as partial least squares. Similar to PCR, partial least squares calculates M linear combinations (known as "PLS components") of the original p predictor variables and uses the method of least squares to fit a linear regression model using the PLS components as predictors.

But unlike PCR, partial least squares attempts to find linear combinations that explain the variation in *both* the response variable and the predictor variables.

Steps to Perform Partial Least Squares

In practice, the following steps are used to perform

partial least squares.

1. Standardize the data such that all of the predictor variables and the response variable have a mean of 0 and a standard deviation of 1. This ensures that each variable is measured on the same scale.

2. Calculate $Z_1, \dots, Z_M$ to be the M linear combinations of the original p predictors.

$Z_m = \sum \Phi_{jm} X_j$ for some constants $\Phi_{1m}, \Phi_{2m}, \Phi_{pm}, m = 1, \dots, M$. To calculate Z_1 , set Φ_{j1} equal to the coefficient from the simple linear regression of Y onto X_j is the linear combination of the predictors that captures the most variance possible. To calculate Z_2 , regression each variable on Z_1 and take the residuals. Then calculate Z_2 using this orthogonalized data in exactly the same manner that Z_1 was calculated. Repeat this process M times to obtain the M PLS components.

3. Use the method of least squares to fit a linear regression model using the PLS components $Z_1, \dots, Z_M$ as predictors.

4. Lastly, use k-fold cross-validation to find the optimal

number of PLS components to keep in the model. The "optimal" number of PLS components to keep is typically the number that produces the lowest test mean-squared error (MSE).

Conclusion

In cases where multicollinearity is present in a dataset, partial least squares tends to perform better than ordinary least squares regression. However, it's a good idea to fit several different models so that we can identify the one that generalizes best to unseen data.

In practice, we fit many different types of models (PLS, PCR, Ridge, Lasso, Multiple Linear Regression, etc.) to a dataset and use k-fold cross-validation to identify the model that produces the lowest test MSE on new data.