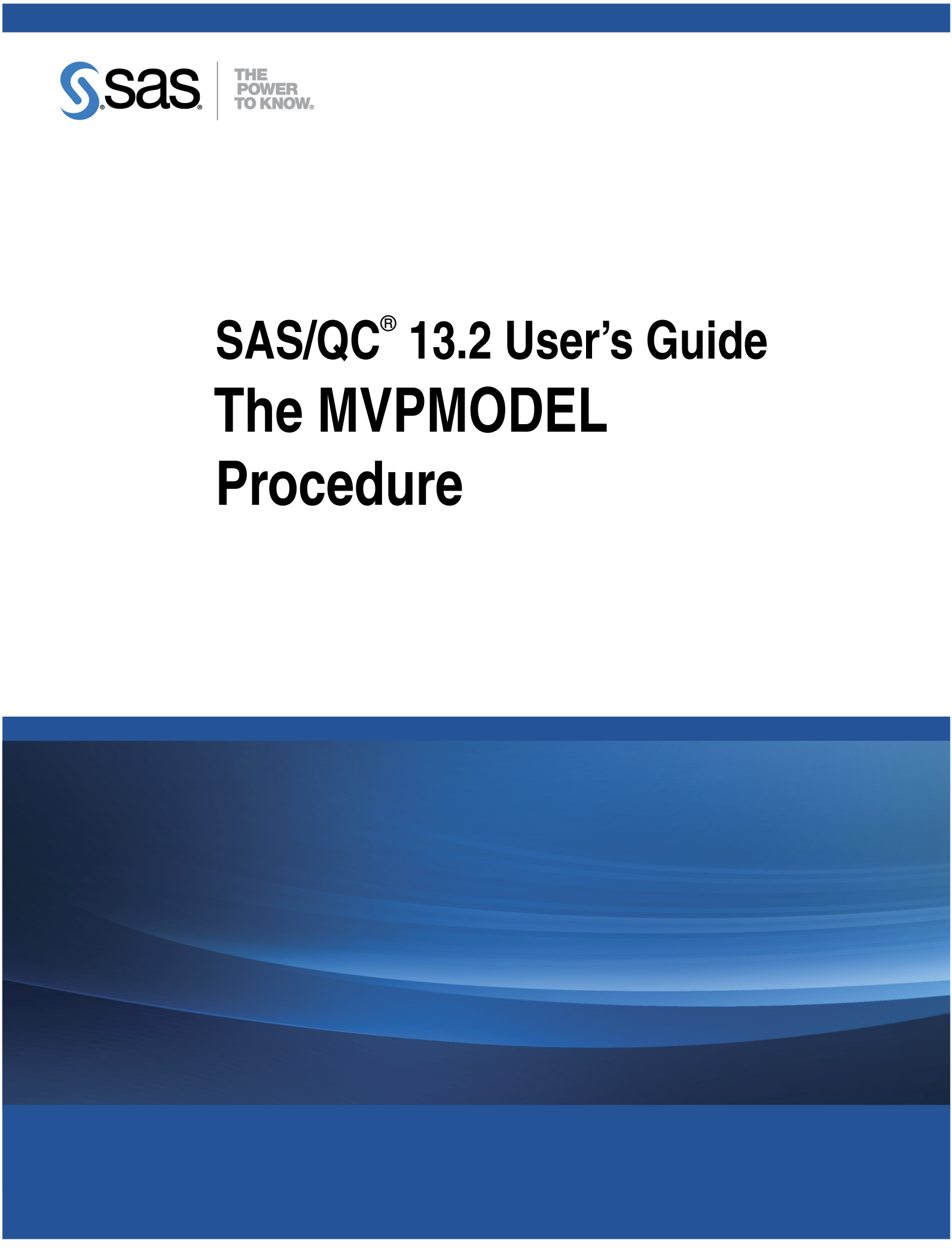


SAS/QC® 13.2 User's Guide
The MVPMODEL Procedure

This document is an individual chapter from *SAS/QC® 13.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2014. *SAS/QC® 13.2 User's Guide*. Cary, NC: SAS Institute Inc.

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

August 2014

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

Chapter 12
The MVPMODEL Procedure

Contents

Overview: MVPMODEL Procedure

The MVPMODEL procedure is used in conjunction with the MVPMONITOR and MVPDIAGNOSE procedures to monitor multivariate process variation over time, to determine whether the process is stable, and to detect and diagnose changes in a stable process. Collectively these three procedures are referred to as the *MVP procedures*. See Chapter 10, "Introduction to Multivariate Process Monitoring Procedures," for a description of how the MVP procedures work together, and Chapter 13, "The MVPMONITOR Procedure," and Chapter 11, "The MVPDIAGNOSE Procedure," for details about the other MVP procedures.

The MVPMODEL procedure provides computational and graphical tools for building a principal component model from multivariate process data in which the measured variables are continuous and correlated. This model then serves as input to the other MVP procedures, described in Chapter 11, "The MVPDIAGNOSE Procedure," and Chapter 13, "The MVPMONITOR Procedure." The MVPMONITOR procedure creates various multivariate control charts, including T^2 charts and SPE (squared prediction error) charts, which are used to detect and diagnose changes in the process. Multivariate control charts can detect unusual variation that would not be detected by individually monitoring the variables with univariate control charts, such as Shewhart charts.

The MVPMODEL procedure implements principal component analysis (PCA) techniques that evolved in the field of chemometrics for monitoring hundreds or even thousands of correlated process variables; see Kourti and MacGregor (1995, 1996) for an introduction. These techniques differ from the classical multivariate T^2 chart in which Hotelling's T^2 statistic is computed as a distance from the multivariate mean scaled by the covariance matrix of the variables; see Alt (1985). Instead, principal component methods compute T^2 based on a small number of principal components that model most of the variation in the data.

One advantage of PCA methods over the classical T^2 chart is that they avoid computational issues that arise when the process measurement variables are collinear and their covariance matrix is nearly singular. A second advantage is that they offer diagnostic tools for interpreting unusual values of T^2. A third advantage is that by projecting the data to a low-dimensional subspace, a principal component model more adequately describes the variation in a multivariate process, which is often driven by a small number of underlying factors that are not directly observable.

Using the MVP Procedures

There are two primary scenarios for using the MVP procedures:

1. To determine whether a process is stable, you can construct T^2 and SPE charts from an existing set of process measurements (this is referred to as a Phase I analysis). First, build a principal component model with the MVPMODEL procedure, saving the measurements and the computed observationwise statistics (including T^2 and SPE) in an OUT= data set. Then specify this data set as a HISTORY= input data set for the MVPMONITOR procedure to create T^2 and SPE charts. Contribution plots indicate which of the original variables are involved in unusual variation displayed by the T^2 and SPE charts. Follow-up action might be needed to adjust the process and eliminate unusual variation signaled by the charts.

2. To detect changes in a stable process, you can construct T^2 and SPE charts from newly acquired data by using the principal component model developed from previous data (this is referred to as a Phase II analysis). You can save information about the model in the OUTLOADINGS= data set created by the MVPMODEL procedure. Specify this data set as a LOADINGS= input data set and specify the new data as a DATA= input data set to create T^2 and SPE charts with the MVPMONITOR procedure.

