

Chapter 59

The VARCLUS Procedure

Chapter Table of Contents

OVERVIEW	3189
Background	3189
GETTING STARTED	3191
SYNTAX	3196
PROC VARCLUS Statement	3196
BY Statement	3201
FREQ Statement	3201
PARTIAL Statement	3202
SEED Statement	3202
VAR Statement	3202
WEIGHT Statement	3202
DETAILS	3203
Missing Values	3203
Using PROC VARCLUS	3203
Output Data Sets	3204
Computational Resources	3206
Interpreting VARCLUS Procedure Output	3207
Displayed Output	3207
ODS Table Names	3209
EXAMPLE	3209
Example 59.1 Correlations among Physical Variables	3209
REFERENCES	3216

Chapter 59

The VARCLUS Procedure

Overview

The VARCLUS procedure divides a set of numeric variables into either disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster, which may be either the first principal component or the centroid component. The first principal component is a weighted average of the variables that explains as much variance as possible. See Chapter 47, “The PRINCOMP Procedure,” for further details. Centroid components (the CENTROID option) are unweighted averages of either the standardized variables (the default) or the raw variables (if you specify the COV option).

PROC VARCLUS tries to maximize the sum across clusters of the variance of the original variables that is explained by the cluster components. Either the correlation or the covariance matrix can be analyzed. If correlations are used, all variables are treated as equally important. If covariances are used, variables with larger variances have more importance in the analysis.

PROC VARCLUS creates an output data set that can be used with the SCORE procedure to compute component scores for each cluster. A second output data set can be used by the TREE procedure to draw a tree diagram of hierarchical clusters.

Background

The VARCLUS procedure attempts to divide a set of variables into nonoverlapping clusters in such a way that each cluster can be interpreted as essentially unidimensional. For each cluster, PROC VARCLUS computes a component that can be either the first principal component or the centroid component and tries to maximize the sum across clusters of the variation accounted for by the cluster components. PROC VARCLUS is a type of oblique component analysis related to multiple group factor analysis (Harman 1976).

The VARCLUS procedure can be used as a variable-reduction method. A large set of variables can often be replaced by the set of cluster components with little loss of information. A given number of cluster components does not generally explain as much variance as the same number of principal components on the full set of variables, but the cluster components are usually easier to interpret than the principal components, even if the latter are rotated.

For example, an educational test might contain fifty items. PROC VARCLUS can be used to divide the items into, say, five clusters. Each cluster can then be treated as a subtest, with the subtest scores given by the cluster components. If the cluster

components are centroid components of the covariance matrix, each subtest score is simply the sum of the item scores for that cluster.

By default, PROC VARCLUS begins with all variables in a single cluster. It then repeats the following steps:

1. A cluster is chosen for splitting. Depending on the options specified, 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).
2. The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvectors), and assigning each variable to the rotated component with which it has the higher squared correlation.
3. Variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components. The reassignment may be required to maintain a hierarchical structure.

The procedure stops when each cluster satisfies a user-specified criterion involving either the percentage of variation accounted for or the second eigenvalue of each cluster. By default, PROC VARCLUS stops when each cluster has only a single eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying factor dimension. The iterative reassignment of variables to clusters proceeds in two phases. The first is a nearest component sorting (NCS) phase, similar in principle to the nearest centroid sorting algorithms described by Anderberg (1973). In each iteration, the cluster components are computed, and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each variable is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested. The NCS phase is much faster than the search phase but is more likely to be trapped by a local optimum.

You can have the iterative reassignment phases restrict the reassignment of variables such that hierarchical clusters are produced. In this case, when a cluster is split, a variable in one of the two resulting clusters can be reassigned to the other cluster resulting from the split but not to a cluster that is not part of the original cluster (the one that is split).

If principal components are used, the NCS phase is an alternating least-squares method and converges rapidly. The search phase is very time consuming for a large number of variables and is omitted by default. If the default initialization method is used, the search phase is rarely able to improve the results of the NCS phase. If random initialization is used, the NCS phase may be trapped by a local optimum from which the search phase can escape.

If centroid components are used, the NCS phase is not an alternating least-squares method and may not increase the amount of variance explained; therefore, it is limited, by default, to one iteration.

Getting Started

This example demonstrates how you can use the VARCLUS procedure to create hierarchical, unidimensional clusters of variables.

The following data, from Hand, et al. (1994), represent amounts of protein consumed from nine food groups for each of 25 European countries. The nine food groups are red meat (RedMeat), white meat (WhiteMeat), eggs (Eggs), milk (Milk), fish (Fish), cereal (Cereal), starch (Starch), nuts (Nuts), and fruits and vegetables (FruitVeg).

Suppose you want to simplify interpretation of the data by reducing the number of variables to a smaller set of variable cluster components. You can use the VARCLUS procedure for this type of variable reduction.

The following DATA step creates the SAS data set Protein:

```

data Protein;
  input Country $18. RedMeat WhiteMeat Eggs Milk
    Fish Cereal Starch Nuts FruitVeg;
  datalines;
Albania      10.1  1.4  0.5   8.9  0.2  42.3  0.6  5.5  1.7
Austria      8.9 14.0  4.3  19.9  2.1  28.0  3.6  1.3  4.3
Belgium     13.5  9.3  4.1  17.5  4.5  26.6  5.7  2.1  4.0
Bulgaria     7.8  6.0  1.6   8.3  1.2  56.7  1.1  3.7  4.2
Czechoslovakia 9.7 11.4  2.8  12.5  2.0  34.3  5.0  1.1  4.0
Denmark     10.6 10.8  3.7  25.0  9.9  21.9  4.8  0.7  2.4
E Germany    8.4 11.6  3.7  11.1  5.4  24.6  6.5  0.8  3.6
Finland     9.5  4.9  2.7  33.7  5.8  26.3  5.1  1.0  1.4
France     18.0  9.9  3.3  19.5  5.7  28.1  4.8  2.4  6.5
Greece     10.2  3.0  2.8  17.6  5.9  41.7  2.2  7.8  6.5
Hungary     5.3 12.4  2.9   9.7  0.3  40.1  4.0  5.4  4.2
Ireland    13.9 10.0  4.7  25.8  2.2  24.0  6.2  1.6  2.9
Italy       9.0  5.1  2.9  13.7  3.4  36.8  2.1  4.3  6.7
Netherlands 9.5 13.6  3.6  23.4  2.5  22.4  4.2  1.8  3.7
Norway      9.4  4.7  2.7  23.3  9.7  23.0  4.6  1.6  2.7
Poland      6.9 10.2  2.7  19.3  3.0  36.1  5.9  2.0  6.6
Portugal    6.2  3.7  1.1   4.9 14.2  27.0  5.9  4.7  7.9
Romania     6.2  6.3  1.5  11.1  1.0  49.6  3.1  5.3  2.8
Spain       7.1  3.4  3.1   8.6  7.0  29.2  5.7  5.9  7.2
Sweden     9.9  7.8  3.5   4.7  7.5  19.5  3.7  1.4  2.0
Switzerland 13.1 10.1  3.1  23.8  2.3  25.6  2.8  2.4  4.9
UK         17.4  5.7  4.7  20.6  4.3  24.3  4.7  3.4  3.3
USSR       9.3  4.6  2.1  16.6  3.0  43.6  6.4  3.4  2.9
W Germany   11.4 12.5  4.1  18.8  3.4  18.6  5.2  1.5  3.8
Yugoslavia  4.4  5.0  1.2   9.5  0.6  55.9  3.0  5.7  3.2
;

