SAS/STAT® 13.1 User's Guide
The VARCLUS Procedure

This document is an individual chapter from *SAS/STAT® 13.1 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2013. *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

December 2013

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** or call 1-800-727-3228.

Chapter 104
The VARCLUS Procedure

Contents

Overview: VARCLUS Procedure

The VARCLUS procedure divides a set of numeric variables into disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster. This linear combination can be either the first principal component (the default) or the centroid component (if you specify the CENTROID option). The first principal component is a weighted average of the variables that explains as much variance as possible. See Chapter 77, "The PRINCOMP Procedure," for further details. Centroid components are unweighted averages of either the standardized variables (the default) or the raw variables (if you specify the COVARIANCE option). PROC VARCLUS tries to maximize the variance that is explained by the cluster components, summed over all the clusters.

The cluster components are oblique, not orthogonal, even when the cluster components are first principal components. In an ordinary principal component analysis, all components are computed from the same variables, and the first principal component is orthogonal to the second principal component and to every other principal component. In PROC VARCLUS, each cluster component is computed from a set of variables that is different from all the other cluster components. The first principal component of one cluster might be correlated with the first principal component of another cluster. Hence, the PROC VARCLUS algorithm is a type of oblique component analysis.

As in principal component analysis, 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 displays a dendrogram (tree diagram of hierarchical clusters) by using ODS Graphics. PROC VARCLUS can also can create an output data set that can be used by the TREE procedure to draw the dendrogram. A second output data set can be used with the SCORE procedure to compute component scores for each cluster.

PROC VARCLUS 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 50 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.

The VARCLUS algorithm is both divisive and iterative. 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 PROPORTION= 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; Harris and Kaiser 1964), and assigning each variable to the rotated component with which it has the higher squared correlation.

3. Variables are iteratively reassigned to clusters to try to maximize the variance accounted for by the cluster components. You can require the reassignment algorithms to maintain a hierarchical structure for the clusters.

The procedure stops splitting when either of the following conditions holds:

- The number of clusters is greater than or equal to the maximum number of clusters as specified by the MAXCLUSTERS= option is reached.
- Every cluster satisfies the stopping criteria specified by the PROPORTION= option (percentage of variation explained) or the MAXEIGEN= option (second eigenvalue) or both.

By default, VARCLUS stops splitting when every cluster has only one eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying 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.

If principal components are used, the NCS phase is an alternating least squares method and converges rapidly. The search phase can be very time-consuming for a large number of variables. But if the default initialization method is used, the search phase is rarely able to substantially improve the results of the NCS phase, so the search takes few iterations. If random initialization is used, the NCS phase might 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 might not increase the amount of variance explained; therefore it is limited, by default, to one iteration.

You can have VARCLUS do the clustering hierarchically by restricting the reassignment of variables such that the clusters maintain a tree structure. In this case, when a cluster is split, a variable in one of the two resulting clusters can be reassigned to the other cluster that results from the split but not to a cluster that is not part of the original cluster (the one that is split).

Getting Started: VARCLUS Procedure

This example demonstrates how you can use PROC VARCLUS to cluster variables.

The following data are job ratings of police officers. The officers were rated by their supervisors on 13 job skills on a scale from 1 to 9. There is also an overall rating that is not used in this analysis. The following DATA step creates the SAS data set JobRat:

```
data JobRat;
   input
      (Communication_Skills
       Problem_Solving
       Learning_Ability
       Judgement_under_Pressure
       Observational_Skills
       Willingness_to_Confront_Problems
       Interest_in_People
       Interpersonal_Sensitivity
       Desire_for_Self_Improvement
       Appearance
       Dependability
       Physical_Ability
       Integrity
       Overall_Rating)
      (1.);
   datalines;
26838853879867
74758876857667
56757863775875
67869777988997
99997798878888

   ... more lines ...

99997899799799
99899899899899
76656399567486
;
```

The following statements cluster the variables:

```
proc varclus data=JobRat maxclusters=3;
   var Communication_Skills--Integrity;
run;
```

The DATA= option specifies the SAS data set JobRat as input.

The MAXCLUSTERS=3 option specifies that no more than three clusters be computed. By default, PROC VARCLUS splits and optimizes clusters until all clusters have a second eigenvalue less than one. In this example, the default setting would produce only two clusters, but going to three clusters produces a more interesting result.

The VAR statement lists the numeric variables (Communication_Skills -- Integrity) to be used in the analysis. The overall rating is omitted from the list of variables.

Although PROC VARCLUS displays output for one cluster, two clusters, and three clusters, the following figures display only the final analysis for three clusters.

For each cluster, Figure 104.1 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 104.1 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, 0.6715, indicates that about 67% of the total variation in the data can be accounted for by the three cluster components.

