

Clustering Variables

The VARCLUS Procedure

Variable Reduction

- The VARCLUS procedure attempts to divide a set of variables into non-overlapping clusters, such that each cluster can be interpreted as essentially one-dimensional.
- This means that a large set of variables can be replaced by a single member of each cluster (to act as a representative), often with very little loss of information.

Variable Selection

You can use outside knowledge to guide the selection, or you can use the $1 - R^2$ ratio to determine which variables are the best candidates.

$$1 - R^2 \text{ ratio} = \frac{1 - R^2 \text{ own cluster}}{1 - R^2 \text{ next closest cluster}}$$

Small values of this ratio indicate that the variable has a strong correlation with its own cluster and a weak correlation with the other clusters.

Variable Reduction

- Variable clustering (PROC VARCLUS) reduces the number of variables, not just the number of dimensions.

Cluster	Variable	R-squared with		1-R**2 Ratio
		Own Cluster	Next Closest	
Cluster 1	RedMeat	0.5350	0.2185	0.5950
	WhiteMeat	0.4544	0.3331	0.8181
	Milk	0.5529	0.2721	0.6142
Cluster 2	Cereal	0.8255	0.4630	0.3250
	Nuts	0.8255	0.4549	0.3201
Cluster 3	Fish	0.7019	0.1365	0.3452
	Starch	0.7019	0.3075	0.4304
Cluster 4	FruitVeg	1.0000	0.0538	0.0000

← choose one

← choose one

← choose one

The Algorithm used in VARCLUS

1. A cluster is chosen for splitting. The selected cluster has either the smallest percentage of variation explained by its cluster component (using the PERCENT option) or the largest eigenvalue associated with the second principal component (using the MAXEIGEN option).

The Algorithm used in VARCLUS

2. The chosen cluster is split into two clusters by finding the first two principal components, rotating the components, and assigning each variable to the rotated component with which it has the higher squared correlation.
 - In PROC VARCLUS, the principal components are rotated using an orthoblique rotation (that is, raw quartimax rotation on the eigenvectors).

The Algorithm used in VARCLUS

3. The variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components.

The VARCLUS Procedure

Divisive Clustering

2nd Eigenvalue