Functionality of the MVPMODEL Procedure

The MVPMODEL procedure performs principal component analysis (PCA) on multivariate process measurement data that consist of p continuous variables that are assumed to be correlated. The input data set for

PROC MVPMODEL provides the values of the p variables that are to be analyzed.

The MVPMODEL procedure computes the following quantities:

- the loadings from the principal component analysis
- the eigenvalues from the principal component analysis, which are the variances of the principal component variables
- the scores from the principal component analysis
- the T^2 statistic for each observation
- the SPE (squared prediction error) statistic for each observation, also known as SSE, Q, or DModX

By default, principal components are computed from the correlation matrix of the variables. Optionally, they can be computed from their covariance matrix instead. The number of principal components in the model (denoted by j, where $j \leq p$) can be specified or determined by one of several cross validation methods.

By default, PROC MVPMODEL outputs the correlation matrix of the input variables and the eigenvalues of the correlation matrix. When ODS Graphics is enabled, the output can also include the following plots:

- a scree plot and a variance-explained plot of the principal components (these plots are created by default)
- when using cross validation, plots of W and root mean PRESS (predicted residual sum of squares) for each principal component
- pairwise score plots of principal component scores
- pairwise loading plots of principal component loadings

PROC MVPMODEL saves information about the principal component model in the following two output data sets, which can subsequently serve as inputs to the MVPMONITOR and MVPDIAGNOSE procedures:

- an output data set which contains all the variables and observations in the input data set together with observationwise statistics, such as scores, residuals, T^2, and SPE
- an output data set that contains the j loadings for each process variable and the eigenvalues associated with each of the principal components

Getting Started: MVPMODEL Procedure

This example illustrates the basic features of the MVPMODEL procedure by using airline flight delay data available from the U.S. Bureau of Transportation Statistics at `http://www.transtats.bts.gov`. The example applies multivariate process monitoring to flight delays.

Suppose you want to use a principal component model to create T^2 and SPE charts to monitor the variation in flight delays. These charts are appropriate because the data are multivariate and correlated.

The following statements create a SAS data set named MWflightDelays to contain the average flight delays for flights that originate in the midwestern United States by airline. The data set contains variables for nine airlines: AA (American Airlines), CO (Continental Airlines), DL (Delta Airlines), F9 (Frontier Airlines), FL (AirTran Airways), NW (Northwest Airlines), UA (United Airlines), US (US Airways), and WN (Southwest Airlines).

```
data MWflightDelays;
   format flightDate MMDDYY8.;
   label flightDate='Date';
   input flightDate :MMDDYY8. AA CO DL F9 FL NW UA US WN;
   datalines;
02/01/07 14.9  7.1  7.9  8.5 14.8  4.5  5.1 13.4  5.1
02/02/07 14.3  9.6 14.1  6.2 12.8  6.0  3.9 15.3 11.4
02/03/07 23.0  6.1  1.7  0.9 11.9 15.2  9.5 18.4  7.6
02/04/07  6.5  6.3  3.9 -0.2  8.4 18.8  6.2  8.8  8.0
02/05/07 12.0 14.1  3.3 -1.3 10.0 13.1 22.8 16.5 11.5
02/06/07 31.9  8.6  4.9  2.0 11.9 21.9 29.0 15.5 15.2

   ... more lines ...

02/16/07 31.2 20.8 15.2 20.1  9.1 12.9 22.9 36.4 16.4
;
```

The observations for a given date are the average flight delays in minutes of flights that depart from the Midwest. For example, on February 2, 2007, F9 (Frontier Airlines) flights departed an average of 6.2 minutes late.

Preliminary Analysis

The following statements use the MVPMODEL procedure to conduct a preliminary principal component analysis:

```
ods graphics on;
proc mvpmodel data=MWflightDelays;
   var AA CO DL F9 FL NW UA US WN;
run;
```

The DATA= option specifies the input data set, which contains the process measurement variables. The VAR statement specifies the process measurement variables to be analyzed. The ODS GRAPHICS ON statement enables ODS Graphics, which is used to produce plots for interpreting the model.

The procedure first outputs a summary of the model and the data, as shown in Figure 12.1.

Figure 12.1 Summary of Model and Data Information

The MVPMODEL Procedure

Data Set	WORK.MWFLIGHTDELAYS
Number of Variables	9
Missing Value Handling	Exclude
Number of Observations Read	16
Number of Observations Used	16
Number of Principal Components	9

This output includes the number of principal components in the model and the number of variables. In this case the procedure produces a model with nine principal components by default, because there are nine process variables.

Next, the procedure outputs the correlation matrix shown in Figure 12.2.

Figure 12.2 Correlation Matrix

Correlation Matrix

	AA	CO	DL	F9	FL	NW	UA	US	WN
AA	1.0000	0.5640	0.5206	0.4874	0.5403	0.4860	0.6466	0.7856	0.5506
CO	0.5640	1.0000	0.7855	0.6580	0.8519	0.6421	0.7672	0.8415	0.6526
DL	0.5206	0.7855	1.0000	0.8231	0.7598	0.4782	0.4951	0.7463	0.4525
F9	0.4874	0.6580	0.8231	1.0000	0.5119	0.2279	0.3509	0.6832	0.3914
FL	0.5403	0.8519	0.7598	0.5119	1.0000	0.6807	0.6975	0.8207	0.7186
NW	0.4860	0.6421	0.4782	0.2279	0.6807	1.0000	0.6715	0.5598	0.3970
UA	0.6466	0.7672	0.4951	0.3509	0.6975	0.6715	1.0000	0.7540	0.7736
US	0.7856	0.8415	0.7463	0.6832	0.8207	0.5598	0.7540	1.0000	0.8152
WN	0.5506	0.6526	0.4525	0.3914	0.7186	0.3970	0.7736	0.8152	1.0000

There are strong correlations (greater than 0.8) between variable pairs F9 and DL, CO and FL, and US and WN. This is not surprising, because these pairs of airlines have closely located hubs or focus cities.

The procedure also outputs the eigenvalue and variance information shown in Figure 12.3.