Figure 104.1 Cluster Summary for Three Clusters from PROC VARCLUS

Oblique Principal Component Cluster Analysis

Cluster Summary for 3 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	6	6	3.771349	0.6286	0.7093
2	5	5	3.575933	0.7152	0.5035
3	2	2	1.382005	0.6910	0.6180

Total variation explained = 8.729286 Proportion = 0.6715

Figure 104.2 shows how the variables are clustered. Figure 104.2 also displays the R-square value of each variable with its own cluster and the R-square value with its nearest cluster. The R-square value for a variable with the nearest cluster should be low if the clusters are well separated. The last column displays the ratio of $(1 - R^2_{own})/(1 - R^2_{nearest})$ for each variable. Small values of this ratio indicate good clustering.

Figure 104.2 R-Square Values from PROC VARCLUS

3 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	Communication_Skills	0.6403	0.3599	0.5620
	Problem_Solving	0.5412	0.2895	0.6458
	Learning_Ability	0.6561	0.1692	0.4139
	Observational_Skills	0.6889	0.2584	0.4194
	Willingness_to_Confront_Problems	0.6480	0.3402	0.5335
	Desire_for_Self_Improvement	0.5968	0.3473	0.6177
Cluster 2	Judgement_under_Pressure	0.6263	0.3719	0.5950
	Interest_in_People	0.8122	0.1885	0.2314
	Interpersonal_Sensitivity	0.7566	0.1387	0.2826
	Dependability	0.6163	0.4419	0.6875
	Integrity	0.7645	0.2724	0.3237
Cluster 3	Appearance	0.6910	0.3047	0.4444
	Physical_Ability	0.6910	0.1871	0.3801

Figure 104.3 displays the standardized scoring coefficients that are used to compute the first principal component of each cluster. Since each variable is assigned to one and only one cluster, each row of the scoring coefficients contains only one nonzero value.

Figure 104.3 Standardized Scoring Coefficients from PROC VARCLUS

Standardized Scoring Coefficients

Cluster	1	2	3
Communication_Skills	0.212170	0.000000	0.000000
Problem_Solving	0.195058	0.000000	0.000000
Learning_Ability	0.214781	0.000000	0.000000
Judgement_under_Pressure	0.000000	0.221313	0.000000
Observational_Skills	0.220086	0.000000	0.000000
Willingness_to_Confront_Problems	0.213452	0.000000	0.000000
Interest_in_People	0.000000	0.252025	0.000000
Interpersonal_Sensitivity	0.000000	0.243245	0.000000
Desire_for_Self_Improvement	0.204848	0.000000	0.000000
Appearance	0.000000	0.000000	0.601493
Dependability	0.000000	0.219544	0.000000
Physical_Ability	0.000000	0.000000	0.601493
Integrity	0.000000	0.244507	0.000000

Figure 104.4 displays the cluster structure and the intercluster correlations. The structure table displays the correlation of each variable with each cluster component. The table of intercorrelations contains the correlations between the cluster components.

Figure 104.4 Cluster Correlations and Intercorrelations from PROC VARCLUS

Cluster Structure

Cluster	1	2	3
Communication_Skills	0.800169	0.599909	0.427341
Problem_Solving	0.735630	0.538017	0.425463
Learning_Ability	0.810014	0.411316	0.376333
Judgement_under_Pressure	0.609876	0.791401	0.345399
Observational_Skills	0.830021	0.407807	0.508305
Willingness_to_Confront_Problems	0.805002	0.362927	0.583265
Interest_in_People	0.434138	0.901225	0.387770
Interpersonal_Sensitivity	0.372371	0.869826	0.287658
Desire_for_Self_Improvement	0.772554	0.589334	0.494842
Appearance	0.552003	0.393759	0.831266
Dependability	0.664778	0.785073	0.574460
Physical_Ability	0.432590	0.416070	0.831266
Integrity	0.521876	0.874342	0.477885

Figure 104.4 *continued*

Inter-Cluster Correlations

Cluster	1	2	3
1	1.00000	0.60851	0.59223
2	0.60851	1.00000	0.48711
3	0.59223	0.48711	1.00000

PROC VARCLUS next displays the summary table of statistics for the cluster history (Figure 104.5). The first three columns give the number of clusters, the total variation explained by clusters, and the proportion of variation explained by clusters, respectively.

As displayed in the first row of Figure 104.5, the variation explained by the first principal component of all the variables is 6.547402, and the proportion of variation explained is 0.5036.

When the number of clusters is two, the total variation explained is 7.96775 and the proportion of variation explained by the two clusters is 0.6129. The larger second eigenvalue of the clusters is 0.937902; so by default, PROC VARCLUS would stop splitting clusters at this point. But because the MAXCLUSTERS=3 option was specified in this example, PROC VARCLUS continues to the three-cluster solution.

When the number of clusters increases to three, the total variation explained is 8.729286 and the proportion of variation explained by the two clusters is 0.6715. The largest second eigenvalue of the clusters is 0.709323. The statistical improvement from increasing the number of clusters from two to three seems modest, but the interpretability of the three clusters argues for the three-cluster solution.