```

The data set Protein contains the character variable Country and the nine numeric variables representing the food groups. The \$18. in the INPUT statement specifies that the variable Country is a character variable with a length of 18.

The following statements create the variable clusters.

```
proc varclus data=Protein outtree=tree centroid maxclusters=4;
  var RedMeat-FruitVeg;
run;
```

The DATA= option specifies the SAS data set Protein as input. The OUTTREE= option creates the output SAS data set Tree to contain the tree structure information. When you specify this option, you are implicitly requiring the clusters to be hierarchical rather than disjoint.

The CENTROID option specifies the centroid method of clustering. This means that the calculated cluster components are the unweighted averages of the standardized variables. The MAXCLUSTERS=4 option specifies that no more than four clusters be computed.

The VAR statement lists the numeric variables (RedMeat—FruitVeg) to be used in the analysis.

The results of this analysis are displayed in the following figures.

Although PROC VARCLUS displays output for each step in the clustering process, the following figures display only the final analysis for four clusters. Figure 59.1 displays the final cluster summary.

Oblique Centroid Component Cluster Analysis				
Cluster summary for 4 clusters				
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	4	4	2.173024	0.5433
2	2	2	1.650997	0.8255
3	2	2	1.403853	0.7019
4	1	1	1	1.0000
Total variation explained = 6.227874 Proportion = 0.6920				

Figure 59.1. Final Cluster Summary from the VARCLUS Procedure

For each cluster, Figure 59.1 displays the number of variables in the cluster, the cluster variation, the total explained variation, and the proportion of the total variance explained by the variables in the cluster. The variance explained by the variables in a cluster is similar to the variance explained by a factor in common factor analysis, but it includes contributions only from the variables in the cluster rather than from all variables.

The line labeled 'Total variation explained' in Figure 59.1 gives the sum of the explained variation over all clusters. The final 'Proportion' represents the total explained variation divided by the sum of cluster variation. This value, 0.6920, indicates that about 69% of the total variation in the data can be accounted for by the four clusters.

Figure 59.2 shows how the variables are clustered. The first cluster represents animal protein (RedMeat, WhiteMeat, Eggs, and Milk), the second cluster contains the variables Cereal and Nuts, the third cluster is composed of the variables Fish and Starch, and the last cluster contains the single variable representing fruits and vegetables (FruitVeg).

Oblique Centroid Component Cluster Analysis				
Cluster	Variable	R-squared with		1-R**2 Ratio
		Own Cluster	Next Closest	
Cluster 1	RedMeat	0.4375	0.1518	0.6631
	WhiteMeat	0.6302	0.3331	0.5545
	Eggs	0.7024	0.4902	0.5837
	Milk	0.4288	0.2721	0.7847
Cluster 2	Cereal	0.8255	0.3983	0.2900
	Nuts	0.8255	0.5901	0.4257
Cluster 3	Fish	0.7019	0.1365	0.3452
	Starch	0.7019	0.3075	0.4304
Cluster 4	FruitVeg	1.0000	0.0578	0.0000

Figure 59.2. R-square Values from the VARCLUS Procedure

Figure 59.2 also displays the R^2 value of each variable with its own cluster and the R^2 value with its nearest cluster. The R^2 value for a variable with the nearest cluster should be low if the clusters are well separated. The last column displays the ratio of $1 - R_{own}^2 / 1 - R_{nearest}^2$ for each variable. Small values of this ratio indicate good clustering.

Figure 59.3 displays the cluster structure and the intercluster correlations. The structure table displays the correlation of each variable with each cluster component. This gives an indication of how and to what extent the cluster represents the variable. The table of intercorrelations contains the correlations between the cluster components.

Oblique Centroid Component Cluster Analysis				
Cluster Structure				
Cluster	1	2	3	4
RedMeat	0.66145	-0.38959	0.06450	-0.34109
WhiteMeat	0.79385	-0.57715	0.04760	-0.06132
Eggs	0.83811	-0.70012	0.30902	-0.04552
Milk	0.65483	-0.52163	0.16805	-0.26096
Fish	-0.08108	-0.36947	0.83781	0.26614
Cereal	-0.58070	0.90857	-0.63111	0.04655
Starch	0.41593	-0.55448	0.83781	0.08441
Nuts	-0.76817	0.90857	-0.37089	0.37497
FruitVeg	-0.24045	0.23197	0.20920	1.00000

Inter-Cluster Correlations				
Cluster	1	2	3	4
1	1.00000	-0.74230	0.19984	-0.24045
2	-0.74230	1.00000	-0.55141	0.23197
3	0.19984	-0.55141	1.00000	0.20920
4	-0.24045	0.23197	0.20920	1.00000

Figure 59.3. Cluster Correlations and Intercorrelations

PROC VARCLUS next displays the summary table of statistics for the cluster history (Figure 59.4). The first three columns give the number of clusters, the total variation explained by clusters, and the proportion of variation explained by clusters.

As displayed in Figure 59.4, when the number of allowable clusters is two, the total variation explained is 3.9607, and the cumulative proportion of variation explained by two clusters is 0.4401. When the number of clusters increases to three, the proportion of explained variance increases to 0.5880. When four clusters are computed, the explained variation is 0.6920.

