MACHINE LEARNING

9TH QUARTER

COMPUTACIONAL ROBOTICS ENGINEERING

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

YULIANA ALEJANDRA MOLINA CORTES

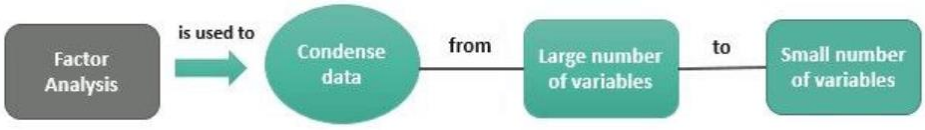
VICTOR ALEJANDRO MOO QUINTAL

BRANDON PACHECO CHAN

JOSHUA ZAMORA RODRIGUEZ

Factor Analysis

Factor Analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) “factors.” These factors typically are viewed as broad concepts or ideas that may describe an observed phenomenon. Factor analysis is used in big data as the data from a large number of variables may be condensed down into a smaller number of variables. In other words, it brings the number of variables in the attribute space down to a more manageable level, making it a method that is not dependent on any other variables (OARC Stats – Statistical Consulting Web Resources, 2023; WallStreetMojo, 2023).

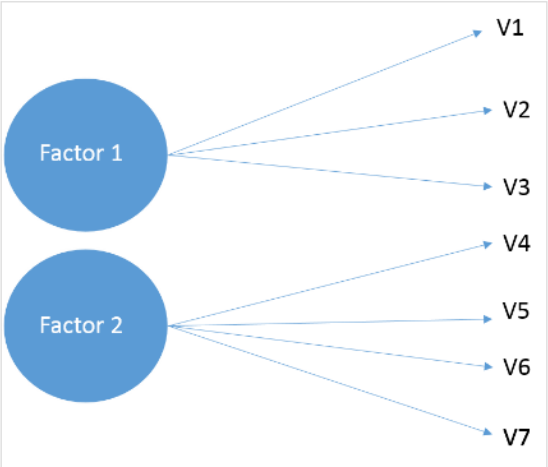


Factor analysis is generally an exploratory/descriptive method that requires many subjective judgments. It is a widely used tool and often controversial because the models, methods, and subjectivity are so flexible that debates about interpretations can occur (Babu, 2020). Factor analysis is one of the unsupervised machine learning algorithms which is used for dimensionality reduction. This algorithm creates factors from the observed variables to represent the common variance i.e., variance due to correlation among the observed variables.

$$x_i - \mu_i = l_1 F_1 + \cdots + l_n F_n + \varepsilon_i$$

x is the variable and F is the factor, and l is the factor loading which, can also be considered as the weight of the factor for the corresponding variable. The number of factors is equal to the number of variables (Babu, 2020). In simpler terms, factor analysis is about categorizing the labels beforehand to simplify the process, which can be achieved by identifying common factors and placing independent correlated variables that can become part of that group.

In simpler terms, factor analysis is about categorizing the labels beforehand to simplify the process, which can be achieved by identifying common factors and placing independent correlated variables that can become part of that group.



Example: Imagine you are an English teacher, and you want to measure how good is a class of students at speaking English. To do so you make some exams that measure:

- Listening
- Speaking
- Writing
- Reading

Those are random variables, and they measure different aspects of how good someone at speaking English is. Each person in your class has one exam note for each skill.

But let's say you want to simplify it; you want to explain the same with less variables. This is where factor analysis comes into action.

What is the purpose of factor analysis? The purpose of factor analysis is that given a set of data (with p variables and n data) explain it in a smaller number of variables losing as little information as possible. These variables are hidden, and they are called "factors".

In the above example if you want to explain these English skills in a single variable (a factor) that would be "English level" since this variable is positively correlated with each one of these skills.

Another way to explain this method is by classifying persons by if they have hair or not, instead of if they have curly, wavy, straight, long, short, blond, red, brown, black or grey hair. Or they don't have any hair. The method will recognize as a common factor hair and will classify all of them using this logic.

Types Of Factor Analysis

#1 – (PCA) Principal Component Analysis

It is the most popular methodology used by researchers because it takes the factors with the highest variance and places them in the first factor. After that, it takes out the variation that can be accounted for by the first component and then isolates the second factor. In addition, this continues right up to the final consideration. It is a dimensionality reduction technique that identifies and extracts the most important patterns and features from data, reducing complexity while retaining meaningful information (WallStreetMojo, 2023)..

#2 – Common Factor Analysis

This is the second most popular method for researchers because it separates the elements that contribute to the most prevalent variation. This method, which is utilized in SEM, does not take into account the interpretation of all of the variables. Common Factor Analysis is a statistical technique used to identify underlying common factors that explain the correlations among observed variables, reducing data complexity for better understanding (WallStreetMojo, 2023)..

#3 – Image Factor Analysis

In order to generate an accurate prediction of the factor in image factoring, it utilizes the OLS regression approach and is based on the correlation matrix as its foundation. Image analysis is a typical factor analysis method used to determine the variability of a group of variables. Image Factor Analysis is applied to image data to uncover underlying patterns or factors that explain variations in multiple images. It's used in fields like computer vision and image processing for dimensionality reduction and feature extraction (WallStreetMojo, 2023)..

#4 – Maximum Likelihood Approach

Maximum likelihood estimation operates on the correlation matrix, but it factors using the maximum likelihood technique. It is a technique used in statistics to estimate the parameters of an assumed probability distribution based on specific observed data. This is accomplished by optimizing a likelihood function in such a way that, according to the statistical model that is being assumed, the observed data has the highest probability.

Advantages and disadvantages

Some advantages of using Factor Analysis are: 1. the possibility of using both objective and subjective qualities; 2. It can be utilized to determine the hidden dimensions or limitations, which may or may not be obvious via direct analysis; 3. its simplicity and economic applicability; 4. The process of naming dimensions and applying them is up to interpretation (WallStreetMojo, 2023)..

However, it can also have some disadvantages:

- The researcher’s capacity to generate a comprehensive and accurate list of product characteristics determines the utility. The value of the technique will be lowered proportionally if crucial characteristics are overlooked.

- It may be challenging to name the components because numerous characteristics may be significantly connected for no obvious reason.
- The factor analysis is unable to provide a meaningful pattern if the variables that have been examined are completely unrelated to one another.
- It is only possible to know what the factors reflect; theorizing is the only way for researchers to gain insight into this topic (WallStreetMojo, 2023).

Conclusion

To conclude, Factor Analysis is a statistical technique used to explore and simplify complex relationships among a set of observed variables. It identifies underlying factors or latent variables that explain the correlations or patterns in the data. By reducing data dimensionality, it helps reveal the essential structures and associations within a dataset, making it easier to interpret and work with. Factor Analysis is widely employed in psychology, sociology, economics, and other fields to understand the latent constructs that influence observed behaviors or phenomena, aiding in hypothesis testing and data reduction while uncovering hidden insights.

References

OARC Stats – Statistical Consulting Web Resources. (2023). *A practical introduction to factor analysis: Exploratory factor analysis*. https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/

Babu, D. (2020, October 19). Factor Analysis-My ML Oreo Detector. *Towards Data Science*. https://towardsdatascience.com/factor-analysis-my-ml-oreo-detector-2e02abc2bb30

WallStreetMojo. (2023). Factor Analysis. https://www.wallstreetmojo.com/factor-analysis/