Figure 104.5 also displays the minimum proportion of variance explained by a cluster, the minimum R square for a variable, and the maximum $(1 - R^2)$ ratio for a variable. The last quantity is the maximum ratio of the value $1 - R^2$ for a variable's own cluster to the value $1 - R^2$ for its nearest cluster.

Figure 104.5 Final Cluster Summary Table from PROC VARCLUS

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	6.547402	0.5036	0.5036	1.772715	0.2995	
2	7.967753	0.6129	0.5475	0.937902	0.3123	0.8026
3	8.729286	0.6715	0.6286	0.709323	0.5412	0.6875

Syntax: VARCLUS Procedure

The following statements are available in the VARCLUS procedure:

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 about 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 invokes the VARCLUS procedure. By default, VARCLUS clusters the numeric variables in the most recently created SAS data set, starting with one cluster and splitting clusters until all clusters have at most one eigenvalue greater than one.

Table 104.1 summarizes the *options* available in the PROC VARCLUS statement.

Table 104.1 Options Available in the PROC VARCLUS Statement

Option	Description
Data Sets	
DATA=	Specifies the input SAS data set
OUTSTAT=	Specifies the output SAS data set to contain statistics
OUTTREE=	Specifies the output SAS data set for use with PROC TREE
Input Data Processing	
COVARIANCE	Uses the covariance matrix instead of the correlation matrix
NOINT	Omits the intercept
VARDEF=	Specifies the divisor for variances
Number of Clusters	
MAXCLUSTERS=	Specifies the maximum number of clusters
MINCLUSTERS=	Specifies the minimum number of clusters
MAXEIGEN=	Specifies the maximum second eigenvalue in a cluster
PROPORTION=	Specifies the minimum proportion of variance explained by a cluster component
Clustering Methods	
CENTROID	Uses centroid components instead of principal components
HIERARCHY	Clusters hierarchically
INITIAL=	Specifies the initialization method

Table 104.1 *continued*

Option	Description
MAXITER=	Specifies the maximum iterations during the alternating least squares phase
MAXSEARCH=	Specifies the maximum iterations during the search phase
MULTIPLEGROUP	Performs a multiple group component analysis
RANDOM=	Specifies the random number seed
Control Displayed Output	
CORR	Displays the correlation matrix
NOPRINT	Suppresses displayed output
PLOTS=	Specifies ODS Graphics details
SHORT	Suppresses display of large matrices
SIMPLE	Displays means and standard deviations
SUMMARY	Suppresses all default displayed output except the final summary table
TRACE	Displays the cluster to which each variable is assigned during the iterations

VARCLUS chooses which cluster to split based on the MAXEIGEN= and PROPORTION= options.

1. If you specify *either* or *both* of these two options, then *only* the specified options affect the choice of the cluster to split.

2. If you specify *neither* of these options, the criterion for choice of cluster to split depends on the CENTROID option:

   a) If you specify CENTROID, VARCLUS splits the cluster with the smallest percentage of variation explained by its cluster component, as if you had specified the PROPORTION= option.

   b) If you do not specify CENTROID, VARCLUS splits the cluster with the largest eigenvalue associated with the second principal component, as if you had specified the MAXEIGEN= option.

The final number of clusters is controlled by three options: MAXCLUSTERS=, MAXEIGEN=, and PROPORTION=.

1. If you specify *any* of these three options, then *only* the options you specify affect the final number of clusters.

2. If you specify *none* of these options, VARCLUS continues to split clusters until the default splitting criterion is satisfied. The default splitting criterion depends on the CENTROID option:

   a) If you specify CENTROID, the default splitting criterion is PROPORTION=0.75.

   b) If you do not specify CENTROID, splitting is based on the MAXEIGEN= criterion, with a default depending on the COVARIANCE option:

      i. For analyzing a correlation matrix (no COVARIANCE option), the default value for MAXEIGEN= is one.

      ii. For analyzing a covariance matrix (using the COVARIANCE option), the default value for MAXEIGEN= is the average variance of the variables being clustered.

VARCLUS continues to split clusters until any of the following conditions holds:

- The number of cluster equals the value specified for MAXCLUSTERS=.
- No cluster qualifies for splitting according to the MAXEIGEN= or PROPORTION= criterion.
- A cluster was chosen for splitting, but after iteratively reassigning variables to clusters, one of the cluster has no members.

The following list gives details about the *options*.

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 COVARIANCE 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 both the CENTROID and MAXEIGEN= options.

CORR
C
displays the correlation matrix.

COVARIANCE
COV
analyzes the covariance matrix instead of the correlation matrix. The COVARIANCE option causes variables with a large variance to have more effect on the cluster components than variables with a small variance.

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 about types of SAS data sets.

HIERARCHY
HI
requires the clusters at different levels to maintain a hierarchical structure. To draw a tree diagram, enable ODS Graphics or use the OUTTREE= option and the TREE procedure.

