Chapter 16
The ACECLUS Procedure

Chapter Table of Contents

Chapter 16
The ACECLUS Procedure

Overview

The ACECLUS (Approximate Covariance Estimation for CLUStering) procedure obtains approximate estimates of the pooled within-cluster covariance matrix when the clusters are assumed to be multivariate normal with equal covariance matrices. Neither cluster membership nor the number of clusters need be known. PROC ACECLUS is useful for preprocessing data to be subsequently clustered by the CLUSTER or the FASTCLUS procedure.

Many clustering methods perform well with spherical clusters but poorly with elongated elliptical clusters (Everitt 1980, 77–97). If the elliptical clusters have roughly the same orientation and eccentricity, you can apply a linear transformation to the data to yield a spherical within-cluster covariance matrix, that is, a covariance matrix proportional to the identity. Equivalently, the distance between observations can be measured in the metric of the inverse of the pooled within-cluster covariance matrix. The remedy is difficult to apply, however, because you need to know what the clusters are in order to compute the sample within-cluster covariance matrix. One approach is to estimate iteratively both cluster membership and within-cluster covariance (Wolfe 1970; Hartigan 1975). Another approach is provided by Art, Gnanadesikan, and Kettenring (1982). They have devised an ingenious method for estimating the within-cluster covariance matrix without knowledge of the clusters. The method can be applied before any of the usual clustering techniques, including hierarchical clustering methods.

First, Art, Gnanadesikan, and Kettenring (1982) obtain a decomposition of the total-sample sum-of-squares-and-cross-products (SSCP) matrix into within-cluster and between-cluster SSCP matrices computed from pairwise differences between observations, rather than differences between observations and means. Then, they show how the within-cluster SSCP matrix based on pairwise differences can be approximated without knowing the number or the membership of the clusters. The approximate within-cluster SSCP matrix can be used to compute distances for cluster analysis, or it can be used in a canonical analysis similar to canonical discriminant analysis. For more information, see Chapter 21, “The CANDISC Procedure.”

Art, Gnanadesikan, and Kettenring demonstrate by Monte Carlo calculations that their method can produce better clusters than the Euclidean metric even when the approximation to the within-cluster SSCP matrix is poor or the within-cluster covariances are moderately heterogeneous. The algorithm used by the ACECLUS procedure differs slightly from the algorithm used by Art, Gnanadesikan, and Kettenring. In the following sections, the PROC ACECLUS algorithm is described first; then, differences between PROC ACECLUS and the method used by Art, Gnanadesikan, and Kettenring are summarized.

Background

It is well known from the literature on nonparametric statistics that variances and, hence, covariances can be computed from pairwise differences instead of deviations from means. (For example, Puri and Sen (1971, pp. 51–52) show that the variance is a U statistic of degree 2.) Let $\mathbf{X} = (x_{ij})$ be the data matrix with n observations (rows) and v variables (columns), and let $\bar{x}_j$ be the mean of the jth variable. The sample covariance matrix $\mathbf{S} = (s_{jk})$ is usually defined as

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

The matrix $\mathbf{S}$ can also be computed as

$$s_{jk} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{h=1}^{i-1} (x_{ij} - x_{hj})(x_{ik} - x_{hk})$$

Let $\mathbf{W} = (w_{jk})$ be the pooled within-cluster covariance matrix, q be the number of clusters, n_c be the number of observations in the cth cluster, and

$$d''_{ic} = \begin{cases} 1 & \text{if observation } i \text{ is in cluster } c \\ 0 & \text{otherwise} \end{cases}$$

The matrix $\mathbf{W}$ is normally defined as

$$w_{jk} = \frac{1}{n-q} \sum_{c=1}^{q} \sum_{i=1}^{n} d''_{ic} (x_{ij} - \bar{x}_{cj})(x_{ik} - \bar{x}_{ck})$$

where $\bar{x}_{cj}$ is the mean of the jth variable in cluster c. Let

$$d'_{ih} = \begin{cases} \frac{1}{n_c} & \text{if observations } i \text{ and } h \text{ are in cluster } c \\ 0 & \text{otherwise} \end{cases}$$

The matrix $\mathbf{W}$ can also be computed as

$$w_{jk} = \frac{1}{n-q} \sum_{i=2}^{n} \sum_{h=1}^{i-1} d'_{ih} (x_{ij} - x_{hj})(x_{ik} - x_{hk})$$

If the clusters are not known, d'_{ih} cannot be determined. However, an approximation to $\mathbf{W}$ can be obtained by using instead

$$d'_{ih} = \begin{cases} 1 & \text{if } \sum_{j=1}^{v} \sum_{k=1}^{v} m_{jk} (x_{ij} - x_{hj})(x_{ik} - x_{hk}) \leq u^2 \\ 0 & \text{otherwise} \end{cases}$$

where u is an appropriately chosen value and $\mathbf{M} = (m_{jk})$ is an appropriate metric. Let $\mathbf{A} = (a_{jk})$ be defined as

$$a_{jk} = \frac{\sum_{i=2}^{n} \sum_{h=1}^{i-1} d_{ih}(x_{ij} - x_{hj})(x_{ik} - x_{hk})}{2 \sum_{i=2}^{n} \sum_{h=1}^{i-1} d_{ih}}$$

If all of the following conditions hold, $\mathbf{A}$ equals $\mathbf{W}$:

- all within-cluster distances in the metric $\mathbf{M}$ are less than or equal to u
- all between-cluster distances in the metric $\mathbf{M}$ are greater than u
- all clusters have the same number of members n_c

If the clusters are of unequal size, $\mathbf{A}$ gives more weight to large clusters than $\mathbf{W}$ does, but this discrepancy should be of little importance if the population within-cluster covariance matrices are equal. There may be large differences between $\mathbf{A}$ and $\mathbf{W}$ if the cutoff u does not discriminate between pairs in the same cluster and pairs in different clusters. Lack of discrimination may occur for one of the following reasons:

- The clusters are not well separated.
- The metric $\mathbf{M}$ or the cutoff u is not chosen appropriately.

In the former case, little can be done to remedy the problem. The remaining question concerns how to choose $\mathbf{M}$ and u. Consider $\mathbf{M}$ first. The best choice for $\mathbf{M}$ is $\mathbf{W}^{-1}$, but $\mathbf{W}$ is not known. The solution is to use an iterative algorithm:

1. Obtain an initial estimate of $\mathbf{A}$, such as the identity or the total-sample covariance matrix. (See the INITIAL= option in the PROC ACECLUS statement for more information.)
2. Let $\mathbf{M}$ equal $\mathbf{A}^{-1}$.
3. Recompute $\mathbf{A}$ using the preceding formula.
4. Repeat steps 2 and 3 until the estimate stabilizes.

Convergence is assessed by comparing values of $\mathbf{A}$ on successive iterations. Let $\mathbf{A}_i$ be the value of $\mathbf{A}$ on the ith iteration and $\mathbf{A}_0$ be the initial estimate of $\mathbf{A}$. Let $\mathbf{Z}$ be a user-specified $v \times v$ matrix. (See the METRIC= option in the PROC ACECLUS statement for more information.) The convergence measure is

$$e_i = \frac{1}{v} \parallel \mathbf{Z}'(\mathbf{A}_i - \mathbf{A}_{i-1})\mathbf{Z} \parallel$$

where $\parallel \cdots \parallel$ indicates the Euclidean norm, that is, the square root of the sum of the squares of the elements of the matrix. In PROC ACECLUS, $\mathbf{Z}$ can be the identity

or an inverse factor of $\mathbf{S}$ or diag($\mathbf{S}$). Iteration stops when e_i falls below a user-specified value. (See the CONVERGE= option or the MAXITER= option in the PROC ACECLUS statement for more information.)

