SAS/STAT® 14.3 User's Guide
The ACECLUS Procedure

This document is an individual chapter from *SAS/STAT® 14.3 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS/STAT® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.3 User's Guide

Copyright © 2017, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

September 2017

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to http://support.sas.com/thirdpartylicenses.

Chapter 24
The ACECLUS Procedure

Contents

Overview: ACECLUS Procedure **893**
Background 894
Getting Started: ACECLUS Procedure **899**
Syntax: ACECLUS Procedure **905**
PROC ACECLUS Statement 905
BY Statement 910
FREQ Statement 911
VAR Statement 911
WEIGHT Statement 911
Details: ACECLUS Procedure **911**
Missing Values 911
Output Data Sets 912
Computational Resources 913
Displayed Output 914
ODS Table Names 915
ODS Graphics 915
Example: ACECLUS Procedure **916**
Example 24.1: Transformation and Cluster Analysis of Fisher Iris Data 916
References **921**

Overview: ACECLUS Procedure

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 needs to be known. PROC ACECLUS is useful for preprocessing data to be subsequently clustered by the CLUSTER or FASTCLUS procedure.

Many clustering methods perform well with spherical clusters but poorly with elongated elliptical clusters (Everitt 1980, pp. 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-crossproducts (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 31, "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 might 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 might 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 $\mathbf{A}$ by 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 by 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 might 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 by 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 the Fisher (1936) iris data, consisting of measurements of petal and sepal length and width for 50 specimens from each of three iris species, are summarized in Table 24.1. 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 20 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 following data:

- 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 those from 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 24.1 Number of Misclassified and Unclassified Observations Using the Fisher (1936) Iris Data

	Clustering Method			
Data	***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 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: ACECLUS Procedure

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 by 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[1] 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 &$20. @@;
   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. Germany

   ... more lines ...

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, respectively. The $20. format in the INPUT statement specifies that the variable Country is a

[1] 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.

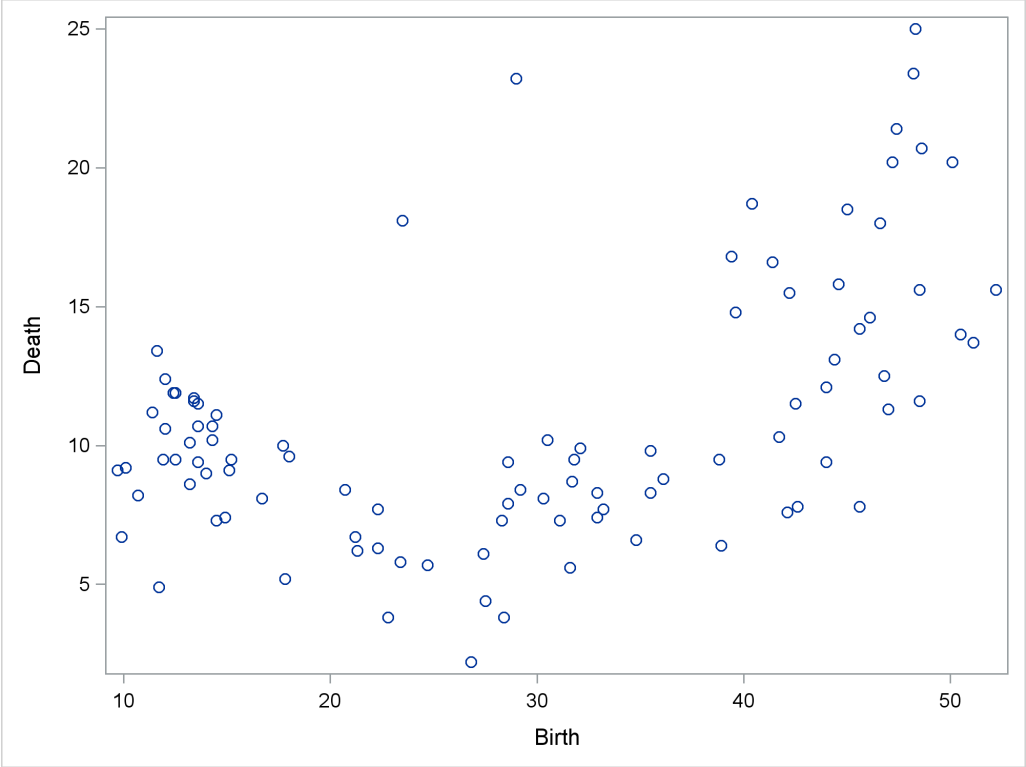
character variable with a length of 20. The preceding & enables the reading of blanks in the middle of the country names. The double trailing at sign (@@) in the INPUT statement specifies that observations are input from each line until all values have been read.