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 (raw quartimax rotation on the eigenvectors; Harris and Kaiser 1964). The following list describes the values for the INITIAL= option:

GROUP obtains the cluster membership of each variable from an observation in the DATA= data set where the _TYPE_ variable has a value of 'GROUP'. In this observation, the variables to be clustered must each have an integer value ranging from one to the number of clusters. You can use this option only if the DATA= data set is a TYPE=CORR, UCORR, COV, UCOV, or FACTOR data set. You can use a data set created either by a previous run of PROC VARCLUS or in a DATA step.

INPUT obtains scoring coefficients for the cluster components from observations in the DATA= data set where the _TYPE_ variable has a value of 'SCORE'. You can use this option only if the DATA= data set is a TYPE=CORR, UCORR, COV, UCOV, or FACTOR data set. 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.

SEED initializes each cluster component to be one of 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 are initially 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. VARCLUS stops splitting clusters after the number of clusters reaches the value of the MAXCLUSTERS= option, regardless of what other splitting options are specified.

MAXEIGEN=*n*

specifies that when choosing a cluster to split, VARCLUS should choose the cluster with the largest second eigenvalue, provided that its second eigenvalue is greater than the MAXEIGEN= value. The MAXEIGEN= option cannot be used with the CENTROID or MULTIPLEGROUP options.

If you do not specify MAXEIGEN=, the default behavior depends on other options as follows:

- If you specify PROPORTION=, CENTROID, or MULTIPLEGROUP, cluster splitting does not depend on the second eigenvalue.
- Otherwise, if you specify MAXCLUSTERS=, the default value for MAXEIGEN= is zero.
- Otherwise, the default value for MAXEIGEN= is either 1.0 if the correlation matrix is analyzed or the average variance if the COVARIANCE option is specified.

If you specify both MAXEIGEN= and MAXCLUSTERS=, the number of clusters will never exceed the value of the MAXCLUSTERS= option.

If you specify both MAXEIGEN= and PROPORTION=, VARCLUS first looks for a cluster to split based on the MAXEIGEN= criterion. If no cluster meets that criterion, VARCLUS then looks for a cluster to split based on the PROPORTION= criterion.

MAXITER=*n*

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

MAXSEARCH=*n*

specifies the maximum number of iterations during the search phase. The default is 1,000 divided by the number of variables.

MINCLUSTERS=*n*

MINC=*n*

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

MULTIPLEGROUP

MG

performs a multiple group component analysis (Harman 1976). You specify which variables belong to which clusters. No clusters are split, and no variables are reassigned to a different cluster. The input data set must be TYPE=CORR, UCORR, COV, UCOV, FACTOR, or SSCP and must contain an observation with _TYPE_='GROUP' that defines the variable groups. Specifying the MULTIPLEGROUP option is equivalent to specifying all of the following options: INITIAL=GROUP, MINC=1, MAXITER=0, MAXSEARCH=0, PROPORTION=0, and MAXEIGEN=large number.

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 displayed output. This option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, "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. The OUTSTAT= data set is TYPE=UCORR if the NOINT option is specified. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Language Reference: Concepts*. For information about types of SAS data sets, see Appendix A, "Special SAS Data Sets."

OUTTREE=*SAS-data-set*

creates an output data set to contain information about the tree structure that can be used by the TREE procedure to display a tree diagram. The OUTTREE= option implies the HIERARCHY option. See Example 104.1 for use of the OUTTREE= option. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Language Reference: Concepts*.

PLOTS < (*global-plot-options***) > < = *plot-request* >**

PLOTS < (*global-plot-options*) > < = (*plot-request* < ... *plot-request* >) >
controls the plots produced through ODS Graphics.

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;

proc varclus plots=dendrogram(height=ncl);
run;

ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

By default, PROC VARCLUS produces a dendrogram.

The *global-plot-options*, UNPACK and ONLY, that are commonly used in the PLOTS= option in other procedures are accepted in PROC VARCLUS, but they currently have no effect since PROC VARCLUS produces only a dendrogram.

The following *plot-requests* can be specified:

ALL
produces all plots, which for PROC VARCLUS is only a dendrogram.

MAXPOINTS=*n*

MAXPTS=*n*
suppresses the dendrogram when the number of variables (clusters) exceeds the *n* value. This prevents an unreadable plot from being produced. The default is MAXPOINTS=200.

DENDROGRAM < (*dendrogram-options*) >
requests a dendrogram and specifies *dendrogram-options*.

Unlike most graphs, the size of the dendrogram can vary as a function of the number of objects that appear in the dendrogram. You can specify the following *dendrogram-options* to control the size and appearance of the dendrogram:

COMPUTEHEIGHT=*a b*

CH=*a b*
specifies the constants for computing the height of the dendrogram. For *n* points being clustered, intercept *a*, and slope *b*, the height is based in part on $a + bn$. For a horizontal dendrogram, the default (given in pixels) is COMPUTEHEIGHT=100 12, the default height in pixels is max(100 + 12n, 480), the default height in inches is max(1.04167 + 0.125n, 5), and the default height in centimeters is max(2.64583 + 0.3175n, 12.7). For a vertical dendrogram, the default height is 480 pixels. The default unit is pixels, and you can use the UNIT= *dendrogram-option* to change the unit to inches or centimeters for this option. Inches equals pixels divided by 96, and centimeters equals inches times 2.54.