The remaining question of how to choose u has no simple answer. In practice, you must try several different values. PROC ACECLUS provides four different ways of specifying u:

- You can specify a constant value for u. This method is useful if the initial estimate of $\mathbf{A}$ is quite good. (See the ABSOLUTE option and the THRESHOLD= option in the PROC ACECLUS statement for more information.)
- You can specify a threshold value $t > 0$ that is multiplied by the root mean square distance between observations in the current metric on each iteration to give u. Thus, the value of u changes from iteration to iteration. This method is appropriate if the initial estimate of $\mathbf{A}$ is poor. (See the THRESHOLD= option in the PROC ACECLUS statement for more information)
- You can specify a value p, $0 < p < 1$, to be transformed into a distance u such that approximately a proportion p of the pairwise Mahalanobis distances between observations in a random sample from a multivariate normal distribution will be less than u in repeated sampling. The transformation can be computed only if the number of observations exceeds the number of variables, preferably by at least 10 percent. This method also requires a good initial estimate of $\mathbf{A}$. (See the PROPORTION= option and the ABSOLUTE option in the PROC ACECLUS statement for more information.)
- You can specify a value p, $0 < p < 1$, to be transformed into a value t that is then multiplied by $1/\sqrt{2v}$ times the root mean square distance between observations in the current metric on each iteration to yield u. The value of u changes from iteration to iteration. This method can be used with a poor initial estimate of $\mathbf{A}$. (See the PROPORTION= option in the PROC ACECLUS statement for more information.)

In most cases, the analysis should begin with the last method using values of p between 0.5 and 0.01 and using the full covariance matrix as the initial estimate of $\mathbf{A}$.

Proportions p are transformed to distances t using the formula

$$t^2 = 2v \left\{ \left[F^{-1}_{v,n-v}(p) \right]^{\frac{n-v}{n-1}} \right\}$$

where $F^{-1}_{v,n-v}$ is the quantile (inverse cumulative distribution) function of an F random variable with v and $n-v$ degrees of freedom. The squared Mahalanobis distance between a single pair of observations sampled from a multivariate normal distribution is distributed as $2v$ times an F random variable with v and $n-v$ degrees of freedom. The distances between two pairs of observations are correlated if the pairs have an observation in common. The quantile function is raised to the power given in the preceding formula to compensate approximately for the correlations among distances between pairs of observations that share a member. Monte Carlo studies indicate that

the approximation is acceptable if the number of observations exceeds the number of variables by at least 10 percent.

If $\mathbf{A}$ becomes singular, step 2 in the iterative algorithm cannot be performed because $\mathbf{A}$ cannot be inverted. In this case, let $\mathbf{Z}$ be the matrix as defined in discussing the convergence measure, and let $\mathbf{Z}'\mathbf{A}\mathbf{Z} = \mathbf{R}'\mathbf{\Lambda}\mathbf{R}$ where $\mathbf{R}'\mathbf{R} = \mathbf{R}\mathbf{R}' = \mathbf{I}$ and $\mathbf{\Lambda} = (\lambda_{jk})$ is diagonal. Let $\mathbf{\Lambda}^* = (\lambda^*_{jk})$ be a diagonal matrix where $\lambda^*_{jj} = \max(\lambda_{jj}, g \text{ trace}(\mathbf{\Lambda}))$, and $0 < g < 1$ is a user-specified singularity criterion (see the SINGULAR= option in the PROC ACECLUS statement for more information). Then $\mathbf{M}$ is computed as $\mathbf{Z}\mathbf{R}'(\mathbf{\Lambda}^*)^{-1}\mathbf{R}\mathbf{Z}'$.

The ACECLUS procedure differs from the method used by Art, Gnanadesikan, and Kettenring (1982) in several respects.

- The Art, Gnanadesikan, and Kettenring method uses the identity matrix as the initial estimate, whereas the ACECLUS procedure enables you to specify any symmetric matrix as the initial estimate and defaults to the total-sample covariance matrix. The default initial estimate in PROC ACECLUS is chosen to yield invariance under nonsingular linear transformations of the data but may sometimes obscure clusters that become apparent if the identity matrix is used.
- The Art, Gnanadesikan, and Kettenring method carries out all computations with SSCP matrices, whereas the ACECLUS procedure uses estimated covariance matrices because covariances are easier to interpret than crossproducts.
- The Art, Gnanadesikan, and Kettenring method uses the m pairs with the smallest distances to form the new estimate at each iteration, where m is specified by the user, whereas the ACECLUS procedure uses all pairs closer than a given cutoff value. Kettenring (1984) says that the m-closest-pairs method seems to give the user more direct control. PROC ACECLUS uses a distance cutoff because it yields a slight decrease in computer time and because in some cases, such as widely separated spherical clusters, the results are less sensitive to the choice of distance cutoff than to the choice of m. Much research remains to be done on this issue.
- The Art, Gnanadesikan, and Kettenring method uses a different convergence measure. Let $\mathbf{A}_i$ be computed on each iteration using the m-closest-pairs method, and let $\mathbf{B}_i = \mathbf{A}_{i-1}^{-1}\mathbf{A}_i - \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The convergence measure is equivalent to trace$(\mathbf{B}_i^2)$.

Analyses of Fisher's (1936) iris data, consisting of measurements of petal and sepal length and width for fifty specimens from each of three iris species, are summarized in Table 16.1. The number of misclassified observations out of 150 is given for four clustering methods:

- k-means as implemented in PROC FASTCLUS with MAXC=3, MAXITER=99, and CONV=0
- Ward's minimum variance method as implemented in PROC CLUSTER

- average linkage on Euclidean distances as implemented in PROC CLUSTER
- the centroid method as implemented in PROC CLUSTER

Each hierarchical analysis is followed by the TREE procedure with NCL=3 to determine cluster assignments at the three-cluster level. Clusters with twenty or fewer observations are discarded by using the DOCK=20 option. The observations in a discarded cluster are considered unclassified.

Each method is applied to

- the raw data
- the data standardized to unit variance by the STANDARD procedure
- two standardized principal components accounting for 95 percent of the standardized variance and having an identity total-sample covariance matrix, computed by the PRINCOMP procedure with the STD option
- four standardized principal components having an identity total-sample covariance matrix, computed by PROC PRINCOMP with the STD option
- the data transformed by PROC ACECLUS using seven different settings of the PROPORTION= (P=) option
- four canonical variables having an identity pooled within-species covariance matrix, computed using the CANDISC procedure

Theoretically, the best results should be obtained by using the canonical variables from PROC CANDISC. PROC ACECLUS yields results comparable to PROC CANDISC for values of the PROPORTION= option ranging from 0.005 to 0.02. At PROPORTION=0.04, average linkage and the centroid method show some deterioration, but k-means and Ward's method continue to produce excellent classifications. At larger values of the PROPORTION= option, all methods perform poorly, although no worse than with four standardized principal components.

Table 16.1. Number of Misclassified and Unclassified Observations Using Fisher's (1936) Iris Data

Data	Clustering Method			
	k-means	Ward's	Average Linkage	Centroid
raw data	16*	16*	25 + 12**	14*
standardized data	25	26	33+4	33+4
two standardized principal components	29	31	30+9	27+32
four standardized principal components	39	27	32+7	45+11
transformed by ACECLUS P=0.32	39	10+9	7+25	
transformed by ACECLUS P=0.16	39	18+9	7+19	7+26
transformed by ACECLUS P=0.08	19	9	3+13	5+16
transformed by ACECLUS P=0.04	4	5	1+19	3+12
transformed by ACECLUS P=0.02	4	3	3	3
transformed by ACECLUS P=0.01	4	4	3	4
transformed by ACECLUS P=0.005	4	4	4	4
canonical variables	3	5	4	4+1

* A single number represents misclassified observations with no unclassified observations.

** Where two numbers are separated by a plus sign, the first is the number of misclassified observations; the second is the number of unclassified observations.

This example demonstrates the following:

- PROC ACECLUS can produce results as good as those from the optimal transformation.
- PROC ACECLUS can be useful even when the within-cluster covariance matrices are moderately heterogeneous.
- The choice of the distance cutoff as specified by the PROPORTION= or the THRESHOLD= option is important, and several values should be tried.
- Commonly used transformations such as standardization and principal components can produce poor classifications.

Although experience with the Art, Gnanadesikan, and Kettenring and PROC ACECLUS methods is limited, the results so far suggest that these methods help considerably more often than they hinder the subsequent cluster analysis, especially with normal-mixture techniques such as k-means and Ward's minimum variance method.

