# Tutorial – 6

**1.**

Correlation Analysis and Covariance Analysis are statistical methods used in SAS Enterprise Miner to look at the relationship between variables. Despite being connected, they are very different from one another.

| **Correlation Analysis** | **Covariance Analysis** |
|---|---|
| a. Correlation analysis measures the strength and direction of the linear relationship between two continuous variables. | a. Covariance analysis, on the other hand, measures the direction and magnitude of the linear relationship between two variables, whether they are continuous or discrete. |
| b. The correlation coefficient ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation. | b. A positive covariance indicates that the variables tend to move in the same direction, while a negative covariance indicates they tend to move in opposite directions. |
| c. Correlation analysis helps identify patterns and dependencies between variables. | c. Covariance alone does not provide a standardized measure of the strength of the relationship, making it harder to compare across different datasets. |

Although they take different approaches, correlation analysis, and covariance analysis can both aid in the identification of redundant features. Here are the benefits and drawbacks of each method for detecting redundant properties.

## Correlation Analysis

**Method:** Correlation analysis measures the linear relationship between two variables using correlation coefficients. High correlation values (close to 1 or -1) indicate a strong linear relationship, while values close to 0 indicate little or no relationship. When examining a set of variables, high correlations among them suggest redundancy.

**Advantages:**

a. Provides a standardized measure of the strength and direction of the relationship.

b. Useful for identifying linear dependencies between variables.

c. Helps identify variables that contribute similar information to the analysis.

**Disadvantages**:

a. Only detects linear relationships and may miss non-linear dependencies.

b. Does not account for the impact of other variables on the relationship.

c. Cannot determine causality.

### Covariance Analysis

**Method:** Covariance analysis calculates the covariance between two variables, which indicates how much they vary together. High positive or negative covariance values imply a strong relationship, while values close to 0 indicate little or no relationship. In the context of redundant attributes, high covariance suggests redundancy.

**Advantages:**

a. Can detect both linear and non-linear relationships.

b. Identifies variables that tend to vary together, even if the relationship is not strictly linear.

c. Considers the variability of variables, which can be informative in certain analyses.

**Disadvantages**:

a. Covariance values are not standardized, making it harder to compare across datasets or variables with different scales.

b. Does not provide a clear measure of the strength and direction of the relationship like correlation does.

c. Can be influenced by the units of measurement, which can make interpretation challenging.

**2.**

The Correlation Matrix Viewer in SAS Enterprise Miner can be used to see how the variables in our dataset are correlated. It is simpler to spot connections and patterns between variables thanks to the Correlation Matrix Viewer, which gives the correlation matrix a graphical representation. This is how we can apply it:

a. Open SAS Enterprise Miner and navigate to the project where our dataset is located.

b. In the diagram workspace, select the node representing your dataset. It is typically labeled as "Data Source" or "Import Data."

c. Right-click on the dataset node and select "Explore" from the context menu. This will open the Explore window.

d. In the Explore window, select the "Correlation Matrix" tab.

e. In the Variables list, select the variables for which we want to calculate the correlation. We can either select individual variables or use the arrow buttons to move variables between the Selected Variables list and the Available Variables list.

f. Once we have selected the variables, click on the "Run" button to generate the correlation matrix.

g. The Correlation Matrix Viewer window will open, displaying a matrix of correlation coefficients between the selected variables. The matrix will have variables listed on both the rows and columns.

h. The correlation coefficients will be color-coded to indicate the strength and direction of the correlation. Positive correlations are typically represented by shades of green, while negative correlations are represented by shades of red.

i. We can hover over each cell in the matrix to view the exact correlation coefficient. Additionally, the matrix viewer provides options to sort the variables by correlation coefficient, zoom in and out, and adjust the color scale.

j. You can also right-click on the matrix and select options such as "Sort Rows by Correlation" or "Sort Columns by Correlation" to reorganize the matrix based on correlation strength.

There are a few general rules that we can follow when reading correlation coefficients to better comprehend the relationship between variables. Here are a few ways that correlation coefficients can be interpreted.

a. Positive Correlation (0 to +1): A positive correlation indicates that as one variable increases, the other variable also tends to increase. The closer the correlation coefficient is to +1, the stronger the positive correlation. For example, if the correlation coefficient between two variables is +0.8, it suggests a strong positive relationship between them.

b. Negative Correlation (0 to -1): A negative correlation indicates that as one variable increases, the other variable tends to decrease. The closer the correlation coefficient is to -1, the stronger the negative correlation. For example, a correlation coefficient of -0.6 suggests a moderate negative relationship between two variables.

c. No Correlation (0): A correlation coefficient of 0 indicates no linear relationship between the variables. It suggests that the variables are independent of each other and do not move together in any consistent pattern.

d. Strength of Correlation: The magnitude of the correlation coefficient represents the strength of the relationship. Generally, correlation coefficients close to +1 or -1 indicate a strong relationship, while values closer to 0 indicate a weak relationship.