COMPUTEWIDTH=*a b*

CW=*a b*

specifies the constants for computing the width of the dendrogram. For n points being clustered, intercept *a*, and slope *b*, the width is based in part on $a + bn$. For a vertical dendrogram, the default (given in pixels) is COMPUTEWIDTH=100 12, the default width in pixels is max(100 + 12n, 640), the default width in inches is max(1.04167 + 0.125n, 6.66667), and the default width in centimeters is max(2.64583 + 0.3175n, 16.933). For a horizontal dendrogram, the default width is 640 pixels. The default unit is pixels, and you can use the UNIT= *dendrogram-option* to change the unit to inches or centimeters for this option. Inches equals pixels divided by 96, and centimeters equals inches times 2.54.

HEIGHT=PROPORTION | NCL | VAREXP

H=P | N | V

specifies the method for drawing the height of the dendrogram. HEIGHT=PROPORTION is the default.

HEIGHT=PROPORTION specifies that the total proportion of variance explained by the clusters at the current level of the tree is used.

HEIGHT=NCL specifies that the number of clusters is used.

HEIGHT=VAREXP specifies that the total variance explained by the clusters at the current level of the tree is used.

HORIZONTAL | VERTICAL

specifies either a horizontal dendrogram with the objects on the vertical axis (HORIZONTAL) or a vertical dendrogram with the objects on the horizontal axis (VERTICAL). The default is HORIZONTAL.

SETHEIGHT=*height*

SH=*height*

specifies the height of the dendrogram. By default, the height is based on the COMPUTEHEIGHT= option. The default unit is pixels, and you can use the UNIT= *dendrogram-option* to change the unit to inches or centimeters for this *dendrogram-option*.

SETWIDTH=*width*

SW=*width*

specifies the width of the dendrogram. By default, the width is based on the COMPUTEWIDTH= option. The default unit is pixels, and you can use the UNIT= *dendrogram-option* to change the unit to inches or centimeters for this *dendrogram-option*.

UNIT=PX | IN | CM

specifies the unit (pixels, inches, or centimeters) for the SETHEIGHT=, SETWIDTH=, COMPUTEHEIGHT=, and COMPUTEWIDTH= *dendrogram-options*.

NONE

suppresses all plots.

The names of the graphs that PROC VARCLUS generates are listed in Table 104.4, along with the required statements and options.

PROPORTION=*n*

PERCENT=*n*

specifies that when choosing a cluster to split, VARCLUS should choose the cluster with the smallest proportion of variation explained, provided that the proportion of variation explained is less than the PROPORTION= value. Values greater than 1.0 are considered to be percentages, so PROPORTION=0.75 and PERCENT=75 are equivalent.

However, if you specify both MAXEIGEN= and PROPORTION=, VARCLUS first looks for a cluster to split based on the MAXEIGEN= criterion. If no cluster meets that criterion, VARCLUS then looks for a cluster to split based on the PROPORTION= criterion.

If you do not specify PROPORTION=, the default behavior depends on other options as follows:

- If you specify MAXEIGEN=, cluster splitting does not depend on the proportion of variation explained.
- Otherwise, if you specify CENTROID and MAXCLUSTERS=, the default value for PROPORTION= is 1.0.
- Otherwise, if you specify CENTROID without MAXCLUSTERS=, the default value is PROPORTION=0.75 or PERCENT=75.
- Otherwise, cluster splitting does not depend on the proportion of variation explained.

If you specify both PROPORTION= and MAXCLUSTERS=, the number of clusters will never exceed the value of the MAXCLUSTERS= option.

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 pseudorandom number sequence.

SHORT

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

SIMPLE

S

displays means and standard deviations.

SUMMARY

suppresses all default displayed output except the final summary table.

TRACE

displays 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 of observations in groups that are 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 you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base 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 *variables* **;**

If you want to base the clustering on partial correlations, list the variables to be partialed 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: VARCLUS Procedure

Missing Values

Observations that contain missing values are omitted from the analysis.