Getting Started

The following example demonstrates how you can use the ACECLUS procedure to obtain approximate estimates of the pooled within-cluster covariance matrix and to compute canonical variables for subsequent analysis. You use PROC ACECLUS to preprocess data before you cluster it using the FASTCLUS or CLUSTER procedure.

Suppose you want to determine whether national figures for birth rates, death rates, and infant death rates can be used to determine certain types or categories of countries. You want to perform a cluster analysis to determine whether the observations can be formed into groups suggested by the data. Previous studies indicate that the clusters computed from this type of data can be elongated and elliptical. Thus, you need to perform a linear transformation on the raw data before the cluster analysis.

The following data* from Rouncefield (1995) are the birth rates, death rates, and infant death rates for 97 countries. The following statements create the SAS data set Poverty:

```
data poverty;
   input Birth Death InfantDeath Country $15. @@;
   datalines;
24.7  5.7  30.8 Albania         12.5 11.9  14.4 Bulgaria
13.4 11.7  11.3 Czechoslovakia  12   12.4   7.6 Former_E._Germa
11.6 13.4  14.8 Hungary         14.3 10.2    16 Poland
13.6 10.7  26.9 Romania           14    9  20.2 Yugoslavia
17.7   10    23 USSR            15.2  9.5  13.1 Byelorussia
13.4 11.6    13 Ukrainian_SSR   20.7  8.4  25.7 Argentina
46.6   18   111 Bolivia         28.6  7.9    63 Brazil
23.4  5.8  17.1 Chile           27.4  6.1    40 Columbia
```

*These data have been compiled from the United Nations Demographic Yearbook 1990 (United Nations publications, Sales No. E/F.91.XII.1, copyright 1991, United Nations, New York) and are reproduced with the permission of the United Nations.

```
32.9  7.4    63 Ecuador          28.3  7.3    56 Guyana
34.8  6.6    42 Paraguay         32.9  8.3 109.9 Peru
  18  9.6  21.9 Uruguay          27.5  4.4  23.3 Venezuela
  29 23.2    43 Mexico             12 10.6   7.9 Belgium
13.2 10.1   5.8 Finland          12.4 11.9   7.5 Denmark
13.6  9.4   7.4 France           11.4 11.2   7.4 Germany
10.1  9.2    11 Greece           15.1  9.1   7.5 Ireland
 9.7  9.1   8.8 Italy            13.2  8.6   7.1 Netherlands
14.3 10.7   7.8 Norway           11.9  9.5  13.1 Portugal
10.7  8.2   8.1 Spain            14.5 11.1   5.6 Sweden
12.5  9.5   7.1 Switzerland      13.6 11.5   8.4 U.K.
14.9  7.4     8 Austria           9.9  6.7   4.5 Japan
14.5  7.3   7.2 Canada           16.7  8.1   9.1 U.S.A.
40.4 18.7 181.6 Afghanistan      28.4  3.8    16 Bahrain
42.5 11.5 108.1 Iran             42.6  7.8    69 Iraq
22.3  6.3   9.7 Israel           38.9  6.4    44 Jordan
26.8  2.2  15.6 Kuwait           31.7  8.7    48 Lebanon
45.6  7.8    40 Oman             42.1  7.6    71 Saudi_Arabia
29.2  8.4    76 Turkey           22.8  3.8    26 United_Arab_Emr
42.2 15.5   119 Bangladesh       41.4 16.6   130 Cambodia
21.2  6.7    32 China            11.7  4.9   6.1 Hong_Kong
30.5 10.2    91 India            28.6  9.4    75 Indonesia
23.5 18.1    25 Korea            31.6  5.6    24 Malaysia
36.1  8.8    68 Mongolia         39.6 14.8   128 Nepal
30.3  8.1 107.7 Pakistan        33.2  7.7    45 Philippines
17.8  5.2   7.5 Singapore        21.3  6.2  19.4 Sri_Lanka
22.3  7.7    28 Thailand         31.8  9.5    64 Vietnam
35.5  8.3    74 Algeria          47.2 20.2   137 Angola
48.5 11.6    67 Botswana         46.1 14.6    73 Congo
38.8  9.5  49.4 Egypt            48.6 20.7   137 Ethiopia
39.4 16.8   103 Gabon            47.4 21.4   143 Gambia
44.4 13.1    90 Ghana              47 11.3    72 Kenya
 44  9.4    82 Libya           48.3   25   130 Malawi
35.5  9.8    82 Morocco            45 18.5   141 Mozambique
  44 12.1   135 Namibia          48.5 15.6   105 Nigeria
48.2 23.4   154 Sierra_Leone     50.1 20.2   132 Somalia
32.1  9.9    72 South_Africa     44.6 15.8   108 Sudan
46.8 12.5   118 Swaziland        31.1  7.3    52 Tunisia
52.2 15.6   103 Uganda           50.5   14   106 Tanzania
45.6 14.2    83 Zaire            51.1 13.7    80 Zambia
41.7 10.3    66 Zimbabwe
;
```

The data set Poverty contains the character variable Country and the numeric variables Birth, Death, and InfantDeath, which represent the birth rate per thousand, death rate per thousand, and infant death rate per thousand. The $15. in the INPUT statement specifies that the variable Country is a character variable with a length of 15. The double trailing at sign (@@) in the INPUT statement specifies that observations are input from each line until all values have been read.

It is often useful when beginning a cluster analysis to look at the data graphically. The following statements use the GPLOT procedure to make a scatter plot of the variables Birth and Death.

```
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
proc gplot data=poverty;
   plot Birth*Death/
   frame cframe=ligr vaxis=axis1 haxis=axis2;
run;
```

The plot, displayed in Figure 16.1, indicates the difficulty of dividing the points into clusters. Plots of the other variable pairs (not shown) display similar characteristics. The clusters that comprise these data may be poorly separated and elongated. Data with poorly separated or elongated clusters must be transformed.

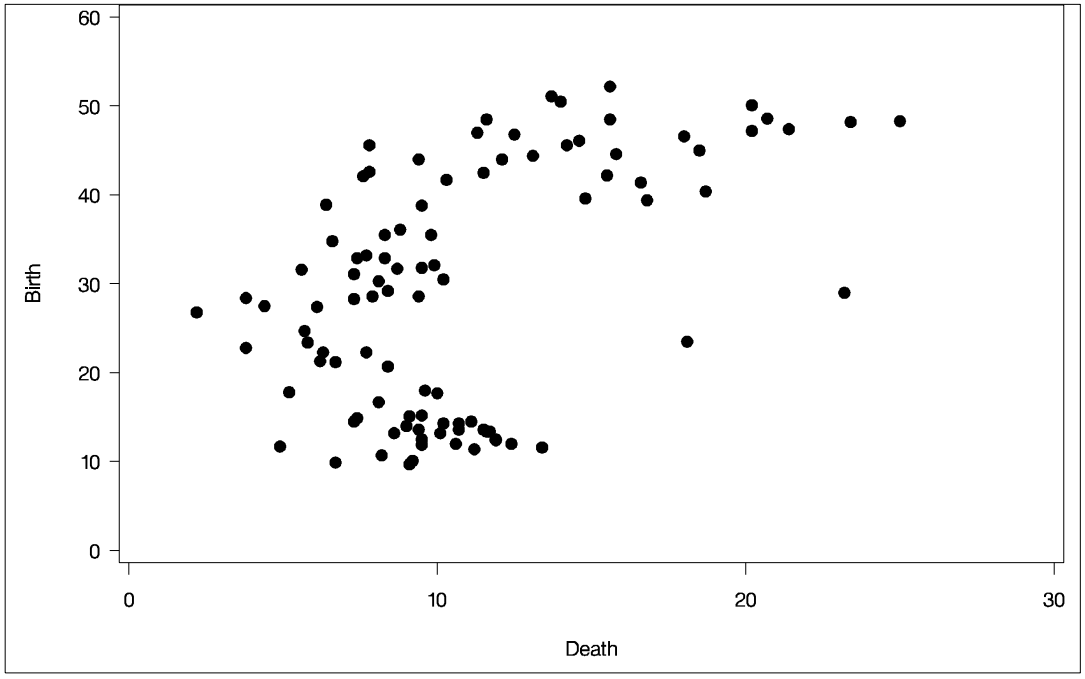


Figure 16.1. Scatter Plot of Original Poverty Data: Birth Rate versus Death Rate

If you know the within-cluster covariances, you can transform the data to make the clusters spherical. However, since you do not know what the clusters are, you cannot calculate exactly the within-cluster covariance matrix. The ACECLUS procedure estimates the within-cluster covariance matrix to transform the data, even when you have no knowledge of cluster membership or the number of clusters.

