Multivariate Techniques: Principal Component Analysis.

Learning Objectives.

By the end of this lesson, the student should be able to:

- Understand what multivariate analysis is.
- Know some of the different multivariate techniques.
- Understand the need for data reduction.
- Understand what principal component analysis is, and how it works
- Reduce large datasets down to relevant factors using PCA
- Know how to apply PCA to machine learning models.
- Recognize when to use principal component analysis and factor analysis.

Overview.

When dealing with data that contains more than two variables, you'll use multivariate analysis. Multivariate techniques allow you to gain a deeper understanding of your data in relation to a specific business or real-world scenarios.

Multivariate Techniques.

1. What is multivariate analysis?

Multivariate analysis is a statistical method that measures relationships between two or more response variables. Multivariate techniques attempt to model reality where each situation, product, or decision involves more than a single factor. For example, the decision to purchase a car may take into consideration price, safety features, color, and functionality. There are several different multivariate techniques to choose from, based on assumptions about the nature of the data and the type of association under analysis. Each technique tests the theoretical models of a research question about associations against the observed data. The theoretical models are based on facts plus new hypotheses about plausible associations between variables.

2. Multivariate data analysis techniques.

Some of the common multivariate techniques include:

1. Multiple Regression Analysis

   Multiple regression is the most commonly utilized multivariate technique. It examines the relationship between a single metric dependent variable and two or more metric independent variables. The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity and equal variance are carefully observed. Multiple regression is often used as a forecasting tool.

2. Logistic Regression Analysis

   This technique is a variation of multiple regression that allows for the prediction of an event. It is allowable to utilize nonmetric (typically binary) dependent variables, as the objective is to arrive at a probabilistic assessment of a binary choice. The independent variables can be either discrete or continuous. A contingency table is produced, which shows the classification of observations as to whether the observed and predicted events match. The sum of events that were predicted to

occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events, is a measure of the effectiveness of the model. This tool helps predict the choices consumers might make when presented with alternatives.

3. Reduction Techniques

Reduction techniques help to perform dimensionality reduction which reduces the no. of variables considered during analysis. There are three techniques used; factor analysis, principal component analysis, and discriminant analysis.

a. Discriminant Analysis

The purpose of discriminant analysis is to correctly classify observations or people into homogeneous groups. The independent variables must be metric and must have a high degree of normality. The discriminant analysis builds a linear discriminant function, which can then be used to classify the observations. The overall fit is assessed by looking at the degree to which the group means differ (Wilkes Lambda or D2) and how well the model classifies. To determine which variables have the most impact on the discriminant function, it is possible to look at partial F values. The higher the partial F, the more impact that variable has on the discriminant function. This tool helps categorize people, like buyers and nonbuyers.

b. Factor Analysis

When there are many variables in a research design, it is often helpful to reduce the variables to a smaller set of factors. This is an independence technique, in which there is no dependent variable. Rather, the researcher is looking for the underlying structure of the data matrix. Ideally, the independent variables are normal and continuous, with at least three to five variables loading onto a factor.

c. Principal Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

Principal Component Analysis (PCA)

PCA is a data reduction technique, which means it's a way of capturing the variance in many variables in a smaller, easier-to-work-with set of variables. When dealing with a dataset with many variables, in order to interpret the data in a more meaningful form, it is necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component.

Principal components are new variables constructed as a linear combination of the initial variables. The components are uncorrelated and most of the explained variance in the data is found in the first principal component. They represent the direction of the data that explain the maximal amount of variance in the data.

When to use PCA:

- When you want to reduce the number of features
- When the dataset is too large to be handled
- When visualizing the data in lower dimensions

How to compute the principal components

1. **Standardization.**

   Since PCA is very sensitive to the variance in the variables, you need to first standardize the range of initial continuous variables so that each contributes equally to the analysis.

2. **Covariance Matrix Computation.**

   If the variables in a dataset are highly correlated, it shows that there's is redundant information in the dataset. We use the covariance matrix to check the correlation between the variables. The entries of the covariance matrix show how the variables are are varying from the mean with respect to each other. If the sign of the covariance is positive, in meant that the variables are correlated, if negative, they are inversely correlated.

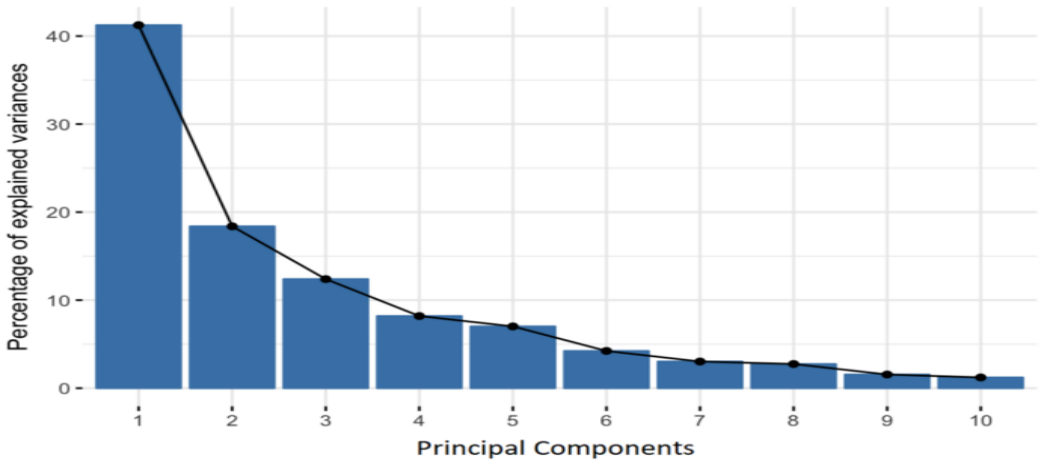
3. **Compute Eigenvectors and Eigenvalues.**

   Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

   Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set. PCA tries to put maximum possible information (variance) in the first component, then maximum remaining information in the second, and so on. Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, i.e the lines that capture most information of the data.

   Eigenvectors and ***eigenvalues*** always come in pairs, so every eigenvector has an eigenvalue, and their number is equal to the number of dimensions of the data. Eigenvectors of the covariance matrix are the directions of the axes where there is the most variance and eigenvalues are the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

   To compute the percentage of variance accounted for by each component, divide the eigenvalue of each component by the sum of eigenvalues. By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.



4. **Feature vector.**

   A feature vector is a matrix that has as columns the eigenvectors of the components that you decide to keep.

5. **Recast the data along the principal component axis.**

   Use the feature vector formed to reorient the data from the original axes to the ones represented by the principal components. This is done by multiplying the transpose of the feature vector with the transpose of the original dataset.

We use PCA on several occasions; when you want to reduce the number of features, when our dataset is too large to be handled, or when visualizing the data in lower dimensions.

Additional Resources.

1. An Introduction to Multivariate Analysis: Link
2. Eleven Multivariate Techniques: Link
3. A Step-by-Step Explanation of Principal Component Analysis (PCA): Link
4. Principal Component Analysis (PCA): Link
5. Principal component analysis: Link
6. Kaggle Example: Link

Python Exercise.

- Multivariate Techniques: Principal Component Analysis: Link