**3**.

The Principal Component Analysis (PCA) node in SAS Enterprise Miner can be used to find redundant attributes in a dataset. A set of data is transformed into a new set of uncorrelated variables called principal components using the dimensionality reduction technique PCA. By analysing how each characteristic contributes to the major components, redundant attributes can be found.

To use the PCA node in SAS Enterprise Miner for identifying redundant attributes,

a. Open SAS Enterprise Miner and create a new diagram or open an existing one.

b. Drag and drop the PCA node from the "Variables" tab onto the diagram.

c. Connect the PCA node to the appropriate data source node containing the dataset we want to analyze.

d. Double-click on the PCA node to open its properties window.

e. In the properties window, specify the input and output variables. Select the variables we want to analyze for redundancy by moving them from the "Available" list to the "Selected" list.

f. Optionally, we can also specify the number of principal components to retain or the proportion of variance to explain. This will determine the dimensionality of the transformed data.

g. Click on the "Results" tab in the properties window to specify the output options. We can choose to save the transformed data, loadings, and scores, which will help in interpreting the results.

h. Once you have set the desired options, click "OK" to close the properties window.

i. Connect the PCA node to the desired analysis or visualization nodes to further explore the results.

j. Run the diagram to execute the PCA analysis.

In PCA, there are several common measures of variance and covariance that are used to analyze and interpret the results. These measures help in understanding the importance of each principal component and the relationships between variables. Here are the key measures:

a. Eigenvalues: Eigenvalues represent the amount of variance explained by each principal component. Higher eigenvalues indicate that the corresponding principal components capture more of the data's variability. Typically, eigenvalues are sorted in descending order, and a scree plot is used to visualize the magnitude of eigenvalues. It helps determine how many principal components should be retained for further analysis.

b. Explained Variance Ratio: The explained variance ratio shows the proportion of total variance explained by each principal component. It is calculated by dividing each eigenvalue by the sum of all eigenvalues. This measure helps identify the principal components that contribute the most to the dataset's variability.

c. Loadings: Loadings represent the correlation between the original variables and the principal components. They indicate the strength and direction of the relationship between variables and principal components. Positive or negative loadings suggest whether variables are positively or negatively associated with a particular principal component. Loadings close to 0 indicate little or no influence on the principal component.

d. Covariance Matrix: The covariance matrix is a square matrix that shows the covariances between pairs of variables in the dataset. It provides information about the strength and direction of the linear relationship between variables. Diagonal elements of the covariance matrix represent the variances of individual variables, while off-diagonal elements represent the covariances between variable pairs.

**4.**

To use the factor analysis node in SAS Enterprise Miner to identify redundant attributes in a dataset,

a. Open SAS Enterprise Miner and create a new project or open an existing one.

b. Import dataset into the project. Ensure that the dataset contains the attributes you want to analyze for redundancy.

c. In the Diagram Workspace, locate the Factor Analysis node from the Toolbox pane on the left side. Drag and drop the Factor Analysis node onto the workspace.

d. Connect the dataset node to the Factor Analysis node by clicking on the output port of the dataset node and dragging the arrow to the input port of the Factor Analysis node.

e. Double-click on the Factor Analysis node to open its properties window.

f. In the properties window, select the variables or attributes that we want to analyze for redundancy. We can either select individual variables or choose to analyze all variables in the dataset.

g. Configure the options for the factor analysis. This includes specifying the number of factors to extract, the rotation method (such as Varimax or Promax), and other settings related to the analysis. Adjust these options based on your specific requirements.

h. Optionally, we can enable the "Save Scoring Code" option to generate SAS code that can be used to score new data based on the factor analysis results.

i. Click the Run button to execute the Factor Analysis node.

j. Once the analysis is complete, examine the output provided by the Factor Analysis node. It typically includes statistical summaries, factor loadings, eigenvalues, and other relevant information.

k. Look for variables with high loadings on the same factors. High loadings indicate strong associations between variables and factors. If multiple variables have high loadings on the same factor(s), it suggests redundancy among those variables.

l. Based on the factor analysis results, we can identify redundant attributes by considering variables that have high loadings on the same factor(s). Redundant attributes are those that provide similar or redundant information, and we can choose to remove or consolidate them from your dataset.