The following statements perform the ACECLUS procedure transformation using the SAS data set Poverty.

```
proc aceclus data=poverty out=ace proportion=.03;
   var Birth Death InfantDeath;
run;
```

The OUT= option creates an output data set called Ace to contain the canonical variable scores. The PROPORTION= option specifies that approximately three percent

of the pairs are included in the estimation of the within-cluster covariance matrix. The VAR statement specifies that the variables Birth, Death, and InfantDeath are used in computing the canonical variables.

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

Figure 16.2 displays the number of observations, the number of variables, and the settings for the PROPORTION and CONVERGE options. The PROPORTION option is set at 0.03, as specified in the previous statements. The CONVERGE parameter is set at its default value of 0.001.

```
                    The ACECLUS Procedure

       Approximate Covariance Estimation for Cluster Analysis


         Observations          97    Proportion      0.0300
         Variables              3    Converge       0.00100


                Means and Standard Deviations
                                          Standard
             Variable           Mean     Deviation

             Birth           29.2299       13.5467
             Death           10.8361        4.6475
             InfantDeath     54.9010       45.9926


                COV: Total Sample Covariances

                      Birth             Death        InfantDeath

Birth              183.512951         30.610056         534.794969
Death               30.610056         21.599205         139.925900
InfantDeath        534.794969        139.925900        2115.317811

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix


                   Threshold =    0.292815
```

Figure 16.2. Means, Standard Deviations, and Covariance Matrix from the ACECLUS Procedure

Figure 16.2 next displays the means, standard deviations, and sample covariance matrix of the analytical variables.

The type of matrix used for the initial within-cluster covariance estimate is displayed in Figure 16.3. In this example, that initial estimate is the full covariance matrix. The threshold value that corresponds to the PROPORTION=0.03 setting is given as 0.292815.

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Iteration History

Iteration	RMS Distance	Distance Cutoff	Pairs Within Cutoff	Convergence Measure
1	2.449	0.717	385.0	0.552025
2	12.534	3.670	446.0	0.008406
3	12.851	3.763	521.0	0.009655
4	12.882	3.772	591.0	0.011193
5	12.716	3.723	628.0	0.008784
6	12.821	3.754	658.0	0.005553
7	12.774	3.740	680.0	0.003010
8	12.631	3.699	683.0	0.000676

Algorithm converged.

Figure 16.3. Table of Iteration History from the ACECLUS Procedure

Figure 16.3 displays the iteration history. For each iteration, PROC ACECLUS displays the following measures:

- root mean square distance between all pairs of observations
- distance cutoff for including pairs of observations in the estimate of within-cluster covariances (equal to RMS*Threshold)
- number of pairs within the cutoff
- convergence measure

Figure 16.4 displays the approximate within-cluster covariance matrix and the table of eigenvalues from the canonical analysis. The first column of the eigenvalues table contains numbers for the eigenvectors. The next column of the table lists the eigenvalues of Inv(ACE)*(COV-ACE).

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

ACE: Approximate Covariance Estimate Within Clusters

	Birth	Death	InfantDeath
Birth	5.94644949	-0.63235725	6.28151537
Death	-0.63235725	2.33464129	1.59005857
InfantDeath	6.28151537	1.59005857	35.10327233

Eigenvalues of Inv(ACE)*(COV-ACE)

	Eigenvalue	Difference	Proportion	Cumulative
1	63.5500	54.7313	0.8277	0.8277
2	8.8187	4.4038	0.1149	0.9425
3	4.4149		0.0575	1.0000

Figure 16.4. Approximate Within–Cluster Covariance Estimates

The next three columns of the eigenvalue table (Figure 16.4) display measures of the relative size and importance of the eigenvalues. The first column lists the difference between each eigenvalue and its successor. The last two columns display the individual and cumulative proportions that each eigenvalue contributes to the total sum of eigenvalues.

The raw and standardized canonical coefficients are displayed in Figure 16.5. The coefficients are standardized by multiplying the raw coefficients with the standard deviation of the associated variable. The ACECLUS procedure uses these standardized canonical coefficients to create the transformed canonical variables, which are the linear transformations of the original input variables, Birth, Death, and InfantDeath.

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Eigenvectors (Raw Canonical Coefficients)

	Can1	Can2	Can3
Birth	0.125610	0.457037	0.003875
Death	0.108402	0.163792	0.663538
InfantDeath	0.134704	-.133620	-.046266

Standardized Canonical Coefficients

	Can1	Can2	Can3
Birth	1.70160	6.19134	0.05249
Death	0.50380	0.76122	3.08379
InfantDeath	6.19540	-6.14553	-2.12790

Figure 16.5. Raw and Standardized Canonical Coefficients from the ACECLUS Procedure

The following statements invoke the CLUSTER procedure, using the SAS data set Ace created in the previous ACECLUS procedure.

```
proc cluster data=ace outtree=tree noprint method=ward;
   var can1 can2 can3 ;
   copy Birth--Country;
run;
```

The OUTTREE= option creates the output SAS data set Tree that is used in subsequent statements to draw a tree diagram. The NOPRINT option suppresses the display of the output. The METHOD= option specifies Ward's minimum-variance clustering method.

The VAR statement specifies that the canonical variables computed in the ACECLUS procedure are used in the cluster analysis. The COPY statement specifies that all the variables from the SAS data set Poverty (Birth—Country) are added to the output data set Tree.

The following statements use the TREE procedure to create an output SAS data set called New. The NCLUSTERS= option specifies the number of clusters desired in the SAS data set New. The NOPRINT option suppresses the display of the output.

```
proc tree data=tree out=new nclusters=3 noprint;
   copy Birth Death InfantDeath can1 can2 ;
   id Country;
run;
```

The COPY statement copies the canonical variables CAN1 and CAN2 (computed in the preceding ACECLUS procedure) and the original analytical variables Birth, Death, and InfantDeath into the output SAS data set New.

The following statements invoke the GPLOT procedure, using the SAS data set created by PROC TREE:

```
legend1 frame cframe=ligr cborder=black
        position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
proc gplot data=new;
   plot Birth*Death=cluster/
        frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
   plot can2*can1=cluster/
        frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;
```

The first plot statement requests a scatter plot of the two variables Birth and Death, using the variable CLUSTER as the identification variable.

The second PLOT statement requests a plot of the two canonical variables, using the value of the variable CLUSTER as the identification variable.