Figure 12.3 Eigenvalue and Variance Information

Eigenvalues of the Correlation Matrix

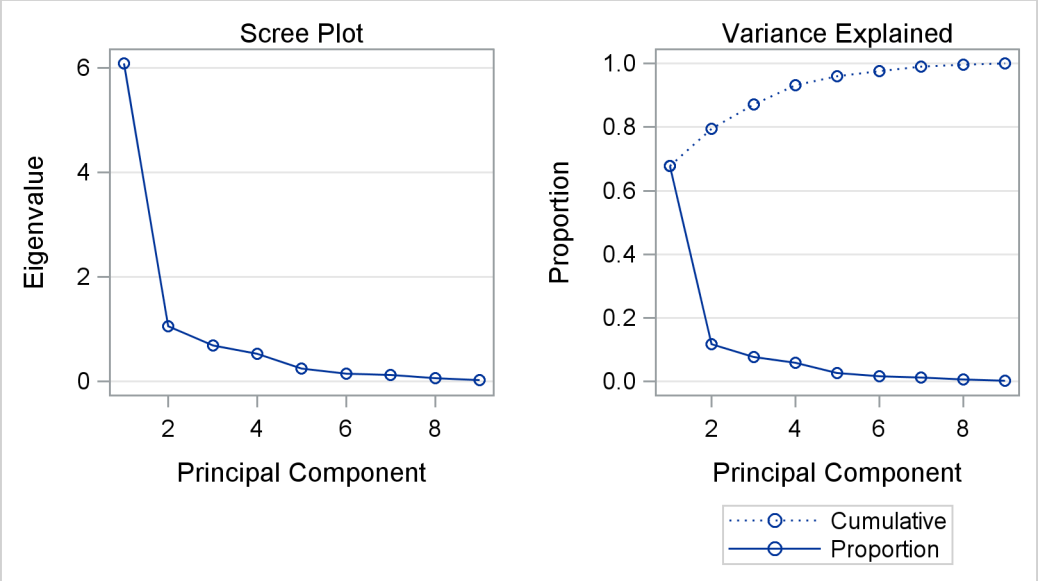
	Eigenvalue	Difference	Proportion	Cumulative
1	6.09006397	5.02872938	0.6767	0.6767
2	1.06133459	0.36642409	0.1179	0.7946
3	0.69491050	0.16102099	0.0772	0.8718
4	0.53388951	0.28357563	0.0593	0.9311
5	0.25031387	0.09537517	0.0278	0.9589
6	0.15493870	0.03339131	0.0172	0.9762
7	0.12154739	0.06166364	0.0135	0.9897
8	0.05988375	0.02676604	0.0067	0.9963
9	0.03311771		0.0037	1.0000

The eigenvalues are the variances of the principal components, and the proportions reflect the relative amount of variance explained by each component. The eigenvalues and the proportions are ordered from largest to smallest. Recall that principal components are orthogonal linear combinations of the variables that maximize variance in orthogonal directions.

More than 85% of the variance is explained by the first three principal components, as shown in the cumulative variance column. This suggests that a model with three principal components is adequate; this is confirmed by the plots in Figure 12.4.

Figure 12.4 shows a paneled display, with a scree plot in the left panel and a variance-explained plot in the right panel.

Figure 12.4 Scree Plot and Variance-Explained Plot



The scree plot shows the eigenvalues for each principal component. Traditionally, the scree plot has been recommended as an aid in selecting the number of principal components for the model by examining the "knee" in the plot (Mardia, Kent, and Bibby 1979). The variance-explained plot shows both the proportion of variance and the cumulative variance explained by the principal components.

Building a Principal Component Model

To build a model that has only three principal components, you can use the NCOMP= option as shown in the following statements:

```
proc mvpmodel data=MWflightDelays ncomp=3 plots=(all score(labels=on))
              out=outDelays;
   var AA CO DL F9 FL NW UA US WN;
run;
```

The PLOTS=ALL option requests all possible plots, which include pairwise plots of the principal component scores and loadings in addition to the default scree plot and variance-explained plot. The OUT= option produces an output data set called outDelays that contains principal component scores, T^2 statistics, SPE statistics, residuals, and more, as described in the section "Output Data Sets" on page 935. Note that ODS Graphics is still enabled, so you do not need to specify the ODS GRAPHICS ON statement here.

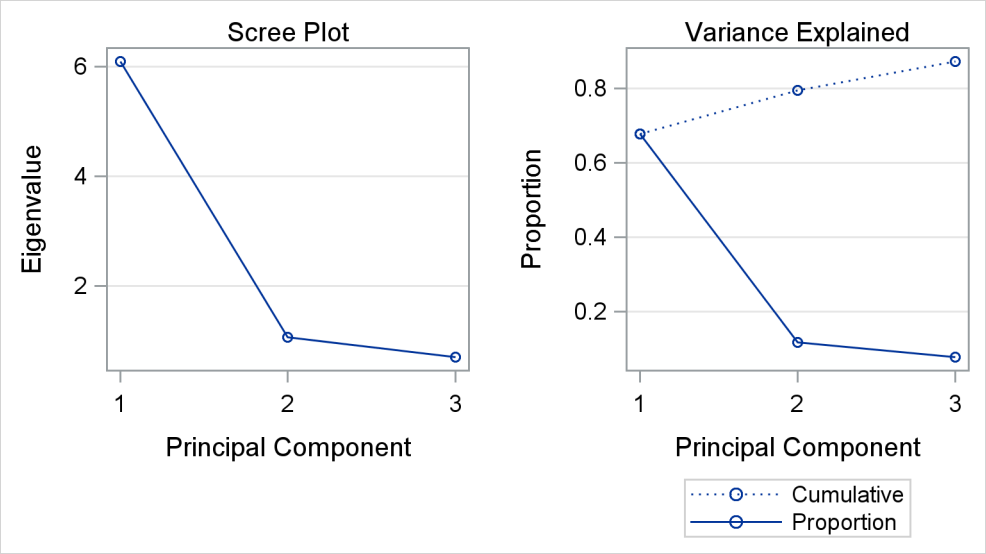
The correlation matrix is the same as in Figure 12.2. The eigenvalue information, scree plot, and variance-explained plot are similar to those in Figure 12.3 and Figure 12.4. However, the use of the NCOMP=3 option results in outputs that show information only for the three components in the model, as seen in Figure 12.5 and Figure 12.6.

Figure 12.5 Eigenvalue and Variance Information

The MVPMODEL Procedure

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.09006397	5.02872938	0.6767	0.6767
2	1.06133459	0.36642409	0.1179	0.7946
3	0.69491050		0.0772	0.8718

Figure 12.6 Scree Plot and Variance-Explained Plot



Also, the model summary, shown in Figure 12.7, is different because there are now only three principal components in the model.

Figure 12.7 Summary of Model and Data Information

The MVPMODEL Procedure

Data Set	WORK.MWFLIGHTDELAYS
Number of Variables	9
Missing Value Handling	Exclude
Number of Observations Read	16
Number of Observations Used	16
Number of Principal Components	3

The outDelays output data set that is partially listed in Figure 12.8 contains T^2 and SPE statistics based on the model that has three principal components, in addition to the original variables and other observationwise statistics.

Figure 12.8 Partial Listing of Output Data Set outDelays

flightDate	AA	CO	DL	F9	FL	NW	UA	US	WN	Prin1	Prin2	Prin3	_NOBS_	_TSQUARE_
02/01/07	14.9	7.1	7.9	8.5	14.8	4.5	5.1	13.4	5.1	-1.08708	1.20953	-0.03839	16	1.57457
02/02/07	14.3	9.6	14.1	6.2	12.8	6.0	3.9	15.3	11.4	-0.65786	1.26249	0.11447	16	1.59169
02/03/07	23.0	6.1	1.7	0.9	11.9	15.2	9.5	18.4	7.6	-0.86457	-0.73183	0.29270	16	0.75065
02/04/07	6.5	6.3	3.9	-0.2	8.4	18.8	6.2	8.8	8.0	-1.50578	-0.69718	1.32511	16	3.35709
02/05/07	12.0	14.1	3.3	-1.3	10.0	13.1	22.8	16.5	11.5	-0.63903	-1.11141	0.38617	16	1.44549

