SAS ACECLUS Procedure: An Overview with an Example

The SAS ACECLUS (Approximate Covariance Estimation for Clustering) procedure is a powerful tool used primarily in cluster analysis. Its main function is to transform multivariate data into a space where cluster analysis becomes more effective by adjusting for correlations among variables. This transformation is accomplished by approximating the within-cluster covariance matrix, making the data more suitable for methods like k-means clustering or hierarchical clustering.

The ACECLUS procedure is particularly helpful when the original variables are correlated or have different scales, as it improves the clustering results by creating an uncorrelated and normalized dataset.

Key Features of PROC ACECLUS

1. Covariance-Based Transformation: It transforms the input data to reduce the influence of correlations among variables, leading to a better cluster separation.

2. Eigenvalue Control: Users can specify the number of principal components to retain through eigenvalue truncation, which improves clustering by focusing on the most relevant dimensions.

3. Iterative Method: The procedure iterates between estimating the cluster covariance and transforming the data, refining the approximation for better cluster separation.

Syntax Overview

The basic syntax for PROC ACECLUS looks like this:

```
Sas proc aceclus data=dataset out=output maxiter=number maxeigen=value;
  var variables;
run;
```

- data: The input dataset containing the variables for clustering.

- out: The output dataset with the transformed variables.

- maxiter: Specifies the maximum number of iterations.

- maxeigen: Limits the eigenvalues to control dimensionality reduction.

- var: Lists the variables to be used in the analysis.

Example: Transforming Data for Clustering

Suppose you have a dataset iris containing measurements of flower characteristics, and you want to prepare this data for clustering.

```
sas
proc aceclus data=sashelp.iris out=iris_transformed maxiter=10 maxeigen=2;
  var SepalLength SepalWidth PetalLength PetalWidth;
run;


proc print data=iris_transformed;
run;
```

Explanation:

- Input Dataset: The procedure uses the sashelp.iris dataset, which contains measurements of iris flowers (sepal length, sepal width, petal length, and petal width).

- Transformation: The maxiter=10 option allows up to 10 iterations to refine the covariance estimation. The maxeigen=2 option restricts the number of dimensions based on the top two eigenvalues, making the output suitable for 2-dimensional visualization or simplified clustering.

- Output Dataset: The transformed data is saved in iris_transformed, which will now contain variables that are more suitable for clustering due to the adjustments made by ACECLUS.

Conclusion

The SAS ACECLUS procedure is a valuable tool for transforming multivariate data into a more cluster-friendly form by adjusting for correlations and reducing dimensionality. This preprocessing step can significantly improve the quality of clustering results, making it a key part of any cluster analysis pipeline.