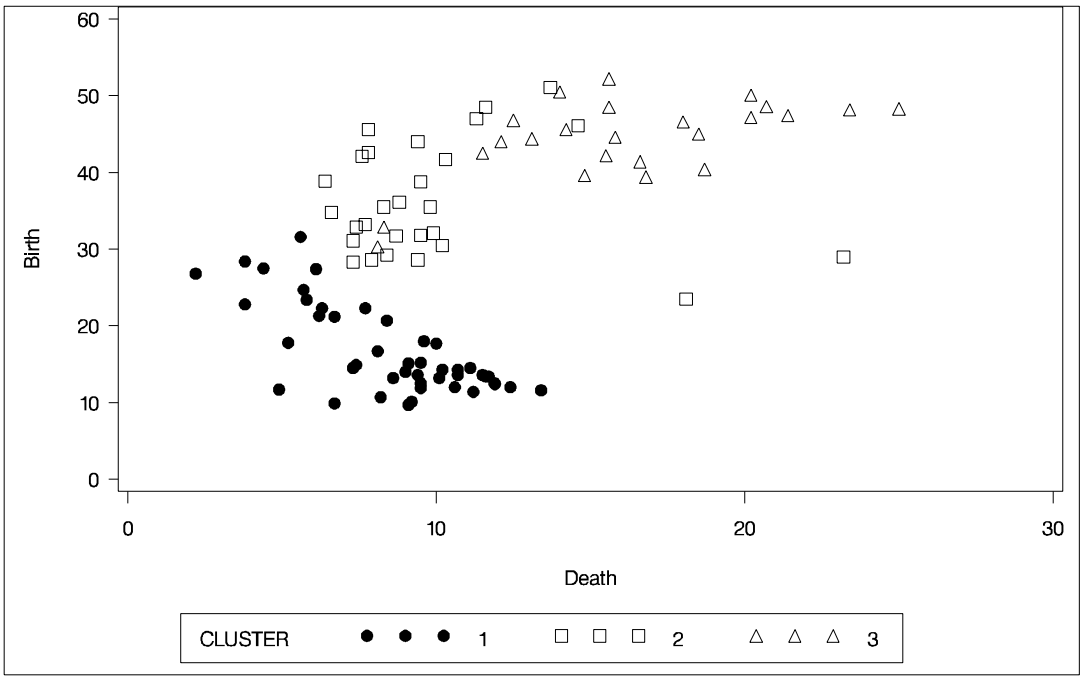


Figure 16.6. Scatter Plot of Poverty Data, Identified by Cluster

Figure 16.6 and Figure 16.7 display the separation of the clusters when three clusters are calculated.

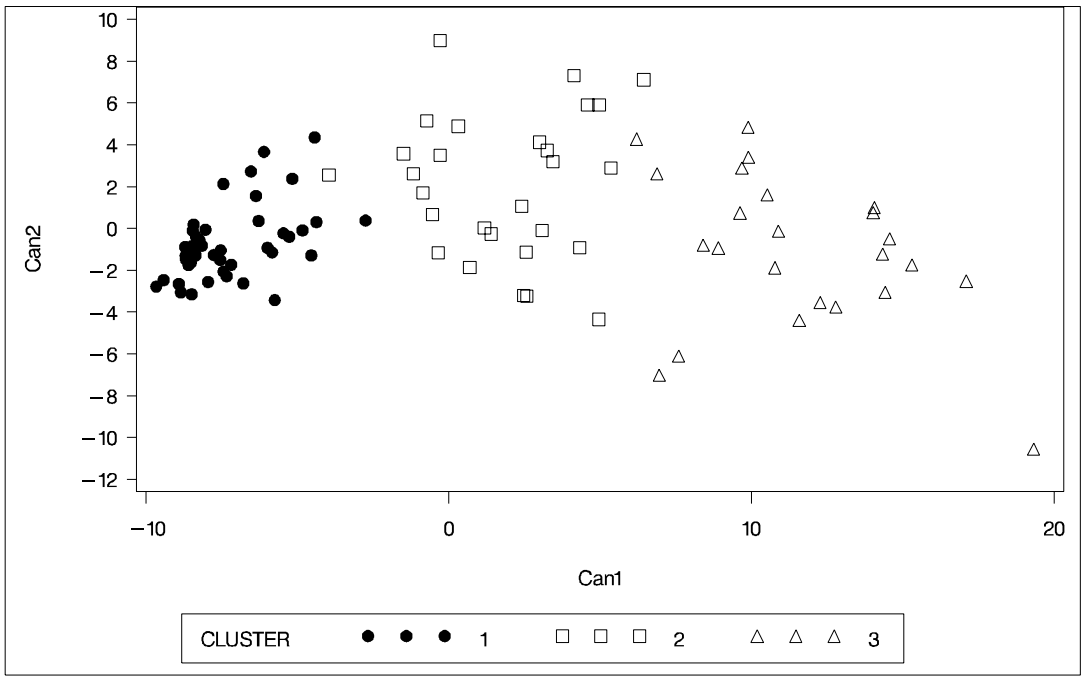


Figure 16.7. Scatter Plot of Canonical Variables

Syntax

The following statements are available in the ACECLUS procedure.

PROC ACECLUS *PROPORTION=p* **|** *THRESHOLD=t* < *options* > **;**
BY *variables* **;**
FREQ *variable* **;**
VAR *variables* **;**
WEIGHT *variable* **;**

Usually you need only the VAR statement in addition to the required PROC ACECLUS statement. The optional BY, FREQ, VAR, and WEIGHT statements are described in alphabetical order after the PROC ACECLUS statement.

PROC ACECLUS Statement

PROC ACECLUS *PROPORTION=p | THRESHOLD=t* < *options* > **;**

The PROC ACECLUS statement starts the ACECLUS procedure. The options available with the PROC ACECLUS statement are summarized in Table 16.2 and discussed in the following sections. Note that, if you specify the METHOD=COUNT option, you must specify either the PROPORTION= or the MPAIRS= option. Otherwise, you must specify either the PROPORTION= or THRESHOLD= option.

Table 16.2. Summary of PROC ACECLUS Statement Options

Task	Options	Description
Specify clustering options		
	METHOD=	specify the clustering method
	MPAIRS=	specify number of pairs for estimating within-cluster covariance (when you specify the option METHOD=COUNT)
	PROPORTION=	specify proportion of pairs for estimating within-cluster covariance
	THRESHOLD=	specify the threshold for including pairs in the estimation of the within-cluster covariance
Specify input and output data sets		
	DATA=	specify input data set name
	OUT=	specify output data set name
	OUTSTAT=	specify output data set name containing various statistics
Specify iteration options		
	ABSOLUTE	use absolute instead of relative threshold
	CONVERGE=	specify convergence criterion
	INITIAL=	specify initial estimate of within-cluster covariance matrix
	MAXITER=	specify maximum number of iterations
	METRIC=	specify metric in which computations are performed
	SINGULAR=	specify singularity criterion
Specify canonical analysis options		
	N=	specify number of canonical variables
	PREFIX=	specify prefix for naming canonical variables
Control displayed output		
	NOPRINT	suppress the display of the output
	PP	produce PP-plot of distances between pairs from last iteration
	QQ	produce QQ-plot of power transformation of distances between pairs from last iteration
	SHORT	omit all output except for iteration history and eigenvalue table

The following list provides details on the options. The list is in alphabetical order.

ABSOLUTE
causes the THRESHOLD= value or the threshold computed from the PROPORTION= option to be treated absolutely rather than relative to the root mean square distance between observations. Use the ABSOLUTE option only when you are confident that the initial estimate of the within-cluster covariance matrix is close to the final estimate, such as when the INITIAL= option specifies a data set created by a previous execution of PROC ACECLUS using the OUTSTAT= option.

CONVERGE=*c*
specifies the convergence criterion. By default, CONVERGE= 0.001. Iteration stops when the convergence measure falls below the value specified by the CONVERGE= option or when the iteration limit as specified by the MAXITER= option is exceeded, whichever happens first.

DATA=*SAS-data-set*
specifies the SAS data set to be analyzed. By default, PROC ACECLUS uses the most recently created SAS data set.

INITIAL=*name*
specifies the matrix for the initial estimate of the within-cluster covariance matrix. Valid values for *name* are as follows:

DIAGONAL | D
: uses the diagonal matrix of sample variances as the initial estimate of the within-cluster covariance matrix.

FULL | F
: uses the total-sample covariance matrix as the initial estimate of the within-cluster covariance matrix.

IDENTITY | I
: uses the identity matrix as the initial estimate of the within-cluster covariance matrix.

INPUT=*SAS-data-set*
: specifies a SAS data set from which to obtain the initial estimate of the within-cluster covariance matrix. The data set can be TYPE=CORR, COV, UCORR, UCOV, SSCP, or ACE, or it can be an ordinary SAS data set. (See Appendix 1, “Special SAS Data Sets,” for descriptions of CORR, COV, UCORR, UCOV, and SSCP data sets. See the section “Output Data Sets” on page 325 for a description of ACE data sets.)

  If you do not specify the INITIAL= option, the default is the matrix specified by the METRIC= option. If neither the INITIAL= nor the METRIC= option is specified, INITIAL=FULL is used if there are enough observations to obtain a nonsingular total-sample covariance matrix; otherwise, INITIAL=DIAGONAL is used.

MAXITER=*n*
specifies the maximum number of iterations. By default, MAXITER=10.