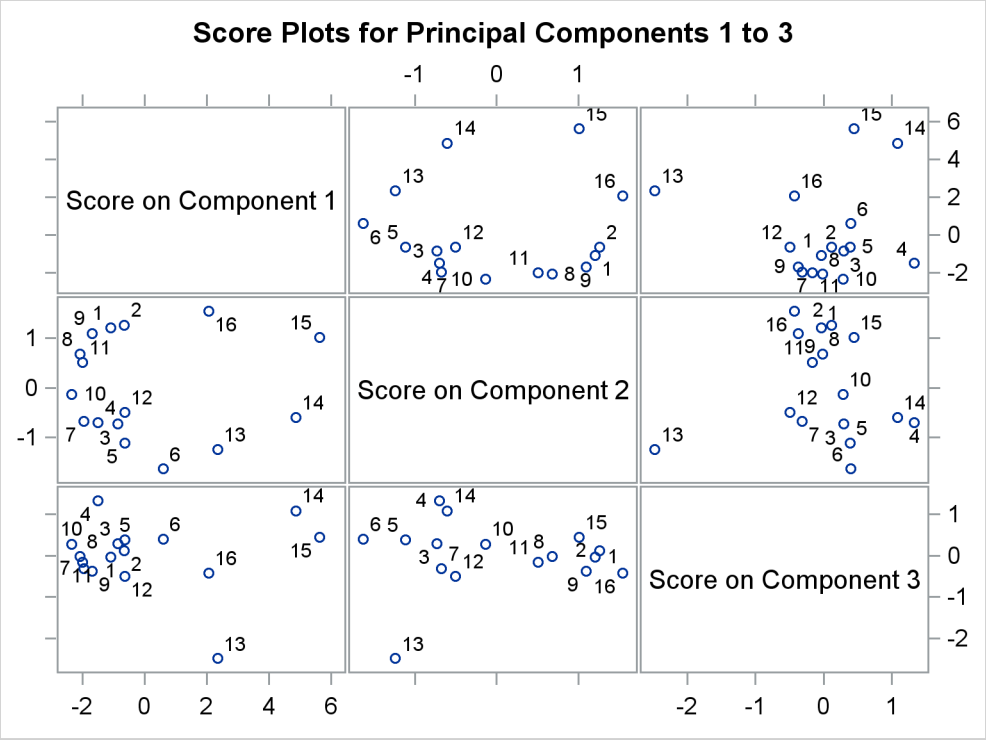
R_AA	R_CO	R_DL	R_F9	R_FL	R_NW	R_UA	R_US	R_WN	_SPE_
-0.05779	-0.18178	-0.01835	-0.15280	0.87457	-0.37864	-0.06037	-0.12896	-0.02300	0.98911
-0.17802	-0.16663	0.68047	-0.62682	0.49289	-0.35101	-0.27027	-0.14161	0.41169	1.54414
0.54274	-0.30297	-0.20552	-0.02408	0.31360	0.21270	-0.49772	0.24287	-0.23829	0.93626
-0.25729	-0.24974	0.05624	0.05279	-0.02427	0.29305	-0.44076	0.13493	0.50899	0.69253
-0.44128	0.28274	0.09998	-0.15050	0.00176	-0.36866	0.39124	0.07233	-0.06265	0.60545

The variables Prin1, Prin2, and Prin3 contain the principal component scores. Variables R_AA through R_WN are the residuals for the process variables. The contents of an OUT= data set are described in detail in the section "Output Data Sets" on page 935. See the section "Principal Component Analysis" on page 930 for computational details of the results saved in the output data set.

You can use an OUT= data set as an input to the MVPMONITOR and MVPDIAGNOSE procedures. The MVPMONITOR procedure produces control charts for the T^2 and SPE statistics. Control charts that are created from the outDelays data set are shown in Example 12.2 and in the MVPMONITOR procedure chapter.

The PLOTS=ALL option produces score plots for pairs of principal components in the model. By default, the score plots are displayed in a matrix. You can specify the PLOTS(SCORES(UNPACK)) option to display the score plots as separate graphs. The score plot matrix is shown in Figure 12.9.

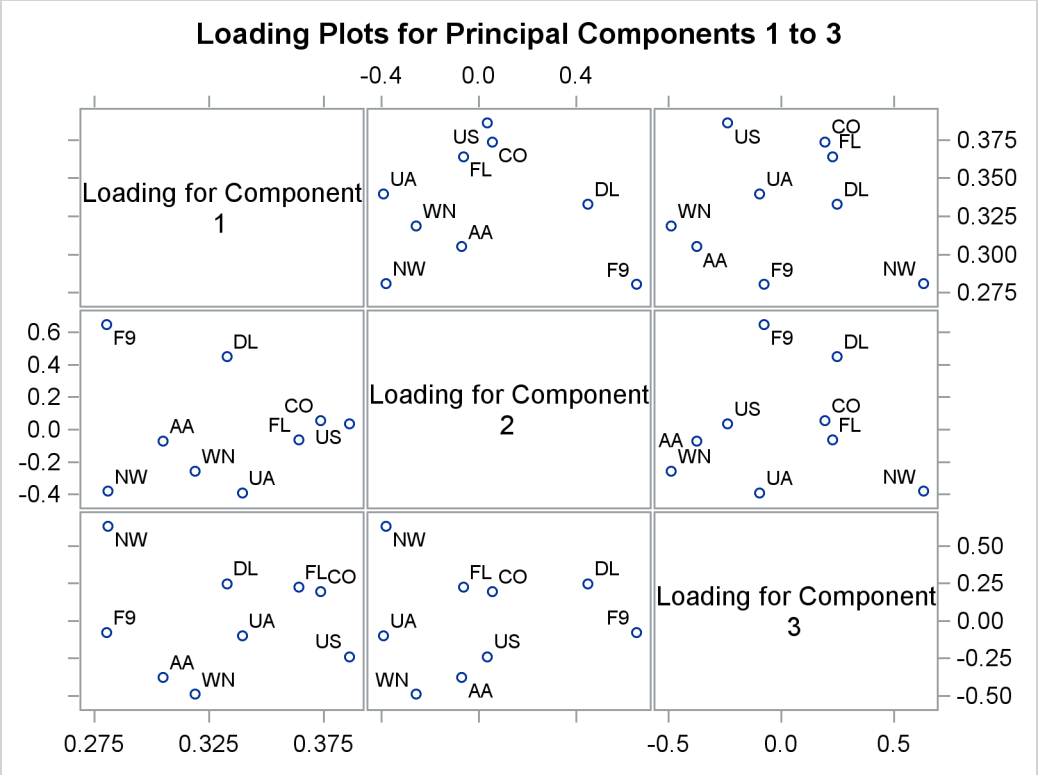
Figure 12.9 Score Plots for Principal Components 1–3



A score plot is a scatter plot of the scores for two principal components. The labels indicate the observation numbers of the points. By examining clusters and outliers in these plots, you can better understand the relationships among the observations and the variation in the process. For example, points 13 through 16 are extreme points in the direction of the first principal component. The directions of the principal components are not uniquely determined, so you need the loadings and external information to interpret them. These points represent flight delays between February 13, 2007, and February 16, 2007, when there was a major winter storm in the Midwest.

Figure 12.10 displays the loading plots that are produced. Loading plots are also displayed in a matrix by default, and they can be unpacked into separate graphs with the PLOT(LOADINGS(UNPACK)) option.

