STA 3180 Statistical Modelling: Factor Analysis

Factor Analysis Lecture Notes

STA 3180 Statistical Modelling

What is Factor Analysis?

Factor Analysis is a statistical technique used to identify the underlying structure of a set of variables. It is used to reduce the number of variables in a dataset by grouping them into a smaller number of factors. The factors are derived from the correlations between the variables, and represent the underlying structure of the data.

Key Concepts:

• Factor Analysis is a statistical technique used to identify the underlying structure of a set of variables.

• It is used to reduce the number of variables in a dataset by grouping them into a smaller number of factors.

• Factors are derived from the correlations between the variables, and represent the underlying structure of the data.

Definitions:

• Variable: A variable is a characteristic or attribute that can take on different values.

• Factor: A factor is a latent variable that is inferred from a set of observed variables.

• Factor Loadings: Factor loadings are the correlations between the observed variables and the factors.

• Factor Scores: Factor scores are the estimated values of the factors for each individual in the dataset.

Rules:

• Factors must be orthogonal (uncorrelated) with each other.

• Factors should explain a significant amount of the variance in the data.

• Factors should have high loadings on several variables.

Examples:

Suppose we have a dataset containing the following variables: height, weight, age, gender, and shoe size. We can use factor analysis to identify the underlying structure of these variables. We might find that there are two factors: one representing physical characteristics (height, weight, and shoe size) and one representing demographic characteristics (age and gender). The factor loadings would indicate how strongly each variable is associated with each factor. For example, height might have a high loading on the physical characteristics factor, while age might have a high loading on the demographic characteristics factor. The factor scores would indicate the estimated value of each factor for each individual in the dataset.