Oblique Centroid Component Cluster Analysis					
Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	0.732343	0.0814	0.0814	0.0875	
2	3.960717	0.4401	0.3743	0.1007	1.0213
3	5.291887	0.5880	0.5433	0.3928	0.7978
4	6.227874	0.6920	0.5433	0.4288	0.7847

Figure 59.4. Final Cluster Summary Table from the VARCLUS Procedure

Figure 59.4 also displays the minimum proportion of variance explained by a cluster, the minimum R^2 for a variable, and the maximum $(1 - R^2)$ ratio for a variable. The last quantity is the ratio of the value $1 - R^2$ for a variable's own cluster to the value $1 - R^2$ for its nearest cluster.

The following statements produce a tree diagram of the cluster structure created by PROC VARCLUS. First, the AXIS1 statement is defined. The ORDER= option specifies the data values in the order in which they should appear on the axis.

```
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none order=(0 to 1 by .2);
proc tree data=tree horizontal vaxis=axis1 haxis=axis2;
height _propor_;
run;
```

Next, the TREE procedure is invoked. The procedure uses the SAS data set *Tree*, created by the OUTTREE= option in the preceding PROC VARCLUS statement. The HORIZONTAL option orients the tree diagram horizontally. The VAXIS and HAXIS options specify the AXIS1 and AXIS2 statements, respectively, to customize the axes of the tree diagram. The HEIGHT statement specifies the use of the variable *_PROPOR_* (the proportion of variance explained) as the height variable.

Figure 59.5 shows how the clusters are created. The ordered variable names are displayed on the vertical axis. The horizontal axis displays the proportion of variance explained at each clustering level.

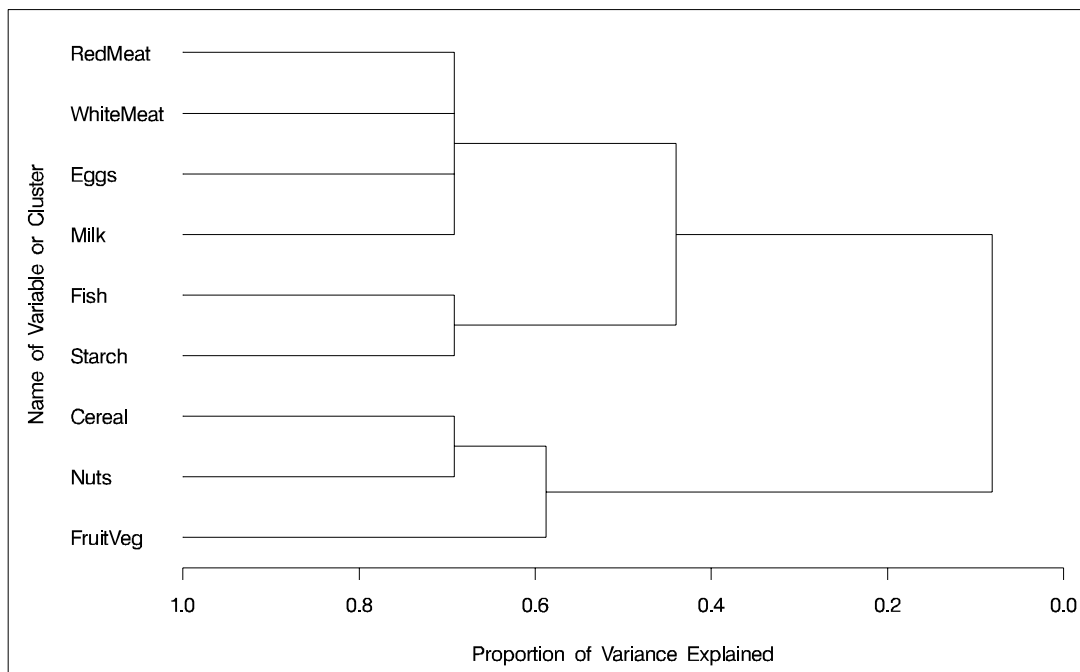


Figure 59.5. Horizontal Tree Diagram from PROC TREE

As you look from left to right in the diagram, objects and clusters are progressively joined until a single, all-encompassing cluster is formed at the right (or root) of the diagram. Clusters exist at each level of the diagram, and every vertical line connects leaves and branches into progressively larger clusters.

For example, when the variables are formed into three clusters, one cluster contains the variables RedMeat, WhiteMeat, Eggs, and Milk; the second cluster contains the variables Fish and Starch; the third cluster contains the variables Cereal, Nuts, and FruitVeg. The proportion of variance explained at that level is 0.5880 (from Figure 59.4). At the next stage of clustering, the third cluster is split as the variable FruitVeg forms the fourth cluster; the proportion of variance explained is 0.6920.

Syntax

The following statements are available in PROC VARCLUS.

```
PROC VARCLUS < options > ;
  VAR variables ;
  SEED variables ;
  PARTIAL variables ;
  WEIGHT variables ;
  FREQ variables ;
  BY variables ;
```

Usually you need only the VAR statement in addition to the PROC VARCLUS statement. The following sections give detailed syntax information for each of the statements, beginning with the PROC VARCLUS statement. The remaining statements are listed in alphabetical order.

PROC VARCLUS Statement

```
PROC VARCLUS < options > ;
```

The PROC VARCLUS statement starts the VARCLUS procedure and optionally identifies a data set or requests particular cluster analyses. By default, the procedure uses the most recently created SAS data set and omits observations with missing values from the analysis. Table 59.1 summarizes some of the options available in the PROC VARCLUS statement.

Table 59.1. Options available on the PROC VARCLUS statement

Task	Options
Specify data sets	DATA= OUTSTAT= OUTTREE=
Determine the number of clusters	MAXCLUSTERS= MINCLUSTERS= MAXEIGEN= PROPORTION=

Table 59.1. (continued)

Task	Options
Specify cluster formation	CENTROID COVARIANCE HIERARCHY INITIAL= MAXITER= MAXSEARCH= MULTIPLEGROUP RANDOM=
Control output	CORR NOPRINT SHORT SIMPLE SUMMARY TRACE
Omit intercept	NOINT
Specify divisor for variances	VARDEF=

The following list gives details on these options. The list is in alphabetical order.

CENTROID

uses centroid components rather than principal components. You should specify centroid components if you want the cluster components to be unweighted averages of the standardized variables (the default) or the unstandardized variables (if you specify the COV option). It is possible to obtain locally optimal clusterings in which a variable is not assigned to the cluster component with which it has the highest squared correlation. You cannot specify the CENTROID option with the MAXEIGEN= option.

CORR**C**

displays the correlation matrix.