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

```
proc sgplot data=poverty;
   scatter y=Death x=Birth;
run;
```

The plot, displayed in Figure 24.1, 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 might be poorly separated and elongated. Data with poorly separated or elongated clusters must be transformed.

Figure 24.1 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 by 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 3 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 Figure 24.2 through Figure 24.5.

Figure 24.2 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. Figure 24.2 next displays the means, standard deviations, and sample covariance matrix of the analysis variables.

Figure 24.2 Means, Standard Deviations, and Covariance Matrix from the ACECLUS Procedure

The ACECLUS Procedure

Approximate Covariance Estimation for Cluster Analysis

Observations	97	**Proportion**	0.0300
Variables	3	**Converge**	0.00100

Means and Standard Deviations

Variable	Mean	Standard 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

The type of matrix used for the initial within-cluster covariance estimate is displayed in Figure 24.3. 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.

Figure 24.3 Table of Iteration History from the ACECLUS Procedure

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Threshold =	0.292815

Figure 24.3 *continued*

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 24.3 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 24.4 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).

Figure 24.4 Approximate Within-Cluster Covariance Estimates

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

The next three columns of the eigenvalue table (Figure 24.4) 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 24.5. 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.

Figure 24.5 Raw and Standardized Canonical Coefficients from the ACECLUS Procedure

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

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 a subsequent step to display cluster membership. 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 PROC TREE 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 analysis variables Birth, Death, and InfantDeath into the output SAS data set New.

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

```
proc sgplot data=new;
   scatter y=Death x=Birth / group=cluster;
   keylegend / title="Cluster Membership";
run;

proc sgplot data=new;
   scatter y=can2 x=can1 / group=cluster;
   keylegend / title="Cluster Membership";
run;
```

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

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

Figure 24.6 and Figure 24.7 display the separation of the clusters when three clusters are calculated.

