# Multivariate Analysis overview

## Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a well-known unsupervised dimensionality reduction technique that constructs relevant features/variables through linear (linear PCA) or non-linear (kernel PCA) combinations of the original variables (features). In this post, we will only focus on the famous and widely used linear PCA method.

The construction of relevant features is achieved by linearly transforming correlated variables into a smaller number of uncorrelated variables. This is done by projecting (dot product) the original data into the reduced PCA space using the eigenvectors of the covariance/correlation matrix aka the principal components (PCs).

The resulting projected data are essentially linear combinations of the original data capturing most of the variance in the data (Jolliffe 2002).

In summary, PCA is an orthogonal transformation of the data into a series of uncorrelated data living in the reduced PCA space such that the first component explains the most variance in the data with each subsequent component explaining less.

### When/Why to use PCA

* PCA technique is particularly useful in processing data where multi-colinearity exists between the features/variables.
* PCA can be used when the dimensions of the input features are high (e.g. a lot of variables).
* PCA can be also used for denoising and data compression.

## Factor Analysis (FA)

Factor Analysis (FA) is a method to reveal relationships between assumed latent variables and manifest variables. A latent variable is a concept that cannot be measured directly but it is assumed to have a relationship with several measurable features in data, called manifest variables. Manifest variables are directly measurable.

There are two main forms of FA: Exploratory and confirmatory FA. Exploratory FA is designed to uncover relationships between manifest variables and factors without any assumption about specific manifest variables being related to specific factors. Confirmatory FA tests if a specific factor in a model provides an adequate fit for the correlations between specific manifest variables.

The FA model is very similar to multiple linear regression because the measurable manifest variables are regressed against the latent variables. In other words, an FA model creates an assumption that observed relationships among manifest variables are due to relationships of the observed manifest variables to the latent variables.

## Multiple Linear Regression (MLR)

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

## Principal Component Regression (PCR)

<https://stats.stackexchange.com/questions/184434/what-is-the-difference-between-principal-component-analysis-pca-and-principal>

“Principal component analysis is a method of data reduction - representing a large number of variable by a (much) smaller number, each of which is a linear combination of the original variables.

One output of PCA is principal component scores. Principal component regression uses those scores as independent variables in a regression.”

## Partial least-squares regression (PLSR)

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models.

PLSR or partial least squares regression is a dimension reduction technique that shares similarities with principal component analysis.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

In principal component regression you seek to obtain a set of new variables (the principal components) that maximize the variance of X and that are uncorrelated to each other.

In PLSR you seek to obtain a set of new variables (the PLS components) that maximize the covariance between X and y and that are uncorrelated to each other.

In both techniques, the new components are uncorrelated. This means that if in your original dataset you were facing a multicolinearity problem (this is, you have predictors in x that are highly correlated between them) by using any of these techniques you will solve the problem, as your components will become uncorrelated.

EDIT: Answer comment

Observe that, in these techniques it is usual to set a threshold on the number of components, so you select the first k components out of a total maximum of p being p≥k

Since PCA maximize the variance of X, the first k components are the variables that best explain X, but it can happen that, when trying to use these variables in the prediction of y, you achieve poor predictive results because the information that related X and y is left in the principal components that you did not select.

On the other hand, PLS maximize covariance between X and y. This means that the first k PLS components are the ones that best explain the relation between X and y. And for this reason, PLS is expected to provide good predictive results.

Regarding your second question, why Multicollinearity of x will be related to the multiple dimension of y

I am not sure if I am understanding it correctly but I will try to provide an answer. In PLSR, as you say, your response variable can be multidimensional, but this has nothing to do with the multicolinearity of X. It is said that there is a multicolinearity problem if there are variables in x that are highly correlated between them, regardless of having a univariate or multivariate y.

<https://stats.stackexchange.com/questions/492235/how-does-plsr-solve-multicollinearity>

## Multidimensional Scaling (MDS)

https://www.statisticshowto.com/multidimensional-scaling/

Multidimensional scaling is a visual representation of distances or dissimilarities between sets of objects. “Objects” can be colors, faces, map coordinates, political persuasion, or any kind of real or conceptual stimuli (Kruskal and Wish, 1978). Objects that are more similar (or have shorter distances) are closer together on the graph than objects that are less similar (or have longer distances). As well as interpreting dissimilarities as distances on a graph, MDS can also serve as a dimension reduction technique for high-dimensional data (Buja et. al, 2007).