COVARIANCE**COV**

analyzes the covariance matrix rather than the correlation matrix.

DATA=SAS-data-set

specifies the input data set to be analyzed. The data set can be an ordinary SAS data set or TYPE=CORR, UCORR, COV, UCOV, FACTOR, or SSCP. If you do not specify the DATA= option, the most recently created SAS data set is used. See Appendix A, "Special SAS Data Sets," for more information on types of SAS data sets.

HIERARCHY**HI**

requires the clusters at different levels to maintain a hierarchical structure.

INITIAL=GROUP**INITIAL=INPUT****INITIAL=RANDOM****INITIAL=SEED**

specifies the method for initializing the clusters. If the INITIAL= option is omitted and the MINCLUSTERS= option is greater than 1, the initial cluster components are obtained by extracting the required number of principal components and performing an orthoblique rotation. The following list describes the values for the INITIAL= option:

- | | |
|---------------|--|
| GROUP | specifies that clusters be initialized by group. You can use this option if the input data set is a TYPE=CORR, UCORR, COV, UCOV, or FACTOR data set. The cluster membership of each variable is obtained from an observation with _TYPE_='GROUP', which contains an integer for each variable ranging from one to the number of clusters. You can use a data set created either by a previous run of PROC VARCLUS or in a DATA step. |
| INPUT | specifies that the input data set is a TYPE=CORR, UCORR, COV, UCOV, or FACTOR data set, in which case scoring coefficients are read from observations where _TYPE_='SCORE'. You can use scoring coefficients from the FACTOR procedure or a previous run of PROC VARCLUS, or you can enter other coefficients in a DATA step. |
| RANDOM | assigns variables randomly to clusters. If you specify INITIAL=RANDOM without the CENTROID option, it is recommended that you specify MAXSEARCH=5, although the CPU time required is substantially increased. |
| SEED | initializes clusters according to the variables named in the SEED statement. Each variable listed in the SEED statement becomes the sole member of a cluster, and the other variables remain unassigned. If you do not specify the SEED statement, the first MINCLUSTERS= variables in the VAR statement are used as seeds. |

MAXCLUSTERS=*n***MAXC=*n***

specifies the largest number of clusters desired. The default value is the number of variables.

MAXEIGEN=*n*

specifies the largest permissible value of the second eigenvalue in each cluster. If you do not specify either the PROPORTION= or the MAXCLUSTERS= option, the default value is the average of the diagonal elements of the matrix being analyzed. This value is either the average variance if a covariance matrix is analyzed, or 1 if the correlation matrix is analyzed (unless some of the variables are constant, in which case the value is the number of nonconstant variables divided by the number of variables). Otherwise, the default is 0. The MAXEIGEN= option cannot be used with the CENTROID option.

MAXSEARCH=*n*

specifies the maximum number of iterations during the search phase. The default is 10 if you specify the CENTROID option; the default is 0 otherwise.

MINCLUSTERS=*n***MINC=*n***

specifies the smallest number of clusters desired. The default value is 2 if INITIAL=RANDOM or INITIAL=SEED; otherwise, the procedure begins with one cluster and tries to split it in accordance with the PROPORTION= or MAXEIGEN= option.

MULTIPLEGROUP**MG**

performs a multiple group component analysis (refer to Harman 1976). The input data set must be TYPE=CORR, UCORR, COV, UCOV, FACTOR or SSCP and must contain an observation with _TYPE_='GROUP' defining the variable groups. Specifying the MULTIPLEGROUP option is equivalent to specifying all of the following options: MINC=1, MAXITER=0, MAXSEARCH=0, MAXEIGEN=0, PROPORTION=0, and INITIAL=GROUP.

NOINT

requests that no intercept be used; covariances or correlations are not corrected for the mean. If you specify the NOINT option, the OUTSTAT= data set is TYPE=UCORR.

NOPRINT

suppresses the output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 14, “Using the Output Delivery System.”

OUTSTAT=*SAS-data-set*

creates an output data set to contain statistics including means, standard deviations, correlations, cluster scoring coefficients, and the cluster structure. If you want to create a permanent SAS data set, you must specify a two-level name. The OUTSTAT= data set is TYPE=UCORR if the NOINT option is specified. For more information on permanent SAS data sets, refer to “SAS Files” and “DATA Step Concepts” in *SAS Language Reference: Concepts*. For information on types of SAS data sets, see Appendix A.

OUTTREE=*SAS-data-set*

creates an output data set to contain information on the tree structure that can be used by the TREE procedure to print a tree diagram. The OUTTREE= option implies the HIERARCHY option. See Example 59.1 for use of the OUTTREE= option. If you want to create a permanent SAS data set, you must specify a two-level name. For more information on permanent SAS data sets, refer to “SAS Files” and “DATA Step Concepts” in *SAS Language Reference: Concepts*.

PROPORTION=*n***PERCENT=*n***

gives the proportion or percentage of variation that must be explained by the cluster component. Values greater than 1.0 are considered to be percentages, so PROPORTION=0.75 and PERCENT=75 are equivalent. If you specify the CENTROID option, the default value is 0.75; otherwise, the default value is 0.

MAXITER=*n*

specifies the maximum number of iterations during the alternating least-squares phase. The default value is 1 if you specify the CENTROID option; the default is 10 otherwise.

RANDOM=*n*

specifies a positive integer as a starting value for use with REPLACE=RANDOM. If you do not specify the RANDOM= option, the time of day is used to initialize the pseudo-random number sequence.

SHORT

suppresses printing of the cluster structure, scoring coefficient, and intercluster correlation matrices.

SIMPLE**S**

displays means and standard deviations.

SUMMARY

suppresses all default output except the final summary table.

TRACE

lists the cluster to which each variable is assigned during the iterations.

VARDEF=DF**VARDEF=N****VARDEF=WDF****VARDEF=WEIGHT | WGT**

specifies the divisor to be used in the calculation of variances and covariances. The default value is VARDEF=DF. The values and associated divisors are displayed in the following table.

Value	Divisor	Formula
DF	degrees of freedom	$n - i$
N	number of observations	n
WDF	sum of weights minus one	$(\sum_j w_j) - 1$
WEIGHT WGT	sum of weights	$\sum_j w_j$

In the preceding table, $i = 0$ if the NOINT option is specified, and $i = 1$ otherwise.