Figure 12.10 Loading Plot for Principal Components 1–3



A loading plot is a scatter plot of the variable loadings for a pair of principal components, and it helps you understand the relationships among the variables. Loadings are the variable coefficients in the eigenvectors (linear combinations of variables) that define the principal component. The loadings explain how variables contribute to the linear combination. Here, the loadings for the first principal component are all positive and all similar in value, which suggests that the first principal component describes the average delay. The second principal component appears to be a contrast between the delays of F9, DL, CO, and US and those of the remaining airlines. See the section "Principal Component Analysis" on page 930 for more information about interpreting principal component loadings and scores.

Syntax: MVPMODEL Procedure

The following statements are available in PROC MVPMODEL:

PROC MVPMODEL < *options* > **;**
BY *variables* **;**
ID *variables* **;**
VAR *variables* **;**

The following sections describe the PROC MVPMODEL statement and then describe the other statements in alphabetical order.

PROC MVPMODEL Statement

PROC MVPMODEL < *options* > **;**

The PROC MVPMODEL statement invokes the MVPMODEL procedure and optionally identifies input and output data sets, specifies details of the analyses performed, and controls displayed output. Table 12.1 summarizes the *options*.

Table 12.1 Summary of PROC MVPMODEL Statement Options

option	**Description**
COV	Computes the principal components from the covariance matrix
CV=	Performs cross validation to select the number of principal components
DATA=	Specifies the input data set
MISSING=	Specifies how observations with missing values are handled
NCOMP=	Specifies the number of principal components to extract
NOCENTER	Suppresses centering of process variables before fitting the model
NOCVSTDIZE	Suppresses re-centering and rescaling of process variables before each model is fit in the cross validation
NOPRINT	Suppresses the display of all output
NOSCALE	Suppresses scaling of process variables before fitting the model
OUT=	Specifies the output data set
OUTLOADINGS=	Specifies the output data set for loadings (eigenvectors)
PLOTS=	Requests and specifies details of plots
PREFIX=	Specifies the prefix for naming principal component score variables in the OUT= data set
RPREFIX=	Specifies the prefix for naming residual variables in the OUT= data set
STDSCORES	Standardizes the principal component scores

You can specify the following *options*.

COV

computes the principal components from the covariance matrix. By default, the correlation matrix is analyzed. The COV option causes variables with large variances to be more strongly associated with components that have large eigenvalues, and it causes variables with small variances to be more strongly associated with components that have small eigenvalues. You should not specify the COV option unless the units in which the variables are measured are comparable or the variables are standardized in some way.

NOTE: Specifying the COV option has the same effect as specifying the NOSCALE option.

CV=ONE

CV=BLOCK < **(***cv-block-options***)** >

CV=SPLIT < **(***cv-split-options***)** >

CV=RANDOM < **(***cv-random-options***)** >

specifies that cross validation be performed to determine the number of principal components and specifies the method to be used. If you do not specify the CV= option, no cross validation is performed.

In cross validation, the input data are repeatedly divided into a *training set*, which is used to compute a model, and a *test set*, which is used to test the model fit. The cross validation that is performed here is along both observations and variables, as described in Eastment and Krzanowski (1982), which is a more detailed version of the "alternative scheme" of Wold (1978). The observations and variables are separately divided into groups. Each test set is the intersection of one observation group and one variable group, so the number of test sets that are used is the product of the number of observation groups and the number of variable groups. See the section "Cross Validation (Experimental)" on page 933 for more information.

NOTE: The CV= option is experimental in this release.

CV=ONE requests *one-at-a-time* cross validation, in which each observation group contains one observation and each variable group contains one variable. This approach is very computationally intensive because it computes $n \times p$ separate principal component models for each potential number of principal components, where n is the number of observations in the input data set and p is the number of process variables.

CV=BLOCK requests *blocked* cross validation, in which observation groups consist of blocks of *nobs* consecutive observations and variable groups consist of blocks of *nvar* consecutive variables. You can specify the following *cv-block-options* in parentheses after the CV=BLOCK option:

NOBS=*nobs*
: specifies that observation groups consist of blocks of *nobs* consecutive observations from the input data. For example, if you specify NOBS=*8*, the first group contains observations 1 through 8, the second group contains observations 9 through 16, and so on. The default is 7.

NVAR=*nvar*
: specifies that variable groups consist of blocks of *nvar* consecutive variables from the input data. For example, if you specify NVAR=*3*, the first group contains variables 1 through 3, the second group contains variables 4 through 6, and so on. The default is 7.

CV=SPLIT requests *split-sample* cross validation, in which observation groups are formed by selecting every *nobs*th observation and variable groups are formed by selecting very *nvar*th variable. You can specify the following *cv-split-options* in parentheses after the CV=SPLIT option:

NOBS=*nobs*
: specifies that observation groups be created by selecting every *nobs*th observation from the input data. For example, if you specify NOBS=*8*, the first group contains observations {1, 9, 17, . . . }, the second group contains observations {2, 10, 18, . . . }, and so on. The default is 7.

NVAR=*nvar*
: specifies that variable groups be created by selecting every *nvar*th variable from the input data. For example, if you specify NVAR=*5*, the first group contains variables {1, 6, 11, . . . }, the second group contains variables {2, 7, 12, . . . }, and so on. The default is 7.

CV=RANDOM requests that observations and variables be assigned to groups randomly. You can specify the following *cv-random-options* in parentheses after the CV=RANDOM option:

NITEROBS=*nogrp*
: specifies the number of observation groups. The default is 10.

NITERVAR=*nvgrp*
specifies the number of variable groups. The default is 10.

NTESTOBS=*nobs*
specifies the number of observations in each observation group. The default is one-tenth the total number of observations.

NTESTVAR=*nvar*
specifies the number of variables in each variable group. The default is one-tenth the total number of variables.

SEED=*n*
specifies an integer used to start the pseudorandom number generator for selecting the random test set. If you do not specify a seed or if you specify a value less than or equal to zero, the seed is generated by default from reading the time of day from the computer's clock.

NOTE: You cannot specify the CV= option together with the NCOMP= option.

DATA=*SAS-data-set*
specifies the input SAS data set to be analyzed. If the DATA= option is omitted, the procedure uses the most recently created SAS data set.

MISSING=AVG | NONE
specifies how observations with missing values are to be handled in computing the fit. MISSING=AVG specifies that the fit be computed by replacing missing values of a process variable with the average of its nonmissing values. The default is MISSING=NONE, which excludes observations with missing values for any process variables from the analysis.

NCOMP=*n* **| ALL**
specifies the number of principal components to extract. The default is $\min\{15, p, N\}$, where p is the number of process variables and N is the number of observations (runs). You can specify NCOMP=ALL to override the limit of 15 principal components. You cannot specify the NCOMP= option together with the CV= option. If the number of nonzero eigenvalues of the correlation matrix is less than the number of components specified, p, then the p will be reset to the number of nonzero eigenvalues.

NOCENTER
suppresses centering of the process variables before fitting. This is useful if the variables are already centered and scaled. See the section "Centering and Scaling" on page 934 for more information.

NOCVSTDIZE
suppresses re-centering and rescaling of the process variables before each model is fit in the cross validation. See the section "Centering and Scaling" on page 934 for more information.

NOPRINT
suppresses the display of all results, both tabular and graphical. This is useful when you want to produce only output data sets.

NOSCALE
suppresses scaling of the process variables before fitting. This is useful if the variables are already centered and scaled.

NOTE: Specifying the NOSCALE option has the same effect as specifying the COV option.