Using the VARCLUS procedure

Default options for PROC VARCLUS often provide satisfactory results. If you want to change the final number of clusters, use one or more of the MAXCLUSTERS=, MAXEIGEN=, or PROPORTION= options. 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 might 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 might want to try one of the following methods to obtain a better solution:

- If the clustering algorithm has not converged, specify larger values for MAXITER= and MAXSEARCH=.
- 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 OUSTAT= data set contains the following variables:

- BY variables
- _NCL_, a numeric variable that gives the number of clusters
- _TYPE_, a character variable that indicates the type of statistic the observation contains
- _NAME_, a character variable that contains 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 104.2 _TYPE_

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 as follows:

```
data Coef2;
   set Coef;
   if _ncl_ = . or _ncl_ = 3;
   drop _ncl_;
run;

proc score data=NewScore score=Coef2;
run;
```

PROC SCORE standardizes the new data by subtracting the original variable means that are stored in the _TYPE_='MEAN' observations and dividing by the original variable standard deviations from the _TYPE_='STD' observations. Then PROC SCORE multiplies the standardized variables by the coefficients from the _TYPE_='SCORE' observations to get the cluster scores.

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 OUTTREE= data set contains the following variables:

- BY variables, if any
- _NAME_, a character variable that gives the name of the node. If the node is a cluster, the name is CLUSn, where n is the number of the cluster. If the node is a single variable, the variable name is used.
- _PARENT_, a character variable that gives the value of _NAME_ of the parent of the node. If the node is the root of the tree, _PARENT_ is blank.
- _LABEL_, a character variable that gives the label of the node. If the node is a cluster, the label is CLUSn, where n is the number of the cluster. If the node is a single variable, the variable label is used.
- _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 the VARCLUS algorithm 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 that contains 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 omitted, the second largest eigenvalue of the cluster is also displayed.

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 labeled "1–R**2 Ratio" gives the ratio of one minus the "Own Cluster" R square to one minus the "Next Closest" R square. 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 following columns:

- 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 the following:

- 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 the following columns:

- Variable, the variables in each cluster
- R square with Own Cluster (the squared correlation of the variable with its own cluster component), and R square with Next Closest (the next highest squared correlation of the variable with a cluster component). Own Cluster values should be higher than the R square 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 the following tables:

- 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 the following:

- 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 square 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 this name to refer to the table when using the Output Delivery System (ODS) to select tables and create output data sets. These ODS table names are listed in Table 104.3. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 104.3 ODS Tables Produced by 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 between variables	CORR
DataOptSummary	Data and options summary table	default
InterClusterCorr	Correlations between cluster components	default
IterHistory	Iteration history	TRACE
RSquare	R squares between variables and clusters	default
SimpleStatistics	Means and standard deviations	SIMPLE
StdScoreCoef	Standardized scoring coefficients	default

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 605 in Chapter 21, "Statistical Graphics Using ODS."

By default, PROC VARCLUS produces a dendrogram.

You can refer to every graph produced through ODS Graphics with a name. The name of the graph that PROC VARCLUS generates is listed in Table 104.4, along with the statement and option required to produce it.

Table 104.4 Graphs Produced by PROC VARCLUS

ODS Graph Name	Plot Description	Statement and Option
Dendrogram	Dendrogram (tree diagram)	PROC VARCLUS PLOTS=DENDROGRAM

Example: VARCLUS Procedure

Example 104.1: Correlations among Physical Variables

The data in this example 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 display a tree diagram. The following statements create the data set and perform the analysis:

```
data phys8(type=corr);
   title 'Eight Physical Measurements on 305 School Girls';
   title2 'Harman: Modern Factor Analysis, 3rd Ed, p22';
   label ArmSpan='Arm Span'              Forearm='Length of Forearm'
         LowerLeg='Length of Lower Leg' BitDiam='Bitrochanteric Diameter'
         Girth='Chest Girth'            Width='Chest Width';
   input _Name_ $ 1-8
         (Height ArmSpan Forearm LowerLeg Weight BitDiam
          Girth Width)(7.);
   _Type_='corr';
   datalines;
Height     1.0    .846   .805   .859   .473   .398   .301   .382
ArmSpan    .846   1.0    .881   .826   .376   .326   .277   .415
Forearm    .805   .881   1.0    .801   .380   .319   .237   .345
LowerLeg   .859   .826   .801   1.0    .436   .329   .327   .365
Weight     .473   .376   .380   .436   1.0    .762   .730   .629
BitDiam    .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 using principal components.

As displayed in Output 104.1.1, when there is only one cluster, the cluster component (by default, the first principal component) explains 58.41% of the total variation of the eight 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, ArmSpan, Forearm, and LowerLeg have been assigned to the first cluster, and that the variables Weight, BitDiam, Girth, and Width have been assigned to the second cluster.

The standardized scoring coefficients in Output 104.1.1 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.

The intercluster correlation table shows that the two cluster components have a moderate correlation of 0.44513.

Output 104.1.1 Principal Component Clusters

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

                Oblique Principal Component Cluster Analysis

            Observations          10000    Proportion          0
            Variables                 8    Maxeigen            1


  Clustering algorithm converged.


                        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 because it has the largest second eigenvalue, 1.770983,
which is greater than the MAXEIGEN=1 value.


  Clustering algorithm converged.
```

Output 104.1.1 *continued*

Cluster Summary for 2 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second 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

