SNS COLLEGE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION)

COIMBATORE – 35

DEPARTMENT OF COMPUTER SIENCE AND ENGINEERING

UNIT 2

Derived Input Directions

- **Principal component regression**
- **Partial Least Squares**

This topic presents **regression methods based on dimension reduction techniques,** which can be very useful when you have a **large data set with multiple correlated predictor variables.**

Generally, all dimension reduction methods work by first summarizing the original predictors into few new variables called principal components (PCs), which are then used as predictors to fit the linear regression model. These methods avoid multicollinearity between predictors, which a big issue in regression setting

When using the dimension reduction methods, it's generally recommended to standardize each predictor to make them comparable. Standardization consists of dividing the predictor by its standard deviation.

Here, we described two well known **regression methods based on dimension reduction: Principal Component Regression** (**PCR**) and **Partial Least Squares** (**PLS**) regression.

Principal component regression

The Principal Component Regression (PCR) first applies **Principal Component Analysis** on the **data set to summarize the original predictor variables** into few new variables also known as principal components (PCs), which are a linear combination of the original data.

These PCs are then used to build the linear regression model. The number of principal components, to incorporate in the model, is chosen by cross-validation (cv). Note that, **PCR is suitable when the data set contains highly correlated predictors.**

Partial least squares regression

A possible drawback of PCR is that we have no guarantee that the selected principal components are associated with the outcome. Here, the selection of the principal components to incorporate in the model is not supervised by the outcome variable.

An alternative to PCR is the **Partial Least Squares** (PLS) regression, which identifies new principal components that not only summarizes the original predictors, but also that are **related to the outcome**. These components are then used to fit the regression model. So, compared to PCR, **PLS uses a dimension reduction strategy that is supervised by the outcome.**

Like PCR, PLS is convenient for data with highly-correlated predictors. The number of PCs used in PLS is generally chosen by cross-validation. Predictors and the outcome variables should be generally standardized, to make the variables comparable.

Reference Links:

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/152-principal-component-and-partial-least-squares-regression-essentials/