OUT=*SAS-data-set*
creates an output data set that contains all the original data from the input data set, principal component scores, and multivariate summary statistics. See the section "Output Data Sets" on page 935 for details.

OUTLOADINGS=*SAS-data-set*
creates an output data set that contains the loadings for the principal components and the eigenvalues of the correlation (or covariance) matrix. See the section "Output Data Sets" on page 935 for details.

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

PLOTS <(*global-plot-options***) > <= (***plot-request* < *(options)* > **<...** *plot-request* < *(options)* > **>) >**
controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

```
plots=none
plots=score
plots=loadings
```

ODS Graphics must be enabled before you request plots. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS" (*SAS/STAT User's Guide*).

You can specify the following *global-plot-options*:

FLIP
interchanges the X-axis and Y-axis dimensions for all score and loading plots.

NCOMP=*n*
specifies that pairwise score and loading plots be produced for the first *n* principal components. The default is 5 or the total number of components $j(\geq 2)$, whichever is smaller. If $n > j$, then the default is NCOMP=*j*. Be aware that the number of score or loading plots produced ($\frac{n\times(n-1)}{2}$) grows quadratically as *n* increases.

ONLY
suppresses the default plots. Only plots specifically requested are displayed. The default plots are the CV plot, when you specify the CV= option, and the scree and variation-explained plots otherwise.

You can specify the following *plot-requests*:

ALL
produces all appropriate plots.

CVPLOT
produces a plot that displays the results of the cross validation and R-square analysis. This plot requires that the CV= option be specified and in that case is displayed by default.

LOADINGS < (*loading-options*) >
produces a matrix of pairwise scatter plots of the principal component loadings. Use NCOMP=*n* to specify the number of principal components for which plots are produced, and use the FLIP option to interchange the default X-axis and Y-axis dimensions.

You can specify the following *loading-options*:

FLIP
flips or interchanges the X-axis and Y-axis dimensions of the loading plots. Specify PLOTS=LOADING(FLIP) to flip the X-axis and Y-axis dimensions.

NCOMP=*n*
specifies that pairwise loading plots be produced for the first *n* principal components. The default is the value specified by the NCOMP= *global-plot-option*. If $n > j$, then the default is NCOMP=*j*. Be aware that the number of loading plots produced ($\frac{n\times(n-1)}{2}$) grows quadratically as *n* increases.

UNPACKPANEL
UNPACK
suppresses paneling of loading plots. By default, all the loading plots appear in a single output panel. Specify UNPACKPANEL to display each loading plot in a separate panel.

NONE
suppresses the display of all plots.

SCORES < (*score-options*) >
produces pairwise scatter plots of the principal component scores. You can use the NCOMP= option to control the number of plots that are displayed.

You can specify the following *score-options*:

ALPHA=*value*
specifies the probability used to compute a prediction ellipse that is overlaid on the score plot. The default is 0.05. If you specify the ALPHA= option, you do not need to specify the ELLIPSE option.

ELLIPSE
requests that a prediction ellipse be overlaid on the principal component score plots. The probability that a new observation falls outside the prediction ellipse is specified by the ALPHA= option.

FLIP
flips or interchanges the X-axis and Y-axis dimensions of the score plots. Specify PLOTS=SCORES(FLIP) to flip the X-axis and Y-axis dimensions.

GROUP=*variable*
specifies a variable in the input data set used to group the points on the score plots. Points with different GROUP= variable values are plotted using different markers and colors to distinguish the groups.

LABELS=ON | OFF | OUTSIDE
specifies which points in the score plots to label. Specify LABELS=ON to label all points and LABELS=OFF to label none of the points. Points are labeled with the values of the first variable listed in the ID statement, or the observation number if no ID statement is specified.

If you specify the ELLIPSE and UNPACKPANEL options, you can specify LABELS=OUTSIDE to label only the points outside the confidence ellipse.

The default is ON if you specify UNPACKPANEL and OFF otherwise.

NCOMP=*n*
specifies that pairwise score plots be produced for the first *n* principal components. The default is the value specified by the NCOMP= *global-plot-option*. If $n > j$, then the default is NCOMP=*j*. Be aware that the number of loading plots produced ($\frac{n \times (n-1)}{2}$) grows quadratically as *n* increases.

UNPACKPANEL
suppresses paneling of score plots. By default, all the score plots appear in a single output panel. Specify UNPACKPANEL to display each score plot in a separate panel.

SCREE < UNPACK >
EIGEN
EIGENVALUE
produces a scree plot of eigenvalues and a variance-explained plot. By default, both plots are produced in a panel. Specify PLOTS= SCREE(UNPACKPANEL) to display each plot in a separate panel. This plot is produced by default unless you specify the CV= option.

PREFIX=*name*
specifies a prefix for naming the principal component scores in the OUT= data set. By default, the names are Prin1, Prin2, . . . , Prin*j*. If you specify PREFIX=ABC, the components are named ABC1, ABC2, ABC3, and so on. The number of characters in the prefix plus the number of digits in *j* should not exceed the current name length defined by the VALIDVARNAME= system option.

RPREFIX=*name*
specifies a prefix for naming the residual variables in the OUT= data set. The default is R_. Residual variable names are formed by appending process variable names to the prefix.

If the length of the resulting residual variable exceeds the maximum name length defined by the VALIDVARNAME= system option, characters are removed from the middle of the process variable name before it is appended to the residual prefix. For example, if you specify RPREFIX=*Residual_*, the maximum variable name length is 32, and there is a process variable named PrimaryThermometerReading, then the corresponding residual variable name is Residual_PrimaryThermeterReading.

STDSCORES
standardizes the principal component scores in the OUT= data set to unit variance. If you omit the STDSCORES option, the variances of the scores are equal to the corresponding eigenvalues. STDSCORES has no effect on the eigenvalues themselves.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC MVPMODEL 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 MVPMODEL procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

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

ID Statement

ID *variables* **;**

The first variable that is specified in the ID statement is used to label observations in score plots for principal components. If you do not specify an ID statement, then score plot points are labeled with their observation numbers.

The values of all ID variables are displayed in tooltips when you create HTML output and specify the IMAGEMAP option in the ODS GRAPHICS statement. See Chapter 21, "Statistical Graphics Using ODS" (*SAS/STAT User's Guide*), for details.

VAR Statement

VAR *variables* **;**

The VAR statement specifies the process variables and their order in the results. By default, if you omit the VAR statement, the MVPMODEL procedure analyzes all numeric variables that are not listed in the BY or ID statement.

Details: MVPMODEL Procedure

Classical T^2 Charts

Classical T^2 charts are defined as follows. Assume that there are n observations for p variables, denoted by $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i$ is a p-dimensional vector. The T^2 statistic for observation i is

$$T_i^2 = \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)' \mathbf{S}^{-1} \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)$$

where

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \quad , \quad \mathbf{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix}, \quad \bar{\mathbf{X}}_n = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

and

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)'$$

For purposes of deriving control limits for the T^2 chart, it is assumed that $\mathbf{X}_i$ has a p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu} = \left(\mu_1, \mu_2, \cdots, \mu_p\right)'$ and covariance matrix $\boldsymbol{\Sigma}$ for $i = 1, 2, \ldots, n$. The classical formulation of the T^2 chart does not involve a principal component model for the data, and it bases the computation of T^2 on the sample covariance matrix $\mathbf{S}$. See Alt (1985) for theoretical details and the section "Multivariate Control Charts" on page 2130 for an example.