BY Statement

BY variables ;

You can specify a BY statement with PROC VARCLUS to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the VARCLUS procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

FREQ Statement

FREQ variable ;

If a variable in your data set represents the frequency of occurrence for the other values in the observation, include the variable's name in a FREQ statement. The procedure then treats the data set as if each observation appears n times, where n is the value of the FREQ variable for the observation. If the value of the FREQ variable is less than 1, the observation is not used in the analysis. Only the integer portion of the value is used. The total number of observations is considered equal to the sum of the FREQ variable.

PARTIAL Statement

PARTIAL *variable* ;

If you want to base the clustering on partial correlations, list the variables to be partialled out in the PARTIAL statement.

SEED Statement

SEED *variables* ;

The SEED statement specifies variables to be used as seeds to initialize the clusters. It is not necessary to use INITIAL=SEED if the SEED statement is present, but if any other INITIAL= option is specified, the SEED statement is ignored.

VAR Statement

VAR *variables* ;

The VAR statement specifies the variables to be clustered. If you do not specify the VAR statement and do not specify TYPE=SSCP, all numeric variables not listed in other statements (except the SEED statement) are processed. The default VAR variable list does not include the variable INTERCEPT if the DATA= data set is TYPE=SSCP. If the variable INTERCEPT is explicitly specified in the VAR statement with a TYPE=SSCP data set, the NOINT option is enabled.

WEIGHT Statement

WEIGHT *variables* ;

If you want to specify relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the weight variable are proportional to the reciprocals of the variances. The WEIGHT variable can take nonintegral values. An observation is used in the analysis only if the value of the WEIGHT variable is greater than zero.

Details

Missing Values

Observations containing missing values are omitted from the analysis.

Using PROC VARCLUS

Default options for PROC VARCLUS often provide satisfactory results. If you want to change the final number of clusters, use the MAXCLUSTERS=, MAXEIGEN=, or PROPORTION= option. The MAXEIGEN= and PROPORTION= options usually produce similar results but occasionally cause different clusters to be selected for splitting. The MAXEIGEN= option tends to choose clusters with a large number of variables, while the PROPORTION= option is more likely to select a cluster with a small number of variables.

Execution time

PROC VARCLUS usually requires more computer time than principal factor analysis, but it can be faster than some of the iterative factoring methods. If you have more than 30 variables, you may want to reduce execution time by one or more of the following methods:

- Specify the MINCLUSTERS= and MAXCLUSTERS= options if you know how many clusters you want.
- Specify the HIERARCHY option.
- Specify the SEED statement if you have some prior knowledge of what clusters to expect.

If computer time is not a limiting factor, you may want to try one of the following methods to obtain a better solution:

- Specify the MAXSEARCH= option with principal components and specify a value of 5 or 10.
- Try several factoring and rotation methods with PROC FACTOR to use as input to PROC VARCLUS.
- Run PROC VARCLUS several times, specifying INITIAL=RANDOM.

Output Data Sets

OUTSTAT= Data Set

The OUTSTAT= data set is TYPE=CORR, and it can be used as input to the SCORE procedure or a subsequent run of PROC VARCLUS. The variables it contains are

- BY variables
- _NCL_, a numeric variable giving the number of clusters
- _TYPE_, a character variable indicating the type of statistic the observation contains
- _NAME_, a character variable containing a variable name or a cluster name, which is of the form CLUS n where n is the number of the cluster
- the variables that are clustered

The values of the _TYPE_ variable are listed in the following table.

Table 59.2. _TYPE_ Value and Statistic

TYPE	Contents
MEAN	means
STD	standard deviations
USTD	uncorrected standard deviations, produced when the NOINT option is specified
N	number of observations
CORR	correlations
UCORR	uncorrected correlation matrix, produced when the NOINT option is specified
MEMBERS	number of members in each cluster
VAREXP	variance explained by each cluster
PROPOR	proportion of variance explained by each cluster
GROUP	number of the cluster to which each variable belongs
RSQUARED	squared multiple correlation of each variable with its cluster component
SCORE	standardized scoring coefficients
USCORE	scoring coefficients to be applied without subtracting the mean from the raw variables, produced when the NOINT option is specified
STRUCTUR	cluster structure
CCORR	correlations between cluster components

The observations with `_TYPE_='MEAN'`, `'STD'`, `'N'`, and `'CORR'` have missing values for the `_NCL_` variable. All other values of the `_TYPE_` variable are repeated for each cluster solution, with different solutions distinguished by the value of the `_NCL_` variable. If you want to specify the `OUTSTAT=` data set with the `SCORE` procedure, you can use a `DATA` step to select observations with the `_NCL_` variable missing or equal to the desired number of clusters. Alternatively, you can use a `WHERE` clause, as follows.

```
proc score score=s (where=(ncl_ =3)) data=newscore;
```

OUTTREE= Data Set

The `OUTTREE=` data set contains one observation for each variable clustered plus one observation for each cluster of two or more variables, that is, one observation for each node of the cluster tree. The total number of output observations is between n and $2n - 1$, where n is the number of variables clustered.

The variables in the `OUTTREE=` data set are

- `BY` variables, if any
- `_NAME_`, a character variable giving the name of the node. If the node is a cluster, the name is `CLUS n` where n is the number of the cluster. If the node is a single variable, the variable name is used.
- `_PARENT_`, a character variable giving the value of `_NAME_` of the parent of the node
- `_NCL_`, the number of clusters
- `_VAREXP_`, the total variance explained by the clusters at the current level of the tree
- `_PROPOR_`, the total proportion of variance explained by the clusters at the current level of the tree
- `_MINPRO_`, the minimum proportion of variance explained by a cluster component
- `_MAXEIG_`, the maximum second eigenvalue of a cluster

Computational Resources

Let

n = number of observations

v = number of variables

c = number of clusters

It is assumed that, at each stage of clustering, the clusters all contain the same number of variables.

Time

The time required for PROC VARCLUS to analyze a given data set varies greatly depending on the number of clusters requested, the number of iterations in both the alternating least-squares and search phases, and whether centroid or principal components are used.

The time required to compute the correlation matrix is roughly proportional to nv^2 .

Default cluster initialization requires time roughly proportional to v^3 . Any other method of initialization requires time roughly proportional to cv^2 .

In the alternating least-squares phase, each iteration requires time roughly proportional to cv^2 if centroid components are used or

$$\left(c + 5\frac{v}{c^2}\right)v^2$$

if principal components are used.

In the search phase, each iteration requires time roughly proportional to v^3/c if centroid components are used or v^4/c^2 if principal components are used. The HIERARCHY option speeds up each iteration after the first split by as much as $c/2$.