SAS Enterprise Miner offers several common factor analysis methods that you can use to analyze your dataset. These methods include:

a. Principal Component Analysis (PCA): PCA is a widely used factor analysis method that aims to extract uncorrelated factors that explain the maximum amount of variance in the original variables. It creates linear combinations of the original variables, known as principal components, which are ordered by the amount of variance they explain.

b. Principal Factor Analysis (PFA): PFA is similar to PCA but focuses on extracting factors that are correlated rather than uncorrelated. It allows for the possibility of correlated factors, which can be useful in certain scenarios.

c. Image Factor Analysis (IFA): IFA is a method used when you have a set of categorical or ordinal variables. It treats the observed variables as indicators of underlying latent factors and estimates the relationships between the observed variables and factors.

d. Alpha Factor Analysis (AFA): AFA is designed for analyzing Likert scale data where the response categories are ordered and represent levels of agreement or disagreement. It estimates factors that represent underlying dimensions and provides information on the reliability and validity of the factors.

**5.**

In SAS Enterprise Miner, attribute selection plays a crucial role in improving model accuracy, interpretability, and generalizability. Evaluating the impact of attribute selection on these aspects involves several steps. Here's a general framework to follow:

a. Define Evaluation Metrics: Start by identifying the evaluation metrics that are relevant to our specific modeling task. For accuracy, we might consider metrics like classification error, misclassification rate, or AUC (Area Under the ROC Curve). Interpretability can be evaluated based on the simplicity and understandability of the selected attributes. Generalizability refers to the ability of the model to perform well on unseen data, so you may use metrics like cross-validation performance or out-of-sample testing to assess generalizability.

b. Data Preparation: Prepare data by removing missing values, handling outliers, and standardizing or normalizing the attributes as necessary. Need to ensure that target variable and attributes are correctly defined within the SAS Enterprise Miner project.

c. Attribute Selection Methods: SAS Enterprise Miner offers various attribute selection techniques, including filter methods (e.g., correlation analysis, information gain) and wrapper methods (e.g., forward selection, backward elimination). Choose the most suitable method(s) based on the characteristics of your data and the goals of your modeling task.

d. Model Building: Create a baseline model using all available attributes to establish a benchmark for comparison. Then, apply the attribute selection technique(s), selected to identify a reduced set of attributes for building alternative models.

e. Model Evaluation: Train and evaluate models using the selected attributes. Compare the performance of these models against the baseline model using the evaluation metrics defined earlier. Assess changes in accuracy, interpretability, and generalizability based on the results.

f. Interpretability Analysis: Examine the selected attributes and the resulting model to assess interpretability. Evaluate whether the reduced set of attributes provides meaningful insights and whether the model is easy to understand and explain to stakeholders.

g. Generalizability Assessment: To evaluate generalizability, perform additional testing on unseen data or use cross-validation techniques. This step helps determine if the models built with selected attributes perform well on data they have not been exposed to during training.

There are several common measures of model performance that can be used to compare different models with different sets of attributes. Here are some key measures:

a. Accuracy: Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of the number of correct predictions to the total number of predictions. Accuracy is suitable for balanced datasets but can be misleading when dealing with imbalanced datasets.

b. Precision: Precision measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It focuses on the model's ability to minimize false positives. Precision is valuable in situations where false positives are costly or undesirable.

c. Recall (Sensitivity or True Positive Rate): Recall measures the proportion of correctly predicted positive instances out of the total actual positive instances. It focuses on the model's ability to capture all positive instances and minimize false negatives.

d. Specificity: Specificity measures the proportion of correctly predicted negative instances out of the total actual negative instances. It complements recall and focuses on the model's ability to correctly identify negative instances.

e. F1 Score: The F1 score combines precision and recall into a single metric. It is the harmonic mean of precision and recall and provides a balanced evaluation of both measures. The F1 score is useful when you want to consider both false positives and false negatives.

f. Area Under the ROC Curve (AUC-ROC): The AUC-ROC measures the model's ability to distinguish between positive and negative instances across different probability thresholds. It provides an aggregate measure of the model's performance and is suitable for evaluating binary classification models.

g. Mean Squared Error (MSE): MSE is commonly used for regression tasks and measures the average squared difference between the predicted and actual values. Lower MSE values indicate better model performance.