A classical T^2 chart is equivalent to a T^2 chart based on a full principal component model (with p components), as discussed in the section "Relationship of Principal Components to Multivariate Control Charts" on page 931. See Example 12.2 for more information.

Principal Component Analysis

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). The application of principal components is discussed by Rao (1964), Cooley and Lohnes (1971), Gnanadesikan (1977), and Jackson (1991). Excellent statistical treatments of principal components are found in Kshirsagar (1972), Morrison (1976), and Mardia, Kent, and Bibby (1979).

Principal component modeling focuses on the number of components used. The analysis begins with an eigenvalue decomposition of the sample covariance matrix, $\mathbf{S}$,

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)'$$

as

$$\begin{aligned} \mathbf{S} &= \mathbf{PLP}' \\ \mathbf{P}'\mathbf{SP} &= \mathbf{L} \end{aligned}$$

where $\mathbf{L}$ is a diagonal matrix and $\mathbf{P}$ is an orthogonal matrix (Jackson 1991; Mardia, Kent, and Bibby 1979). The columns of $\mathbf{P}$ are the eigenvectors, and the diagonal elements of $\mathbf{L}$ are the eigenvalues. The eigenvectors are customarily scaled so that they have unit length.

A principal component, $\mathbf{t}_i$, is a linear combination of the original variables. The coefficients are the eigenvectors of the covariance matrix. The principal component scores for the ith observation are computed as

$$\mathbf{t}_i = \mathbf{P}'(\mathbf{x}_i - \bar{\mathbf{x}})$$

The principal components are sorted by descending order of the eigenvalues, which are equal to the variances of the components.

The eigenvectors are the principal component loadings. The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables. The scores on the first j principal components have the highest possible generalized variance of any set of j unit-length linear combinations of the original variables.

The first j principal components provide a least squares solution to the model

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

where $\mathbf{X}$ is an $n \times p$ matrix of the centered observed variables, $\mathbf{T}$ is the $n \times j$ matrix of scores on the first j principal components, $\mathbf{P}'$ is the $j \times p$ matrix of eigenvectors, and $\mathbf{E}$ is an $n \times p$ matrix of residuals. The first j principal components are the vectors (rows of $\mathbf{P}'$) that minimize trace($\mathbf{E}'\mathbf{E}$), the sum of all the squared elements in $\mathbf{E}$.

The first j principal components are the best linear predictors of the process variables among all possible sets of j variables, although any nonsingular linear transformation of the first j principal components provides equally good prediction. The same result is obtained by minimizing the determinant or the Euclidean norm of $\mathbf{E}'\mathbf{E}$ rather than the trace.

Relationship of Principal Components to Multivariate Control Charts

Multivariate control charts typically plot the T^2 statistic, which is a summary of multivariate variation. The classical T^2 statistic is defined in "Classical T^2 Charts" on page 930. When there is high correlation among the process variables, the correlation matrix is nearly singular. The subspace in which the process varies can be adequately explained by fewer variables than the original p variables. Thus, the principal component approach to multivariate control charts is to project the original p variables into a lower-dimensional subspace by using a model based on j principal components, where $j < p$.

The key to the relationship between principal components and multivariate control charts is the decomposition of the sample covariance matrix, $\mathbf{S}$, into the form $\mathbf{S} = \mathbf{PLP}'$, where $\mathbf{L}$ is a diagonal matrix (Jackson 1991; Mardia, Kent, and Bibby 1979). This is also the eigenvalue decomposition of $\mathbf{S}$, where the columns of $\mathbf{P}$ are the eigenvectors and the diagonal elements of $\mathbf{L}$ are the eigenvalues.

Equivalence of T^2 Statistics

The T^2 statistic that is produced by the full principal component model is equivalent to the classical T^2 statistic. This is seen in the matrix representation of the T^2 statistic computed from a principal component model that uses all p components,

$$T_i^2 = \left(\mathbf{t}_i - \bar{\mathbf{t}}_n\right)' \mathbf{L}_n^{-1} \left(\mathbf{t}_i - \bar{\mathbf{t}}_n\right)$$

Because $\bar{\mathbf{t}}_n$ is the zero matrix by construction, then

$$T_i^2 = \mathbf{t}_i' \mathbf{L}_n^{-1} \mathbf{t}_i$$

Because $\mathbf{t}_i = \mathbf{P}' \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)$, then

$$\begin{aligned} T_i^2 &= \mathbf{t}_i' \mathbf{L}_n^{-1} \mathbf{t}_i \\ &= \left(\mathbf{P}' \left(\mathbf{x_i} - \bar{\mathbf{x}}\right)\right)' \mathbf{L}_n^{-1} \left(\mathbf{P}' \left(\mathbf{x_i} - \bar{\mathbf{x}}\right)\right) \\ &= \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{P} \mathbf{L}_n^{-1} \mathbf{P}' \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \\ &= \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \end{aligned}$$

which is the classical form. Consequently the classical T^2 statistic can be expressed as a sum of squares,

$$T_i^2 = \frac{t_{i1}^2}{l_1^2} + \ldots + \frac{t_{ip}^2}{l_p^2}$$

where l_k^2 is the variance of the kth principal component.

Computing the T^2 and SPE Statistics

Creating a T^2 chart that is based on a principal component model begins with choosing the number (j) of principal components. Effectively, this involves selecting a subspace in $j < p$ dimensions and then creating a T^2 statistic based on that j-component model.

The T^2 statistic is meant to monitor variation in the model space. However, if variation appears in the $p - j$ subspace orthogonal to model space, then the model assumptions and physical process should be reexamined. Variation outside the model space can be detected with an SPE chart.

In a model with j principal components, the T^2 statistic is calculated as

$$T_i^2 = \frac{t_{i1}^2}{l_1^2} + \ldots + \frac{t_{ij}^2}{l_j^2}$$

where t_{ik} is the principal component score for the kth principal component of the ith observation and l_k is the standard deviation of t_{ik}.

The information in the remaining $p - j$ principal components is monitored with charts for the SPE statistic, which is calculated as

$$\begin{aligned} \mathrm{SPE}_i &= \sum_{k=j+1}^{p} e_{ik}^2 \\ &= \sum_{k=j+1}^{p} \left(x_{ik} - \hat{x}_{ik}\right)^2 \end{aligned}$$

Cross Validation (Experimental)

NOTE: The CV= option is experimental in this release.

You can use cross validation to choose the number of principal components in the model to avoid overfitting.

One method of choosing the number of principal components is to fit the model to only part of the available data (the *training set*) and to measure how well models with different numbers of extracted components fit the other part of the data (the *test set*). This is called *test set validation*. However, it is rare that you have enough make both parts large enough for pure test set validation to be useful. Alternatively, you can make several different divisions of the observed data into a training set and a test set. This is called *cross validation*. The MVPMODEL procedure supports four types of cross validation. In *one-at-a-time* cross validation, the first observation is held out as a single-element test set, with all other observations as the training set; next, the second observation is held out, then the third, and so on. Another method is to hold out successive blocks of observations as test sets—for example, observations 1 through 7, then observations 8 through 14, and so on; this is known as *blocked* validation. A similar method is *split-sample* cross validation, in which successive groups of widely separated observations are held out as the test set—for example, observations {1, 11, 21, ... }, then observations {2, 12, 22, ... }, and so on. Finally, test sets can be selected from the observed data randomly; this is known as *random-sample* cross validation.