2 Clusters		R-squared with			
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	ArmSpan	0.9002	0.1658	0.1196	Arm Span
	Forearm	0.8661	0.1413	0.1560	Length of Forearm
	LowerLeg	0.8652	0.1829	0.1650	Length of Lower Leg
	Height	0.8777	0.2088	0.1545	
Cluster 2	BitDiam	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
	Weight	0.8477	0.1974	0.1898	

Standardized Scoring Coefficients

Cluster		1	2
ArmSpan	Arm Span	0.270377	0.000000
Forearm	Length of Forearm	0.265194	0.000000
LowerLeg	Length of Lower Leg	0.265057	0.000000
BitDiam	Bitrochanteric Diameter	0.000000	0.294591
Girth	Chest Girth	0.000000	0.286407
Width	Chest Width	0.000000	0.272710
Height		0.266977	0.000000
Weight		0.000000	0.315597

Cluster Structure

Cluster		1	2
ArmSpan	Arm Span	0.948813	0.407210
Forearm	Length of Forearm	0.930624	0.375865
LowerLeg	Length of Lower Leg	0.930142	0.427715
BitDiam	Bitrochanteric Diameter	0.366201	0.859404
Girth	Chest Girth	0.304779	0.835529
Width	Chest Width	0.402430	0.795572
Height		0.936881	0.456908
Weight		0.444281	0.920686

Output 104.1.1 *continued*

```
                    Inter-Cluster Correlations

                  Cluster             1             2

                  1             1.00000       0.44513
                  2             0.44513       1.00000


No cluster meets the criterion for splitting.


             Total   Proportion     Minimum     Maximum                 Maximum
          Variation           of  Proportion      Second     Minimum    1-R**2
   Number Explained    Variation   Explained  Eigenvalue  R-squared     Ratio
       of        by    Explained        by a        in a       for a     for a
 Clusters  Clusters  by Clusters     Cluster     Cluster   Variable  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
```

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;
```

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 shown previously in Output 104.1.1.

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. Since the centroid component for the one-cluster solution explains only 57.89% of the variation as shown in Output 104.1.2, the variables are split into two clusters. The resulting clusters are the same two clusters created by the principal component method. Recall that this outcome was suggested by the similar standardized scoring coefficients in the principal cluster component solution.

In the two-cluster solution, the centroid component of the second cluster explains only 72.75% of the total variation of the cluster. Since this percentage is less than 75%, the second cluster is split.

In the R-square 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.

Each cluster component 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.

The centroid method stops at the three-cluster solution. 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 squared correlation between the variables and their own cluster component is 0.7482.

If the PROPORTION= option were set to a value between 0.5789 (the proportion of variance explained in the one-cluster solution) and 0.7275 (the minimum proportion of variance explained in the two-cluster solution),

PROC VARCLUS would stop at the two-cluster solution, and the centroid solution would find the same clusters as the principal components solution, although the cluster components would be slightly different.

Output 104.1.2 Centroid Component Clusters

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

            Oblique Centroid Component Cluster Analysis

        Observations          10000    Proportion       0.75
        Variables                 8    Maxeigen            0


  Clustering algorithm converged.


                     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 1 will be split because it has the smallest proportion of variation
explained, 0.578875, which is less than the PROPORTION=0.75 value.


  Clustering algorithm converged.


                     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
```

Output 104.1.2 *continued*

2 Clusters		R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	ArmSpan	0.8994	0.1669	0.1208	Arm Span
	Forearm	0.8663	0.1410	0.1557	Length of Forearm
	LowerLeg	0.8658	0.1824	0.1641	Length of Lower Leg
	Height	0.8778	0.2075	0.1543	
Cluster 2	BitDiam	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
	Weight	0.8368	0.1975	0.2033	

Standardized Scoring Coefficients

Cluster		1	2
ArmSpan	Arm Span	0.266918	0.000000
Forearm	Length of Forearm	0.266918	0.000000
LowerLeg	Length of Lower Leg	0.266918	0.000000
BitDiam	Bitrochanteric Diameter	0.000000	0.293105
Girth	Chest Girth	0.000000	0.293105
Width	Chest Width	0.000000	0.293105
Height		0.266918	0.000000
Weight		0.000000	0.293105

Cluster Structure

Cluster		1	2
ArmSpan	Arm Span	0.948361	0.408589
Forearm	Length of Forearm	0.930744	0.375468
LowerLeg	Length of Lower Leg	0.930477	0.427054
BitDiam	Bitrochanteric Diameter	0.366212	0.856453
Girth	Chest Girth	0.304821	0.835936
Width	Chest Width	0.402246	0.804574
Height		0.936883	0.455485
Weight		0.444419	0.914781

Inter-Cluster Correlations

Cluster	1	2
1	1.00000	0.44484
2	0.44484	1.00000

Output 104.1.2 *continued*

Cluster 2 will be split because it has the smallest proportion of variation explained, 0.7275, which is less than the PROPORTION=0.75 value.

Clustering algorithm converged.