METHOD= COUNT | C
METHOD= THRESHOLD | T

specifies the clustering method. The METHOD=THRESHOLD option requests a method (also the default) that uses all pairs closer than a given cutoff value to form the estimate at each iteration. The METHOD=COUNT option requests a method that uses a number of pairs, m, with the smallest distances to form the estimate at each iteration.

METRIC=*name*

specifies the metric in which the computations are performed, implies the default value for the INITIAL= option, and specifies the matrix $\mathbf{Z}$ used in the formula for the convergence measure e_i and for checking singularity of the $\mathbf{A}$ matrix. Valid values for *name* are as follows:

DIAGONAL | D	uses the diagonal matrix of sample variances $\mathrm{diag}(\mathbf{S})$ and sets $\mathbf{Z} = \mathrm{diag}(\mathbf{S})^{-\frac{1}{2}}$, where the superscript $-\frac{1}{2}$ indicates an inverse factor.
FULL | F	uses the total-sample covariance matrix $\mathbf{S}$ and sets $\mathbf{Z} = \mathbf{S}^{-\frac{1}{2}}$.
IDENTITY | I	uses the identity matrix $\mathbf{I}$ and sets $\mathbf{Z} = \mathbf{I}$.

If you do not specify the METRIC= option, METRIC=FULL is used if there are enough observations to obtain a nonsingular total-sample covariance matrix; otherwise, METRIC=DIAGONAL is used.

The option METRIC= is rather technical. It affects the computations in a variety of ways, but for well-conditioned data the effects are subtle. For most data sets, the METRIC= option is not needed.

MPAIRS=*m*

specifies the number of pairs to be included in the estimation of the within-cluster covariance matrix when METHOD=COUNT is requested. The values of m must be greater than 0 but less than or equal to $(totfq \times (totfq - 1))/2$, where *totfq* is the sum of nonmissing frequencies specified in the FREQ statement. If there is no FREQ statement, *totfq* equals the number of total nonmissing observations.

N=*n*

specifies the number of canonical variables to be computed. The default is the number of variables analyzed. N=0 suppresses the canonical analysis.

NOPRINT

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, "Using the Output Delivery System."

OUT=*SAS-data-set*

creates an output SAS data set that contains all the original data as well as the canonical variables having an estimated within-cluster covariance matrix equal to the identity matrix. If you want to create a permanent SAS data set, you must specify a

two-level name. See Chapter 16, "SAS Data Files" in *SAS Language Reference: Concepts* for information on permanent SAS data sets.

OUTSTAT=*SAS-data-set*
specifies a TYPE=ACE output SAS data set that contains means, standard deviations, number of observations, covariances, estimated within-cluster covariances, eigenvalues, and canonical coefficients. If you want to create a permanent SAS data set, you must specify a two-level name. See Chapter 16, "SAS Data Files" in *SAS Language Reference: Concepts* for information on permanent SAS data sets.

PROPORTION=*p*
PERCENT=*p*
P=*p*
specifies the percentage of pairs to be included in the estimation of the within-cluster covariance matrix. The value of p must be greater than 0. If p is greater than or equal to 1, it is interpreted as a percentage and divided by 100; PROPORTION=0.02 and PROPORTION=2 are equivalent. When you specify METHOD=THRESHOLD, a threshold value is computed from the PROPORTION= option under the assumption that the observations are sampled from a multivariate normal distribution.

When you specify METHOD=COUNT, the number of pairs, m, is computed from PROPORTION=p as

$$m = \text{floor}\left(\frac{p}{2} \times \textit{totfq} \times (\textit{totfq} - 1)\right)$$

where *totfq* is the number of total non-missing observations.

PP
produces a PP probability plot of distances between pairs of observations computed in the last iteration.

PREFIX=*name*
specifies a prefix for naming the canonical variables. By default the names are CAN1, CAN2, . . . , CANn. If you specify PREFIX=ABC, the variables are named ABC1, ABC2, ABC3, and so on. The number of characters in the prefix plus the number of digits required to designate the variables should not exceed the name length defined by the VALIDVARNAME= system option. For more information on the VALIDVARNAME= system option, refer to *SAS Language Reference: Dictionary*.

QQ
produces a QQ probability plot of a power transformation of the distances between pairs of observations computed in the last iteration. **Caution:** The QQ plot may require an enormous amount of computer time.

SHORT
omits all items from the standard output except for the iteration history and the eigenvalue table.

SINGULAR=*g*
SING=*g*

specifies a singularity criterion $0 < g < 1$ for the total-sample covariance matrix $\mathbf{S}$ and the approximate within-cluster covariance estimate $\mathbf{A}$. The default is SINGULAR=1E−4.

THRESHOLD=*t*
T=*t*

specifies the threshold for including pairs of observations in the estimation of the within-cluster covariance matrix. A pair of observations is included if the Euclidean distance between them is less than or equal to t times the root mean square distance computed over all pairs of observations.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC ACECLUS 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 ACECLUS 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.

If you specify the INITIAL=INPUT= option and the INITIAL=INPUT= data set does not contain any of the BY variables, the entire INITIAL=INPUT= data set provides the initial value for the matrix $\mathbf{A}$ for each BY group in the DATA= data set.

If the INITIAL=INPUT= data set contains some but not all of the BY variables, or if some BY variables do not have the same type or length in the INITIAL=INPUT= data set as in the DATA= data set, then PROC ACECLUS displays an error message and stops.

If all the BY variables appear in the INITIAL=INPUT= data set with the same type and length as in the DATA= data set, then each BY group in the INITIAL=INPUT= data set provides the initial value for $\mathbf{A}$ for the corresponding BY group in the DATA= data set. All BY groups in the DATA= data set must also appear in the INITIAL= INPUT= data set. The BY groups in the INITIAL=INPUT= data set must be in

the same order as in the DATA= data set. If you specify NOTSORTED in the BY statement, identical BY groups must occur in the same order in both data sets. If you do not specify NOTSORTED, some BY groups can appear in the INITIAL= INPUT= data set, but not in the DATA= data set; such BY groups are not used in the analysis.

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 observation, include the name of that variable in the 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 a value of the FREQ variable is not integral, it is truncated to the largest integer not exceeding the given value. Observations with FREQ values less than one are not included in the analysis. The total number of observations is considered equal to the sum of the FREQ variable.

VAR Statement

VAR *variables* **;**

The VAR statement specifies the numeric variables to be analyzed. If the VAR statement is omitted, all numeric variables not specified in other statements are analyzed.

WEIGHT Statement

WEIGHT *variable* **;**

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 that variable 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 values of the WEIGHT variable can be non-integral and are not truncated. An observation is used in the analysis only if the value of the WEIGHT variable is greater than zero.

The WEIGHT and FREQ statements have a similar effect, except in calculating the divisor of the $\mathbf{A}$ matrix.

Details

Missing Values

Observations with missing values are omitted from the analysis and are given missing values for canonical variable scores in the OUT= data set.

Output Data Sets

OUT= Data Set

The OUT= data set contains all the variables in the original data set plus new variables containing the canonical variable scores. The N= option determines the number of new variables. The OUT= data set is not created if N=0. The names of the new variables are formed by concatenating the value given by the PREFIX= option (or the prefix CAN if the PREFIX= option is not specified) and the numbers 1, 2, 3, and so on. The OUT= data set can be used as input to PROC CLUSTER or PROC FASTCLUS. The cluster analysis should be performed on the canonical variables, not on the original variables.

OUTSTAT= Data Set

The OUTSTAT= data set is a TYPE=ACE data set containing the following variables.

- the BY variables, if any
- the two new character variables, _TYPE_ and _NAME_
- the variables analyzed, that is, those in the VAR statement, or, if there is no VAR statement, all numeric variables not listed in any other statement

Each observation in the new data set contains some type of statistic as indicated by the _TYPE_ variable. The values of the _TYPE_ variable are as follows:

TYPE	**Contents**
MEAN	mean of each variable
STD	standard deviation of each variable
N	number of observations on which the analysis is based. This value is the same for each variable.
SUMWGT	sum of the weights if a WEIGHT statement is used. This value is the same for each variable.
COV	covariances between each variable and the variable named by the _NAME_ variable. The number of observations with _TYPE_=COV is equal to the number of variables being analyzed.
ACE	estimated within-cluster covariances between each variable and the variable named by the _NAME_ variable. The number of observations with _TYPE_=ACE is equal to the number of variables being analyzed.