Memory

The amount of memory, in bytes, needed by PROC VARCLUS is approximately

$$v^2 + 2vc + 20v + 15c$$

Interpreting VARCLUS Procedure Output

Because PROC VARCLUS is a type of oblique component analysis, its output is similar to the output from the FACTOR procedure for oblique rotations. The scoring coefficients have the same meaning in both PROC VARCLUS and PROC FACTOR; they are coefficients applied to the standardized variables to compute component scores. The cluster structure is analogous to the factor structure containing the correlations between each variable and each cluster component. A cluster pattern is not displayed because it would be the same as the cluster structure, except that zeros would appear in the same places in which zeros appear in the scoring coefficients. The intercluster correlations are analogous to interfactor correlations; they are the correlations among cluster components.

PROC VARCLUS also displays a cluster summary and a cluster listing. The cluster summary gives the number of variables in each cluster and the variation explained by the cluster component. The latter is similar to the variation explained by a factor but includes contributions from only the variables in that cluster rather than from all variables, as in PROC FACTOR. The proportion of variance explained is obtained by dividing the variance explained by the total variance of variables in the cluster. If the cluster contains two or more variables and the CENTROID option is not used, the second largest eigenvalue of the cluster is also printed.

The cluster listing gives the variables in each cluster. Two squared correlations are calculated for each cluster. The column labeled “Own Cluster” gives the squared correlation of the variable with its own cluster component. This value should be higher than the squared correlation with any other cluster unless an iteration limit has been exceeded or the CENTROID option has been used. The larger the squared correlation is, the better. The column labeled “Next Closest” contains the next highest squared correlation of the variable with a cluster component. This value is low if the clusters are well separated. The column headed “1-R**2 Ratio” gives the ratio of one minus the “Own Cluster” R^2 to one minus the “Next Closest” R^2 . A small “1-R**2 Ratio” indicates a good clustering.

Displayed Output

The following items are displayed for each cluster solution unless the NOPRINT or SUMMARY option is specified. The CLUSTER SUMMARY table includes

- the Cluster number
- Members, the number of members in the cluster
- Cluster Variation of the variables in the cluster
- Variation Explained by the cluster component. This statistic is based only on the variables in the cluster rather than on all variables.
- Proportion Explained, the result of dividing the variation explained by the cluster variation

- Second Eigenvalue, the second largest eigenvalue of the cluster. This is displayed if the cluster contains more than one variable and the CENTROID option is not specified

PROC VARCLUS also displays

- Total variation explained, the sum across clusters of the variation explained by each cluster
- Proportion, the total explained variation divided by the total variation of all the variables

The cluster listing includes

- Variable, the variables in each cluster
- R-squared with Own Cluster, the squared correlation of the variable with its own cluster component; and R-squared with Next Closest, the next highest squared correlation of the variable with a cluster component. Own Cluster values should be higher than the R^2 with any other cluster unless an iteration limit is exceeded or you specify the CENTROID option. Next Closest should be a low value if the clusters are well separated.
- $1-R^{*2}$ Ratio, the ratio of one minus the value in the Own Cluster column to one minus the value in the Next Closest column. The occurrence of low ratios indicates well-separated clusters.

If the SHORT option is not specified, PROC VARCLUS also displays

- Standardized Scoring Coefficients, standardized regression coefficients for predicting cluster components from variables
- Cluster Structure, the correlations between each variable and each cluster component
- Inter-Cluster Correlations, the correlations between the cluster components

If the analysis includes partitions for two or more numbers of clusters, a final summary table is displayed. Each row of the table corresponds to one partition. The columns include

- Number of Clusters
- Total Variation Explained by Clusters
- Proportion of Variation Explained by Clusters
- Minimum Proportion (of variation) Explained by a Cluster
- Maximum Second Eigenvalue in a Cluster
- Minimum R-squared for a Variable
- Maximum $1-R^{*2}$ Ratio for a Variable

ODS Table Names

PROC VARCLUS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 14, “Using the Output Delivery System.”

Table 59.3. ODS Tables Produced in PROC VARCLUS

ODS Table Name	Description	Option
ClusterQuality	Cluster quality	default
ClusterStructure	Cluster structure	default
ClusterSummary	Cluster Summary	default
ConvergenceStatus	Convergence status	default
Corr	Correlations	CORR
DataOptSummary	Data and options summary table	default
InterClusterCorr	Inter-cluster correlations	default
IterHistory	Iteration history	TRACE
RSquared	Cluster Rsq	default
SimpleStatistics	Simple statistics	SIMPLE
StandScoreCoeff	Standardized scoring coefficients	default

Example

Example 59.1. Correlations among Physical Variables

The following data are correlations among eight physical variables as given by Harman (1976). The first PROC VARCLUS run clusters on the basis of principal components, the second run clusters on the basis of centroid components. The third analysis is hierarchical, and the TREE procedure is used to print a tree diagram. The results of the analyses follow.

```
data phys8(type=corr);
  title 'Eight Physical Measurements on 305 School Girls';
  title2 'Harman: Modern Factor Analysis, 3rd Ed, p22';
  label height='Height'
        arm_span='Arm Span'
        forearm='Length of Forearm'
        low_leg='Length of Lower Leg'
        weight='Weight'
        bit_diam='Bitrochanteric Diameter'
        girth='Chest Girth'
        width='Chest Width';
```

```

input _name_ $ 1-8
      (height arm_span forearm low_leg weight bit_diam
      girth width)(7.);
_type_='corr';
datalines;
height 1.0      .846  .805  .859  .473  .398  .301  .382
arm_span .846  1.0    .881  .826  .376  .326  .277  .415
forearm .805  .881  1.0    .801  .380  .319  .237  .345
low_leg .859  .826  .801  1.0    .436  .329  .327  .365
weight .473  .376  .380  .436  1.0    .762  .730  .629
bit_diam .398  .326  .319  .329  .762  1.0    .583  .577
girth .301  .277  .237  .327  .730  .583  1.0    .539
width .382  .415  .345  .365  .629  .577  .539  1.0
;

proc varclus data=phys8;
run;

```

The PROC VARCLUS statement invokes the procedure. By default, PROC VARCLUS clusters on the basis of principal components.