Figure 24.6 Scatter Plot of Poverty Data, Identified by Cluster

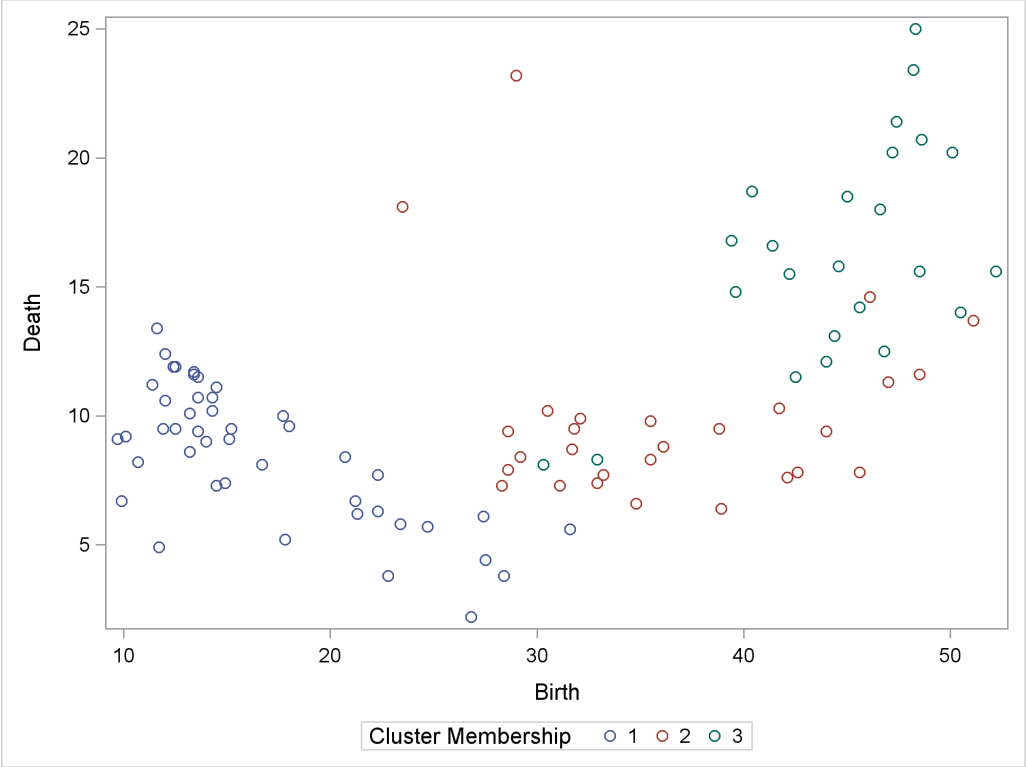
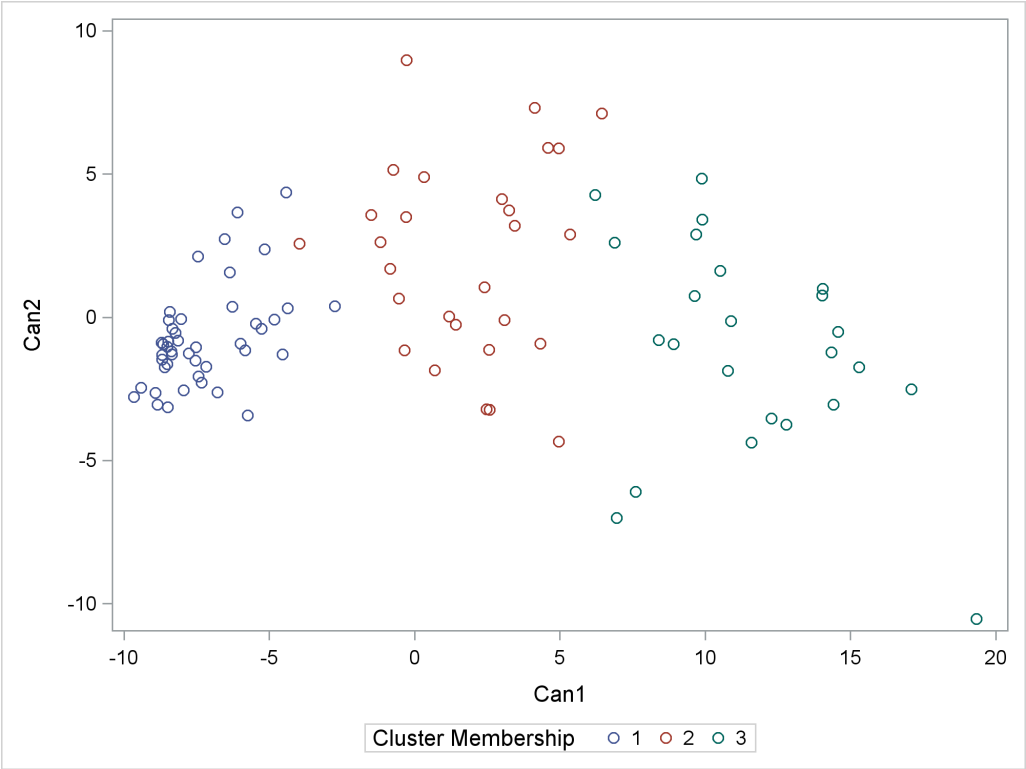


Figure 24.7 Scatter Plot of Canonical Variables



Syntax: ACECLUS Procedure

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 invokes the ACECLUS procedure. Table 24.2 summarizes the options available in the PROC ACECLUS statement. These options are also 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 24.2 Summary of PROC ACECLUS Statement Options

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

The following list provides details about the options.