EIGENVAL	eigenvalues of INV(ACE)*(COV−ACE). If the N= option requests fewer than the maximum number of canonical variables, only the specified number of eigenvalues are produced, with missing values filling out the observation.
SCORE	standardized canonical coefficients. The _NAME_ variable contains the name of the corresponding canonical variable as constructed from the PREFIX= option. The number of observations with _TYPE_=SCORE equals the number of canonical variables computed. To obtain the canonical variable scores, these coefficients should be multiplied by the standardized data.
RAWSCORE	raw canonical coefficients. To obtain the canonical variable scores, these coefficients should be multiplied by the raw (centered) data.

The OUTSTAT= data set can be used

- to initialize another execution of PROC ACECLUS
- to compute canonical variable scores with the SCORE procedure
- as input to the FACTOR procedure, specifying METHOD=SCORE, to rotate the canonical variables

Computational Resources

Let

n = number of observations

v = number of variables

i = number of iterations

Memory

The memory in bytes required by PROC ACECLUS is approximately

$$8(2n(v+1)+21v+5v^2)$$

bytes. If you request the PP or QQ option, an additional $4n(n-1)$ bytes are needed.

Time

The time required by PROC ACECLUS is roughly proportional to

$$2nv^2+10v^3+i\left(\frac{n^2v}{2}+nv^2+5v^3\right)$$

Displayed Output

Unless the SHORT option is specified, the ACECLUS procedure displays the following items:

- Means and Standard Deviations of the input variables
- the **S** matrix, labeled COV: Total Sample Covariances
- the name or value of the matrix used for the Initial Within-Cluster Covariance Estimate
- the Threshold value if the PROPORTION= option is specified

For each iteration, PROC ACECLUS displays

- the Iteration number
- RMS Distance, the root mean square distance between all pairs of observations
- the Distance Cutoff (u) for including pairs of observations in the estimate of the within-cluster covariances, which equals the RMS distance times the threshold
- the number of Pairs Within Cutoff
- the Convergence Measure (e_i) as specified by the METRIC= option

If the SHORT option is not specified, PROC ACECLUS also displays the **A** matrix, labeled ACE: Approximate Covariance Estimate Within Clusters.

The ACECLUS procedure displays a table of eigenvalues from the canonical analysis containing the following items:

- Eigenvalues of Inv(ACE)*(COV−ACE)
- the Difference between successive eigenvalues
- the Proportion of variance explained by each eigenvalue
- the Cumulative proportion of variance explained

If the SHORT option is not specified, PROC ACECLUS displays

- the Eigenvectors or raw canonical coefficients
- the standardized eigenvectors or standard canonical coefficients

ODS Table Names

PROC ACECLUS 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 15, “Using the Output Delivery System.”

Table 16.3. ODS Tables Produced in PROC ACECLUS

ODS Table Name	Description	Statement	Option
ConvergenceStatus	Convergence status	PROC	default
DataOptionInfo	Data and option information	PROC	default
Eigenvalues	Eigenvalues of Inv(ACE)*(COV-ACE)	PROC	default
Eigenvectors	Eigenvectors (raw canonical coefficients)	PROC	default
InitWithin	Initial within-cluster covariance estimate	PROC	INITIAL=INPUT
IterHistory	Iteration history	PROC	default
SimpleStatistics	Simple statistics	PROC	default
StdCanCoef	Standardized canonical coefficients	PROC	default
Threshold	Threshold value	PROC	PROPORTION=
TotSampleCov	Total sample covariances	PROC	default
Within	Approximate covariance estimate within clusters	PROC	default

Example

Example 16.1. Transformation and Cluster Analysis of Fisher Iris Data

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on fifty iris specimens from each of three species, *Iris setosa*, *I. versicolor*, and *I. virginica*. Mezzich and Solomon (1980) discuss a variety of cluster analyses of the iris data.

In this example PROC ACECLUS is used to transform the data, and the clustering is performed by PROC FASTCLUS. Compare this with the example in Chapter 27, "The FASTCLUS Procedure." The results from the FREQ procedure display fewer misclassifications when PROC ACECLUS is used. The following statements produce Output 16.1.1 through Output 16.1.5.

```
proc format;
   value specname
      1='Setosa    '
      2='Versicolor'
      3='Virginica ';
run;

data iris;
   title 'Fisher (1936) Iris Data';
   input SepalLength SepalWidth PetalLength PetalWidth Species @@;
   format Species specname.;
```

```
   label SepalLength='Sepal Length in mm.'
         SepalWidth ='Sepal Width in mm.'
         PetalLength='Petal Length in mm.'
         PetalWidth ='Petal Width in mm.';
   symbol = put(species, specname10.);
   datalines;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2 67 31 56 24 3
63 28 51 15 3 46 34 14 03 1 69 31 51 23 3 62 22 45 15 2
59 32 48 18 2 46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3 58 27 51 19 3
68 32 59 23 3 51 33 17 05 1 57 28 45 13 2 62 34 54 23 3
77 38 67 22 3 63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3 70 32 47 14 2
64 32 45 15 2 61 28 40 13 2 48 31 16 02 1 59 30 51 18 3
55 24 38 11 2 63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3 44 32 13 02 1
67 33 57 21 3 50 35 16 06 1 58 26 40 12 2 44 30 13 02 1
77 28 67 20 3 63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1 51 38 16 02 1
61 30 49 18 3 48 34 19 02 1 50 30 16 02 1 50 32 12 02 1
61 26 56 14 3 64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3 49 30 14 02 1
51 35 14 02 1 56 30 45 15 2 58 27 41 10 2 50 34 16 04 1
46 32 14 02 1 60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3 58 27 51 19 3
57 29 42 13 2 72 30 58 16 3 54 34 15 04 1 52 41 15 01 1
71 30 59 21 3 64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2 55 42 14 02 1
49 31 15 02 1 77 26 69 23 3 60 22 50 15 3 54 39 17 04 1
66 29 46 13 2 52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2 58 27 39 12 2
47 32 13 02 1 46 31 15 02 1 69 32 57 23 3 62 29 43 13 2
74 28 61 19 3 59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3 67 25 58 18 3
49 31 15 01 1 67 31 47 15 2 63 23 44 13 2 54 37 15 02 1
56 30 41 13 2 63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3 69 31 54 21 3
54 39 13 04 1 51 35 14 03 1 72 36 61 25 3 65 32 51 20 3
61 29 47 14 2 56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1 48 30 14 01 1
45 23 13 03 1 57 25 50 20 3 57 38 17 03 1 51 38 15 03 1
55 23 40 13 2 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3 67 30 50 17 2
63 33 60 25 3 53 37 15 02 1
;

proc aceclus data=iris out=ace p=.02 outstat=score;
   var SepalLength SepalWidth PetalLength PetalWidth ;
run;

legend1 frame cframe=ligr cborder=black position=center
        value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
```

```
proc gplot data=ace;
   plot can2*can1=Species/
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
   format Species specname. ;
run;
proc fastclus data=ace maxc=3 maxiter=10 conv=0 out=clus;
   var can:;
run;
proc freq;
   tables cluster*Species;
run;
```

Output 16.1.1. Using PROC ACECLUS to Transform Fisher's Iris Data

```
                    Fisher (1936) Iris Data

                     The ACECLUS Procedure

        Approximate Covariance Estimation for Cluster Analysis


          Observations         150    Proportion      0.0200
          Variables              4    Converge       0.00100


                 Means and Standard Deviations
                              Standard
    Variable           Mean    Deviation    Label

    SepalLength     58.4333       8.2807    Sepal Length in mm.
    SepalWidth      30.5733       4.3587    Sepal Width in mm.
    PetalLength     37.5800      17.6530    Petal Length in mm.
    PetalWidth      11.9933       7.6224    Petal Width in mm.

  Initial Within-Cluster Covariance Estimate = Full Covariance Matrix
```

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

COV: Total Sample Covariances

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	68.5693512	-4.2434004	127.4315436	51.6270694
SepalWidth	-4.2434004	18.9979418	-32.9656376	-12.1639374
PetalLength	127.4315436	-32.9656376	311.6277852	129.5609396
PetalWidth	51.6270694	-12.1639374	129.5609396	58.1006264

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Threshold = 0.334211

Iteration History

Iteration	RMS Distance	Distance Cutoff	Pairs Within Cutoff	Convergence Measure
1	2.828	0.945	408.0	0.465775
2	11.905	3.979	559.0	0.013487
3	13.152	4.396	940.0	0.029499
4	13.439	4.491	1506.0	0.046846
5	13.271	4.435	2036.0	0.046859
6	12.591	4.208	2285.0	0.025027
7	12.199	4.077	2366.0	0.009559
8	12.121	4.051	2402.0	0.003895
9	12.064	4.032	2417.0	0.002051
10	12.047	4.026	2429.0	0.000971

Algorithm converged.