Output 59.1.1. Principal Cluster Components: Cluster Summary

```

Eight Physical Measurements on 305 School Girls
Harman: Modern Factor Analysis, 3rd Ed, p22

Oblique Principal Component Cluster Analysis

Cluster summary for 1 cluster
Cluster  Variation  Proportion  Second
Cluster  Members  Variation  Explained  Explained  Eigenvalue
-----
1         8         8         4.67288    0.5841     1.7710

Total variation explained = 4.67288 Proportion = 0.5841

Cluster 1 will be split.

Cluster summary for 2 clusters
Cluster  Variation  Proportion  Second
Cluster  Members  Variation  Explained  Explained  Eigenvalue
-----
1         4         4         3.509218   0.8773     0.2361
2         4         4         2.917284   0.7293     0.4764

Total variation explained = 6.426502 Proportion = 0.8033

R-squared with
-----
Cluster  Variable  Own  Next  1-R**2  Variable
Cluster  Variable  Cluster  Closest  Ratio  Label
-----
Cluster 1  height  0.8777  0.2088  0.1545  Height
           arm_span  0.9002  0.1658  0.1196  Arm Span
           forearm  0.8661  0.1413  0.1560  Length of Forearm
           low_leg  0.8652  0.1829  0.1650  Length of Lower Leg
-----
Cluster 2  weight  0.8477  0.1974  0.1898  Weight
           bit_diam  0.7386  0.1341  0.3019  Bitrochanteric Diameter
           girth  0.6981  0.0929  0.3328  Chest Girth
           width  0.6329  0.1619  0.4380  Chest Width

No cluster meets the criterion for splitting.
    
```

As displayed in Output 59.1.1, the cluster component (by default, the first principal component) explains 58.41% of the total variation in the 8 variables.

The cluster is split because the second eigenvalue is greater than 1 (the default value of the MAXEIGEN option).

The two resulting cluster components explain 80.33% of the variation in the original variables. The cluster summary table shows that the variables height, arm_span, forearm, and low_leg have been assigned to the first cluster; and that the variables weight, bit_diam, girth, and width have been assigned to the second cluster.

Output 59.1.2. Standard Scoring Coefficients and Cluster Structure Table

Oblique Principal Component Cluster Analysis			
Standardized Scoring Coefficients			
Cluster		1	2
height	Height	0.266977	0.000000
arm_span	Arm Span	0.270377	0.000000
forearm	Length of Forearm	0.265194	0.000000
low_leg	Length of Lower Leg	0.265057	0.000000
weight	Weight	0.000000	0.315597
bit_diam	Bitrochanteric Diameter	0.000000	0.294591
girth	Chest Girth	0.000000	0.286407
width	Chest Width	0.000000	0.272710

Cluster Structure			
Cluster		1	2
height	Height	0.936881	0.456908
arm_span	Arm Span	0.948813	0.407210
forearm	Length of Forearm	0.930624	0.375865
low_leg	Length of Lower Leg	0.930142	0.427715
weight	Weight	0.444281	0.920686
bit_diam	Bitrochanteric Diameter	0.366201	0.859404
girth	Chest Girth	0.304779	0.835529
width	Chest Width	0.402430	0.795572

The standardized scoring coefficients in Output 59.1.2 show that each cluster component has similar scores for each of its associated variables. This suggests that the principal cluster component solution should be similar to the centroid cluster component solution, which follows in the next PROC VARCLUS run.

The cluster structure table displays high correlations between the variables and their own cluster component. The correlations between the variables and the opposite cluster component are all moderate.

Output 59.1.3. Inter-Cluster Correlations

Oblique Principal Component Cluster Analysis			
Inter-Cluster Correlations			
Cluster		1	2
1		1.00000	0.44513
2		0.44513	1.00000

The intercluster correlation table shows that the cluster components are moderately correlated with $\rho = 0.44513$.

In the following statements, the CENTROID option in the PROC VARCLUS statement specifies that cluster centroids be used as the basis for clustering.

```
proc varclus data=phys8 centroid;
run;
```

Output 59.1.4. Centroid Cluster Components: Cluster Summary

```

Oblique Centroid Component Cluster Analysis

      Cluster summary for 1 cluster
      Cluster  Variation  Proportion
Cluster  Members  Variation  Explained  Explained
-----
      1           8           8      4.631      0.5789

Total variation explained = 4.631 Proportion = 0.5789

      Cluster summary for 2 clusters
      Cluster  Variation  Proportion
Cluster  Members  Variation  Explained  Explained
-----
      1           4           4      3.509      0.8773
      2           4           4      2.91       0.7275

Total variation explained = 6.419 Proportion = 0.8024

      R-squared with
      -----
      Own      Next      1-R**2      Variable
Cluster  Variable  Cluster  Closest      Ratio      Label
-----
Cluster 1  height      0.8778      0.2075      0.1543      Height
           arm_span  0.8994      0.1669      0.1208      Arm Span
           forearm   0.8663      0.1410      0.1557      Length of Forearm
           low_leg   0.8658      0.1824      0.1641      Length of Lower Leg
-----
Cluster 2  weight      0.8368      0.1975      0.2033      Weight
           bit_diam  0.7335      0.1341      0.3078      Bitrochanteric Diameter
           girth     0.6988      0.0929      0.3321      Chest Girth
           width     0.6473      0.1618      0.4207      Chest Width
    
```

The first cluster component, which, in the centroid method, is an unweighted sum of the standardized variables, explains 57.89% of the variation in the data. This value is near the maximum possible variance explained, 58.41%, which is attained by the first principal component (Output 59.1.1).

The centroid clustering algorithm splits the variables into the same two clusters created in the principal component method. Recall that this outcome was suggested by the similar standardized scoring coefficients in the principal cluster component solution.

The default behavior in the centroid method is to split any cluster with less than 75% of the total cluster variance explained by the centroid component. In the next step, the second cluster, with a component that explains only 72.75% of the total variation of the cluster, is split.

In the R-squared table for two clusters, the width variable has a weaker relation to its cluster than any other variable; in the three cluster solution this variable is in a cluster of its own.

Output 59.1.5. Standardized Scoring Coefficients

Oblique Centroid Component Cluster Analysis					
Standardized Scoring Coefficients					
Cluster			1	2	
height	Height		0.266918	0.000000	
arm_span	Arm Span		0.266918	0.000000	
forearm	Length of Forearm		0.266918	0.000000	
low_leg	Length of Lower Leg		0.266918	0.000000	
weight	Weight		0.000000	0.293105	
bit_diam	Bitrochanteric Diameter		0.000000	0.293105	
girth	Chest Girth		0.000000	0.293105	
width	Chest Width		0.000000	0.293105	

Each cluster component (Output 59.1.5) is an unweighted average of the cluster's standardized variables. Thus, the coefficients for each of the cluster's associated variables are identical in the centroid cluster component solution.