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 by 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 A, "Special SAS Data Sets," for descriptions of CORR, COV, UCORR, UCOV, and SSCP data sets. See the section "Output Data Sets" on page 912 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 | 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*×(*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 20, "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 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*.

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 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 < =*plot-request* **>**

PLOTS < =(*plot-request* **< ...** *plot-request* **>) >**

specifies options that control the details of the plots. When you specify only one plot request, you can omit the parentheses.

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

```
ods graphics on;

proc aceclus plots=all;
   var _numeric_;
run;

ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 615 in Chapter 21, "Statistical Graphics Using ODS."

The *plot-requests* include the following:

ALL
produces both P-P and Q-Q plots.

NONE
suppresses all plots.

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

QQ
produces a Q-Q plot of a power transformation of the distances between pairs of observations computed in the last iteration. **CAUTION:** The Q-Q plot can require an enormous amount of computer time.

When ODS Graphics is not enabled, you can specify the PP and QQ options (but not PLOTS=(PP QQ)) to create printer plots.

By default, no plots are produced.

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

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

where *totfq* is the number of total nonmissing observations.

PP
produces a P-P 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, ..., CAN*N*. 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 about the VALIDVARNAME= system option, see *SAS System Options: Reference*.

QQ
produces a Q-Q plot of a power transformation of the distances between pairs of observations computed in the last iteration. **CAUTION:** The Q-Q plot can 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 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 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 by using the DATASETS procedure (in Base SAS software).

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 about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *SAS Visual Data Management and Utility 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 nonintegral 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: ACECLUS Procedure

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:

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.
RAWSCORE	raw canonical coefficients. To obtain the canonical variable scores, these coefficients should be multiplied by the raw data centered by means obtained from the observation with _TYPE_='MEAN'.

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, using means obtained from the observation with _TYPE_='MEAN' and standard deviations obtained from the observation with _TYPE_='STD'.

The OUTSTAT= data set can be used in the following conditions:

- 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 following items:

- 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 following items:

- 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 about ODS, see Chapter 20, "Using the Output Delivery System."

Table 24.3 ODS Tables Produced by 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

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 615 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 614 in Chapter 21, "Statistical Graphics Using ODS."

You can reference every graph produced through ODS Graphics with a name. The names of the graphs that PROC ACECLUS generates are listed in Table 24.4.

Table 24.4 Graphs Produced by PROC ACECLUS

ODS Graph Name	Plot Description
PPLOT	P-P plot
QQPLOT	Q-Q plot

Example: ACECLUS Procedure

Example 24.1: 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 50 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 iris data, which is available from the Sashelp library, and the clustering is performed by PROC FASTCLUS. Compare this with the example in Chapter 40, "The FASTCLUS Procedure." The results from the FREQ procedure display fewer misclassifications when PROC ACECLUS is used.

The following statements produce Output 24.1.1 through Output 24.1.5:

```
title 'Fisher (1936) Iris Data';

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

proc sgplot data=ace;
   scatter y=can2 x=can1 / group=Species;
   keylegend / title="Species";
run;

proc fastclus data=ace maxc=3 maxiter=10 conv=0 out=clus;
   var can:;
run;