ACE: Approximate Covariance Estimate Within Clusters

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	11.73342939	5.47550432	4.95389049	2.02902429
SepalWidth	5.47550432	6.91992590	2.42177851	1.74125154
PetalLength	4.95389049	2.42177851	6.53746398	2.35302594
PetalWidth	2.02902429	1.74125154	2.35302594	2.05166735

Output 16.1.2. Eigenvalues, Raw Canonical Coefficients, and Standardized Canonical Coefficients

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Eigenvalues of Inv(ACE)*(COV-ACE)

	Eigenvalue	Difference	Proportion	Cumulative
1	63.7716	61.1593	0.9367	0.9367
2	2.6123	1.5561	0.0384	0.9751
3	1.0562	0.4167	0.0155	0.9906
4	0.6395		0.00939	1.0000

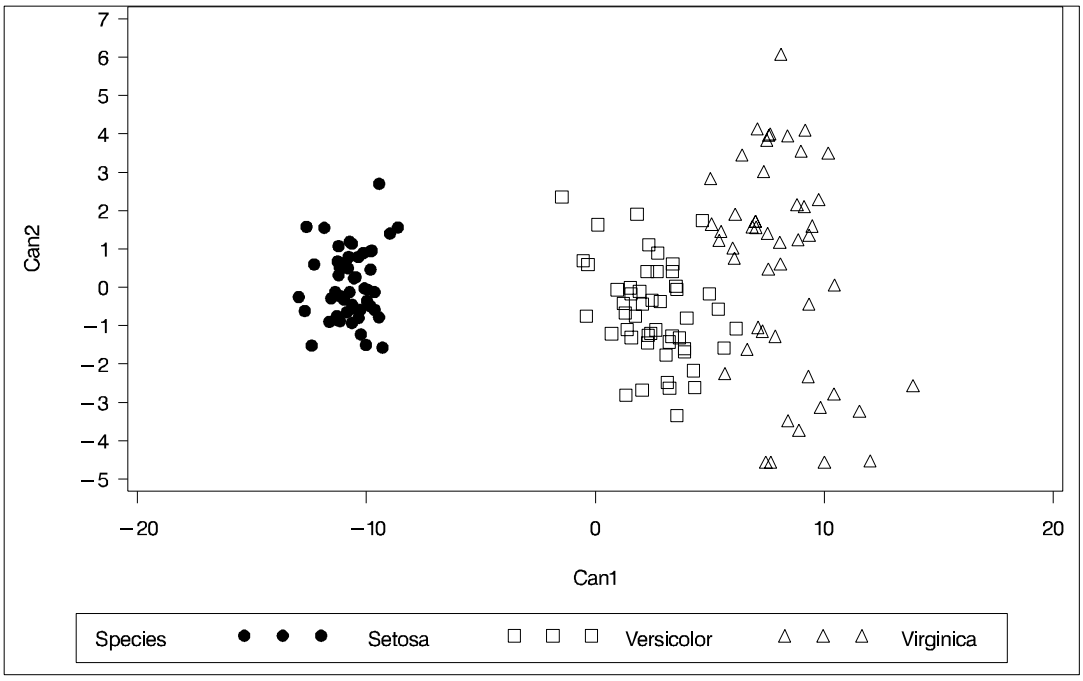
Eigenvectors (Raw Canonical Coefficients)

		Can1	Can2	Can3	Can4
SepalLength	Sepal Length in mm.	-.012009	-.098074	-.059852	0.402352
SepalWidth	Sepal Width in mm.	-.211068	-.000072	0.402391	-.225993
PetalLength	Petal Length in mm.	0.324705	-.328583	0.110383	-.321069
PetalWidth	Petal Width in mm.	0.266239	0.870434	-.085215	0.320286

Standardized Canonical Coefficients

		Can1	Can2	Can3	Can4
SepalLength	Sepal Length in mm.	-0.09944	-0.81211	-0.49562	3.33174
SepalWidth	Sepal Width in mm.	-0.91998	-0.00031	1.75389	-0.98503
PetalLength	Petal Length in mm.	5.73200	-5.80047	1.94859	-5.66782
PetalWidth	Petal Width in mm.	2.02937	6.63478	-0.64954	2.44134

Output 16.1.3. Plot of Transformed Iris Data: PROC PLOT



Output 16.1.4. Clustering of Transformed Iris Data: Partial Output from PROC FASTCLUS

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=3 Maxiter=10 Converge=0

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster
1	50	1.1016	5.2768		3
2	50	1.8880	6.8298		3
3	50	1.4138	5.3152		2

Cluster Summary

Cluster	Distance Between Cluster Centroids
1	13.2845
2	5.8580
3	5.8580

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
Can1	8.04808	1.48537	0.966394	28.756658
Can2	1.90061	1.85646	0.058725	0.062389
Can3	1.43395	1.32518	0.157417	0.186826
Can4	1.28044	1.27550	0.021025	0.021477
OVER-ALL	4.24499	1.50298	0.876324	7.085666

Pseudo F Statistic = 520.80

Approximate Expected Over-All R-Squared = 0.80391

Cubic Clustering Criterion = 5.179

WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	Can1	Can2	Can3	Can4
1	-10.67516964	0.06706906	0.27068819	0.11164209
2	8.12988211	0.52566663	0.51836499	0.14915404
3	2.54528754	-0.59273569	-0.78905317	-0.26079612

Cluster Standard Deviations

Cluster	Can1	Can2	Can3	Can4
1	0.953761025	0.931943571	1.398456061	1.058217627
2	1.799159552	2.743869556	1.270344142	1.370523175
3	1.572366584	1.393565864	1.303411851	1.372050319

Output 16.1.5. Crosstabulation of Cluster by Species for Fisher's Iris Data: PROC FREQ

The FREQ Procedure

Table of CLUSTER by Species

CLUSTER(Cluster) Species

Frequency / Percent / Row Pct / Col Pct	Setosa	Versicolor	Virginica	Total
1	50 33.33 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	50 33.33
2	0 0.00 0.00 0.00	2 1.33 4.00 4.00	48 32.00 96.00 96.00	50 33.33
3	0 0.00 0.00 0.00	48 32.00 96.00 96.00	2 1.33 4.00 4.00	50 33.33
Total	50 33.33	50 33.33	50 33.33	150 100.00

References

Art, D., Gnanadesikan, R., and Kettenring, R. (1982), "Data-based Metrics for Cluster Analysis," *Utilitas Mathematica*, 21A, 75–99.

Everitt, B.S. (1980), *Cluster Analysis*, Second Edition, London: Heineman Educational Books Ltd.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Hartigan, J.A. (1975), *Clustering Algorithms*, New York: John Wiley & Sons, Inc.

Kettenring, R. (1984), personal communication.

Mezzich, J.E and Solomon, H. (1980), *Taxonomy and Behavioral Science*, New York: Academic Press, Inc.

Puri, M.L. and Sen, P.K. (1971), *Nonparametric Methods in Multivariate Analysis*, New York: John Wiley & Sons, Inc.

Rouncefield, M. (1995), "The Statistics of Poverty and Inequality," *Journal of Statistics Education*, 3(2). [Online]: [http://www.stat.ncsu.edu/info/jse], accessed Dec. 19, 1997.

Wolfe, J.H. (1970),"Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research*, 5, 329–350.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999