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

3 Clusters		R-squared with Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	ArmSpan	0.8994	0.1722	0.1215	Arm Span
	Forearm	0.8663	0.1225	0.1524	Length of Forearm
	LowerLeg	0.8658	0.1668	0.1611	Length of Lower Leg
	Height	0.8778	0.1921	0.1513	
Cluster 2	BitDiam	0.7691	0.3329	0.3461	Bitrochanteric Diameter
	Girth	0.7482	0.2905	0.3548	Chest Girth
	Weight	0.8685	0.3956	0.2175	
Cluster 3	Width	1.0000	0.4259	0.0000	Chest Width

Standardized Scoring Coefficients

Cluster		1	2	3
ArmSpan	Arm Span	0.26692	0.00000	0.00000
Forearm	Length of Forearm	0.26692	0.00000	0.00000
LowerLeg	Length of Lower Leg	0.26692	0.00000	0.00000
BitDiam	Bitrochanteric Diameter	0.00000	0.37398	0.00000
Girth	Chest Girth	0.00000	0.37398	0.00000
Width	Chest Width	0.00000	0.00000	1.00000
Height		0.26692	0.00000	0.00000
Weight		0.00000	0.37398	0.00000

Output 104.1.2 *continued*

Cluster Structure

Cluster		1	2	3
ArmSpan	Arm Span	0.94836	0.36613	0.41500
Forearm	Length of Forearm	0.93074	0.35004	0.34500
LowerLeg	Length of Lower Leg	0.93048	0.40838	0.36500
BitDiam	Bitrochanteric Diameter	0.36621	0.87698	0.57700
Girth	Chest Girth	0.30482	0.86501	0.53900
Width	Chest Width	0.40225	0.65259	1.00000
Height		0.93688	0.43830	0.38200
Weight		0.44442	0.93196	0.62900

Inter-Cluster Correlations

Cluster	1	2	3
1	1.00000	0.41716	0.40225
2	0.41716	1.00000	0.65259
3	0.40225	0.65259	1.00000

No cluster meets the criterion for splitting.

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

In the following statements, the MAXC= option computes all clustering solutions, from one to eight clusters, and the SUMMARY option suppresses all output except the final cluster quality table:

```
ods graphics on;

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

The results from PROC VARCLUS are shown in Output 104.1.3.

Output 104.1.3 Hierarchical Clusters and the SUMMARY Option

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

Oblique Principal Component Cluster Analysis

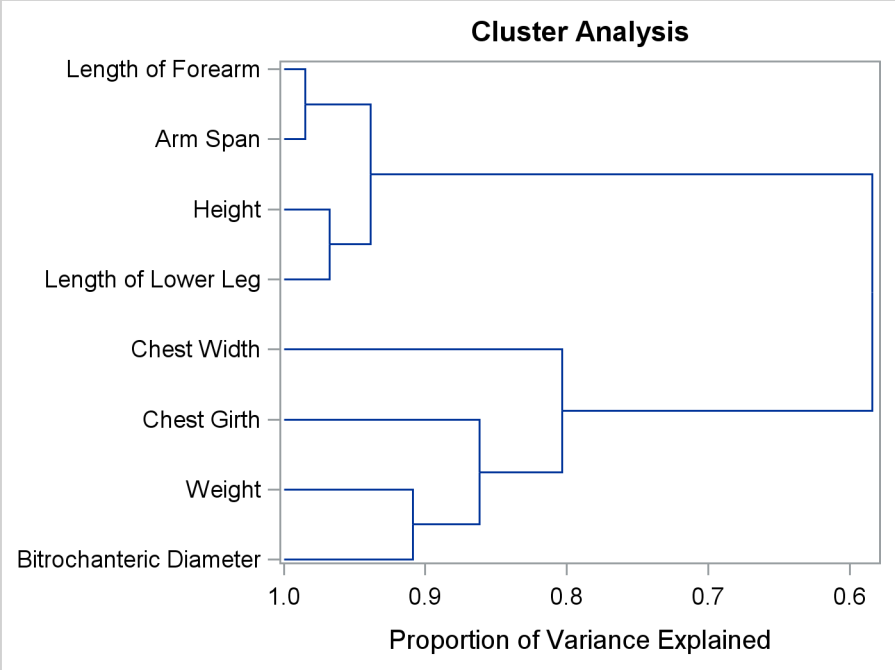
Observations	10000	Proportion	1
Variables	8	Maxeigen	0

Clustering algorithm converged.

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. 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. The tree diagram in Output 104.1.4 displays the cluster hierarchy.

Output 104.1.4 Dendrogram



It appears 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.

Harman, H. H. (1976), *Modern Factor Analysis*, 3rd Edition, Chicago: University of Chicago Press.

Harris, C. W. and Kaiser, H. F. (1964), "Oblique Factor Analytic Solutions by Orthogonal Transformation," *Psychometrika*, 32, 363–379.

Subject Index

Syntax Index