Output 59.1.6. Cluster Summary for Three Clusters

Oblique Centroid Component Cluster Analysis					
Cluster summary for 3 clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	
1	4	4	3.509	0.8773	
2	3	3	2.383333	0.7944	
3	1	1	1	1.0000	
Total variation explained = 6.892333 Proportion = 0.8615					
R-squared with					
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	height	0.8778	0.1921	0.1513	Height
	arm_span	0.8994	0.1722	0.1215	Arm Span
	forearm	0.8663	0.1225	0.1524	Length of Forearm
	low_leg	0.8658	0.1668	0.1611	Length of Lower Leg
Cluster 2	weight	0.8685	0.3956	0.2175	Weight
	bit_diam	0.7691	0.3329	0.3461	Bitrochanteric Diameter
	girth	0.7482	0.2905	0.3548	Chest Girth
Cluster 3	width	1.0000	0.4259	0.0000	Chest Width

The centroid method stops at the three cluster solution. As displayed in Output 59.1.6 and Output 59.1.7, the three centroid components account for 86.15% of the variability in the eight variables, and all cluster components account for at least 79.44% of the total variation in the corresponding cluster. Additionally, the smallest correlation between the variables and their own cluster component is 0.7482.

Output 59.1.7. Cluster Quality Table

Oblique Centroid Component Cluster Analysis					
Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.631000	0.5789	0.5789	0.4306	
2	6.419000	0.8024	0.7275	0.6473	0.4207
3	6.892333	0.8615	0.7944	0.7482	0.3548

Note that, if the proportion option were set to a value between 0.5789 (the proportion of variance explained in the 1-cluster solution) and 0.7275 (the minimum proportion of variance explained in the 2-cluster solution), PROC VARCLUS would stop at a two cluster solution, and the centroid solution would find the same clusters as the principal components solution.

In the following statements, the MAXC= option computes all clustering solutions, from one to eight clusters. The SUMMARY option suppresses all output except the final cluster quality table, and the OUTTREE= option saves the results of the analysis to an output data set and forces the clusters to be hierarchical. The TREE procedure is invoked to produce a graphical display of the clusters.

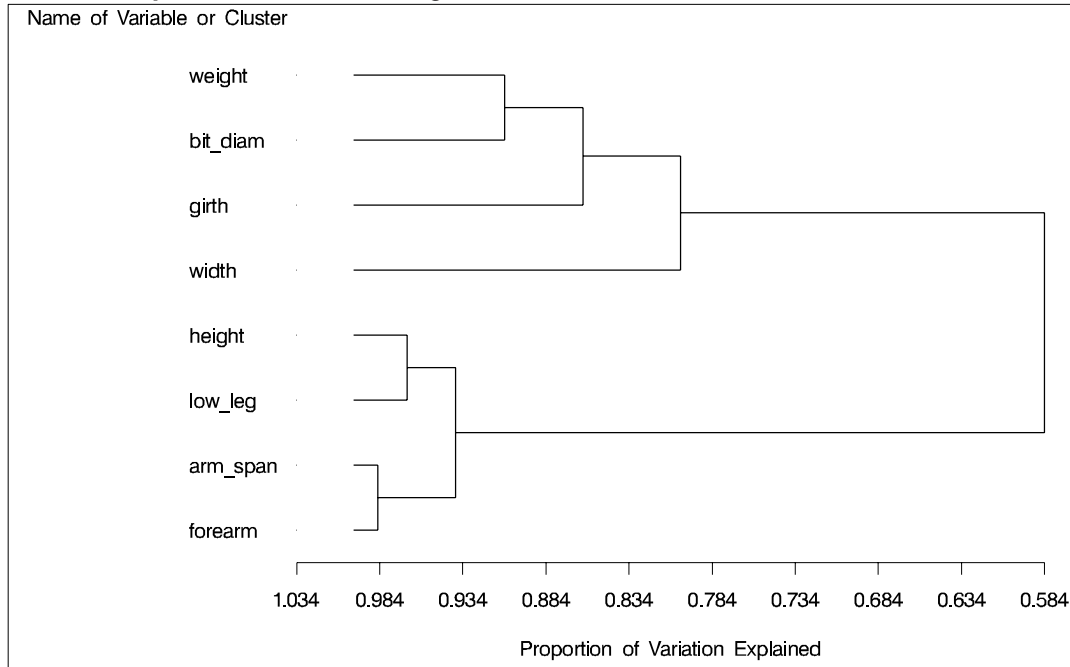
```
proc varclus data=phys8 maxc=8 summary outtree=tree;
run;

goptions ftext=swiss;
axis2 minor=none;
axis1 label=('Proportion of Variation Explained') minor=none;
proc tree horizontal vaxis=axis2 haxis=axis1 lines=(width=2);
    height _propor_;
run;
```

Output 59.1.8. Hierarchical Clusters and the SUMMARY Option

Oblique Principal Component Cluster Analysis						
Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.672880	0.5841	0.5841	1.770983	0.3810	
2	6.426502	0.8033	0.7293	0.476418	0.6329	0.4380
3	6.895347	0.8619	0.7954	0.418369	0.7421	0.3634
4	7.271218	0.9089	0.8773	0.238000	0.8652	0.2548
5	7.509218	0.9387	0.8773	0.236135	0.8652	0.1665
6	7.740000	0.9675	0.9295	0.141000	0.9295	0.2560
7	7.881000	0.9851	0.9405	0.119000	0.9405	0.2093
8	8.000000	1.0000	1.0000	0.000000	1.0000	0.0000

The principal component method first separates the variables into the same two clusters that were created in the first PROC VARCLUS run. Note that, in creating the third cluster, the principal component method identifies the variable width. This is the same variable that is put into its own cluster in the preceding centroid method example.

Output 59.1.9. TREE Diagram from PROC TREE

The tree diagram in Output 59.1.9 displays the cluster hierarchy. It is clear from the diagram that there are two, or possibly three, clusters present. However, the MAXC=8 option forces PROC VARCLUS to split the clusters until each variable is in its own cluster.

References

- Anderberg, M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press, Inc.
- Harman, H.H. (1976), *Modern Factor Analysis*, Third Edition, Chicago: University of Chicago Press.
- Hand, D.J.; Daly, F.; Lunn, A.D.; McConway, K.J.; and Ostrowski E. (1994), *A Handbook of Small Data Sets*, London: Chapman & Hall, 297–298.