proc freq;
   tables cluster*Species;
run;
```

Output 24.1.1 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

Output 24.1.1 *continued*

Means and Standard Deviations

Variable	Mean	Standard Deviation	Label
SepalLength	58.4333	8.2807	Sepal Length (mm)
SepalWidth	30.5733	4.3587	Sepal Width (mm)
PetalLength	37.5800	17.6530	Petal Length (mm)
PetalWidth	11.9933	7.6224	Petal Width (mm)

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.

Output 24.1.2 Eigenvalues, Raw Canonical Coefficients, and Standardized Canonical Coefficients

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

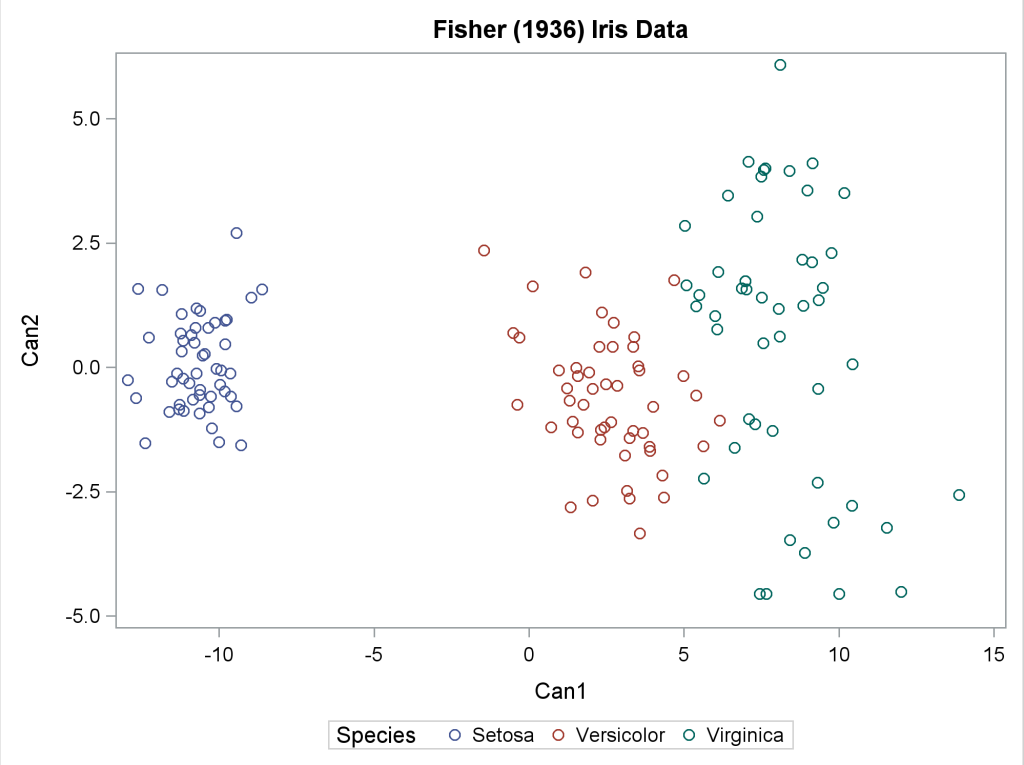
Output 24.1.2 *continued*

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 (mm)	-.012009	-.098074	-.059852	0.402352
SepalWidth	Sepal Width (mm)	-.211068	-.000072	0.402391	-.225993
PetalLength	Petal Length (mm)	0.324705	-.328583	0.110383	-.321069
PetalWidth	Petal Width (mm)	0.266239	0.870434	-.085215	0.320286

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

Output 24.1.3 Plot of Transformed Iris Data: PROC SGPLOT



Output 24.1.4 Clustering of Transformed Iris Data: Partial Output from PROC FASTCLUS

Fisher (1936) Iris Data

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	Distance Between Cluster Centroids
1	50	1.4138	5.3152		2	5.8580
2	50	1.8880	6.8298		1	5.8580
3	50	1.1016	5.2768		1	13.2845

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 values above are invalid for correlated variables.

Cluster Means

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

Cluster Standard Deviations

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

Output 24.1.5 Crosstabulation of Cluster by Species for Fisher's Iris Data: PROC FREQ

Fisher (1936) Iris Data

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of CLUSTER by Species			
		Species(Iris Species)		
CLUSTER(Cluster)	Setosa	Versicolor	Virginica	Total
1	0 0.00 0.00 0.00	48 32.00 96.00 96.00	2 1.33 4.00 4.00	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	50 33.33 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	50 33.33
Total	50 33.33	50 33.33	50 33.33	150 100.00

References

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

Everitt, B. S. (1980). *Cluster Analysis*. 2nd ed. London: Heineman Educational Books.

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.

Kettenring, J. R. (1984). Personal communication.

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

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

Rouncefield, M. (1995). "The Statistics of Poverty and Inequality." Journal of Statistics Education Data Archive. Accessed May 22, 2009. http://www.amstat.org/publications/jse/v3n2/datasets.rouncefield.html.

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

Subject Index

ACECLUS procedure
  analyzing data in groups, 899, 916
  between-cluster SSCP matrix, 894
  clustering methods, 907
  compared with other procedures, 897
  computational resources, 913
  controlling iterations, 896
  decomposition of the SSCP matrix, 894
  eigenvalues and eigenvectors, 902, 908, 912, 914, 915
