**Missing values processing**

Another shortcoming of standard approaches to PCA is that it is not obvious how to deal properly with incomplete data set, in which some of the points are missing. Currently the incomplete points are either discarded or completed using a variety of interpolation methods. However, such approaches are no longer valid when a significant portion of the measurement matrix is unknown. Typically, the training data for PCA is pre-processed in some way. But in some realistic problems where the amount of training data is huge, it becomes impractical to manually verify that all the data is *good*. In general, training data may contain some errors from the underlying data generation method. We view these error points as "outliers". However, the standard PCA algorithm is based on the assumption that data have not been spoiled by outliers.

When the percentage of missing data is very small, it is possible to replace the missing elements with the mean or an extreme value, which is a common strategy in multivariate statistics. However, such an approach is no longer valid when a significant portion of the measurement matrix is unknown. It is not unusual for a large portion of the matrix to be unobservable[1].

## Which are the methods to treat missing values?

1. **Deletion:** It is of two types: List Wise Deletion and Pair Wise Deletion.
    - In list wise deletion, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.
    - In pair wise deletion, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.

List wise deletion

| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| F | . | 280 |
| M | 33 | 332 |
| M | . | 272 |
| F | 25 | . |
| M | 29 | 326 |
| | 26 | 259 |
| M | 32 | 297 |

Pair wise deletion

| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| F | . | 280 |
| M | 33 | 332 |
| M | . | 272 |
| F | 25 | . |
| M | 29 | 326 |
| | 26 | 259 |
| M | 32 | 297 |

    - Deletion methods are used when the nature of missing data is "**Missing completely at random**" else non random missing values can bias the model output.

2. **Mean/ Mode/ Median Imputation:** Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types:-
    - **Generalized Imputation:** In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median. Like in above table, variable "**Manpower"** is missing so we take average of all non missing values of "**Manpower"** (**28.33**) and then replace missing value with it.
    - **Similar case Imputation:** In this case, we calculate average for gender "**Male"** (29.75) and "**Female**" (25) individually of non missing values then replace the missing value based on gender. For "**Male**", we will replace missing values of manpower with 29.75 and for "**Female**" with 25.

3. **Prediction Model**: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set.We can use regression, ANOVA, Logistic regression and various modeling technique to perform this. There are 2 drawbacks for this approach:
    1. The model estimated values are usually more well-behaved than the true values
    2. If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

4. **KNN Imputation:** In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantages and disadvantages.
    - **Advantages:**
        - k-nearest neighbour can predict both qualitative & quantitative attributes
        - Creation of predictive model for each attribute with missing data is not required
        - Attributes with multiple missing values can be easily treated
        - Correlation structure of the data is taken into consideration

- **Disadvantage:**
  - KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances.
  - Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

After dealing with missing values, the next task is to deal with outliers. Often, we tend to neglect outliers while building models. This is a discouraging practice. Outliers tend to make your data skewed and reduces accuracy[2].

[1] Principal Component Analysis With Missing Data and Outliers. *https://www.researchgate.net/publication/2910386_Principal_Component_Analysis_With_Missing_Data_and_Outliersv* (accessed Oct 2016).

[2]Statgraphics® Centurion User Manual by Statpoint Technologies Inc (2014). *https://es.scribd.com/document/315402228/Statgraphics-Centurion-XVII-User-Manual* (accessed Oct 2016).