  initial estimates, 895, 907
  memory requirements, 913
  missing values, 911
  output data sets, 912
  output table names, 915
  time requirements, 913
  within-cluster SSCP matrix, 894
analyzing data in groups
  ACECLUS procedure, 899
approximate covariance estimation
  clustering, 894

between-cluster SSCP matrix
  ACECLUS procedure, 894

cluster
  elliptical, 893
clustering
  approximate covariance estimation, 894
clustering methods
  ACECLUS procedure, 907
computational resources
  ACECLUS procedure, 913
convergence criterion
  ACECLUS procedure, 906

decomposition of the SSCP matrix
  ACECLUS procedure, 894

eigenvalues and eigenvectors
  ACECLUS procedure, 902, 908, 912, 914, 915

initial estimates
  ACECLUS procedure, 907

memory requirements
  ACECLUS procedure, 913
missing values
  ACECLUS procedure, 911

OUT= data sets
  ACECLUS procedure, 912
output data sets
  ACECLUS procedure, 912
output table names
  ACECLUS procedure, 915

pooled within-cluster covariance matrix
  definition, 894

time requirements
  ACECLUS procedure, 913

within-cluster SSCP matrix
  ACECLUS procedure, 894

Syntax Index

ABSOLUTE option
    PROC ACECLUS statement, 906
ACECLUS procedure
    syntax, 905
ACECLUS procedure, BY statement, 910
ACECLUS procedure, FREQ statement, 911
ACECLUS procedure, PROC ACECLUS statement, 905
    ABSOLUTE option, 906
    CONVERGE= option, 906
    DATA= option, 907
    INITIAL= option, 907
    MAXITER= option, 907
    METHOD= option, 907
    METRIC= option, 907
    MPAIRS= option, 908
    N= option, 908
    NOPRINT option, 908
    OUT= option, 908
    OUTSTAT= option, 908
    P= option, 909
    PERCENT= option, 909
    PLOTS= option, 908
    PP option, 909
    PREFIX= option, 909
    PROPORTION= option, 909
    QQ option, 909
    SHORT option, 910
    SINGULAR= option, 910
    T= option, 910
    THRESHOLD= option, 910
ACECLUS procedure, VAR statement, 911
ACECLUS procedure, WEIGHT statement, 911

BY statement
    ACECLUS procedure, 910

CONVERGE= option
    PROC ACECLUS statement, 906

DATA= option
    PROC ACECLUS statement, 907

INITIAL= option
    PROC ACECLUS statement, 907

MAXITER= option
    PROC ACECLUS statement, 907
METHOD= option
    PROC ACECLUS statement, 907
METRIC= option
    PROC ACECLUS statement, 907
MPAIRS= option
    PROC ACECLUS statement, 908

N= option
    PROC ACECLUS statement, 908
NOPRINT option
    PROC ACECLUS statement, 908

OUT= option
    PROC ACECLUS statement, 908
OUTSTAT= option
    PROC ACECLUS statement, 908

P= option
    PROC ACECLUS statement, 909
PERCENT= option
    PROC ACECLUS statement, 909
PLOTS= option
    PROC ACECLUS statement, 908
PP option
    PROC ACECLUS statement, 909
PREFIX= option
    PROC ACECLUS statement, 909
PROC ACECLUS statement, *see* ACECLUS procedure
PROPORTION= option
    PROC ACECLUS statement, 909

QQ option
    PROC ACECLUS statement, 909

SHORT option
    PROC ACECLUS statement, 910
SINGULAR= option
    PROC ACECLUS statement, 910

T= option
    PROC ACECLUS statement, 910
THRESHOLD= option
    PROC ACECLUS statement, 910