Which cross validation method you should use depends on your data. The most common method is one-at-a-time validation (CV=ONE), but it is not appropriate when the observed data are serially correlated. In that case either blocked (CV=BLOCK) or split-sample (CV=SPLIT) validation might be more appropriate; you can select the number of test sets in blocked or split-sample validation by specifying options in parentheses after the CV= option. The numbers in parentheses are the number of test sets over the rows and columns. For more information, see the section "An Alternative Scheme" in Wold (1978), as well as Eastment and Krzanowski (1982), both of which describe the cross validation approach used here in more detail.

CV=ONE is the most computationally intensive of the cross validation methods, because it requires you to recompute the principal component model for every input observation. Using random subset selection with CV=RANDOM might lead different researchers to produce different principal component models from the same data (unless the same seed is used).

Whichever validation method you use, the number of principal components that are chosen is usually the one that optimizes some criterion or selection rule. Choices of a criterion include the ratio described by Wold (1978), the W statistic described by Eastment and Krzanowski (1982), and the predicted residual sum of squares (PRESS). The W statistic is used by the MVPMODEL procedure.

The method of choosing the number of principal components in the MVPMODEL procedure is described in Eastment and Krzanowski (1982). This method is a heuristic based on the ratio of the mean PRESS (MPRESS) to the degrees of freedom for the principal component model. First, the MPRESS is computed for models with 0 to $maxcomp$ principal components. The maximum number of components is $\min(15, nvar, nobs) - 1$ and can be further reduced to the number of nonzero eigenvalues in the covariance matrix. Second, for each of the i possible number of components, the W_i statistic is computed as

$$W_i = \frac{MPRESS(i-1) - MPRESS(i)}{D_i} \div \frac{MPRESS(i)}{D_R}$$

where $MPRESS = \frac{1}{np} PRESS$, D_i is the number of degrees of freedom used to fit the model with i principal components, and D_R is the remaining number of degrees of freedom.

Extracting too many components can lead to an overfit model, one that matches the training data too well, sacrificing predictive ability. Thus, if you specify the number of principal components in the model, you should not use cross validation to select the appropriate number of components for the final model, or you should consider the analysis to be preliminary and examine the results to determine the appropriate number of components for a subsequent analysis.

Centering and Scaling

By default, the variables are centered and scaled to have mean 0 and standard deviation 1. Without centering, both the mean variable value and the variation around that mean are involved in selecting principal component loadings. Scaling serves to place all process variables on an equal footing relative to their variation in the data. For example, if Time and Temp are two of the process variables, then scaling says that a change of std (Time) in Time is roughly equivalent to a change of std (Temp) in Temp.

The formulas that are used to compute the variation in the different centering and scaling cases are defined in the section "Definitional Formulas" in Chapter A, "Special SAS Data Sets" (*SAS/STAT User's Guide*). The definitional formula that is used when either the NOSCALE or COV option is specified is the COV formula. The definitional formula that is used when the NOCENTER option is specified is the UCORR formula. The definitional formula that is used when both the NOCENTER and NOSCALE options are specified is the UCOV formula. The default definitional formula, when no centering or scaling options are specified, is the CORR formula.

Missing Values

By default, observations that have missing process variables are simply excluded from the analysis. If you specify MISSING=AVG in the PROC MVPMODEL statement, then all observations in the input data set contribute to both the analysis and the OUT= data set. With MISSING=AVG, the fit is computed by replacing missing values of a process variable with the average of its nonmissing values.

Input Data Set

The input data set provides the set of process variables that are analyzed. You can specify the input data set by using the DATA= option in the PROC MVPMODEL statement. If you do not specify the DATA= option, the procedure uses the last data set created as its input data set.

The MVPMODEL procedure treats each observation in the DATA= data set as an individual multivariate observation. The observations do not need to be identified or sorted by time because the sequence of the data is not used to build the principal component model. If you provide a time variable in the input data set, it is preserved in the OUT= data set and can be used subsequently by the MVPMONITOR procedure to create control charts.

In basic applications of the MVPMODEL procedure, the observations in the DATA= data set represent measurements from a single process. You can build different principal component models for two or more processes by grouping their measurements in the DATA= data and processing them as BY groups.

In some applications, it is desirable to combine the data from two or more processes and build a common principal component model. This might be the case with processes that are peers in the sense that they are believed to share the same pattern of common cause variation. When you provide the MVPMONITOR procedure with a common model for a set of peer processes, it uses the model to construct identical control limits for each process. This enables you to decide whether a particular process exhibits unusual variation relative to the behavior of its peers.

Output Data Sets

OUT= Data Set

The OUT= data set contains all the variables in the input data set plus new variables that contain the principal component scores, residuals, and other computed values listed in Table 12.2.

The names of the score variables are formed by concatenating the value given by the PREFIX= option (or the default Prin, if PREFIX= is not specified) and the numbers 1, 2, ..., j, where j is the number of principal components in the model.

The names of the residual variables are formed by concatenating the value given by the RPREFIX= option (or the default R_, if RPREFIX= is not specified) and the names of the process variables used in the analysis. Residual variables are created only when the number of principal components in the model is less than the number of process measurement variables in the input data set.

Table 12.2 Computed Variables in the OUT= Data Set

Variable	Description
Prin1–Prin$j$	Principal component scores
R_var1–R_$varp$	Residuals
NOBS	Number of observations used in the analysis
SPE	Squared prediction error (SPE)
TSQUARE	T^2 statistic computed from principal component scores

OUTLOADINGS= Data Set

The OUTLOADINGS= data set contains the eigenvalues of the correlation (or covariance) matrix, the loadings computed for the process variables, and other information about the principal component model. The variables that are saved in the OUTLOADINGS= data set are listed in Table 12.3.

Table 12.3 Variables in the OUTLOADINGS= Data Set

Variable	Description
VALUE	Character variable identifying the type of values in an observation
_PC_	Principal component number
NOBS	Number of observations used in the analysis
process variables	Eigenvalues, means, standard deviations, and loadings for *process variables*

Valid values for the _VALUE_ variable are as follows:

EIGEN	eigenvalues from the principal component analysis
LOADING	principal component loadings
MEAN	process variable means
STD	process variable standard deviations

For an observation where _VALUE_ is equal to LOADING, the _PC_ variable identifies the principal component whose loadings are recorded in that observation.

The process variable means and standard deviations are used by the other MVP procedures to center and scale new data in a Phase II analysis. If you specify the NOCENTER option, the OUTLOADINGS= data set does not contain a MEAN observation. If you specify the NOSCALE option, the OUTLOADINGS= data set does not contain a STD observation.

ODS Table Names

PROC MVPMODEL assigns a name to each table that it creates. You can use these names to refer to the tables when you use the Output Delivery System (ODS) to select tables and create output data sets. The ODS table names are listed in Table 12.4.

Table 12.4 ODS Tables Produced with the PROC MVPMODEL Statement

ODS Table Name	Description	Option
Corr	Correlation matrix	Default
Cov	Covariance matrix	COV or NOSCALE
CVResults	Results of cross validation	CV=
Eigenvalues	Eigenvalues of the correlation or covariance matrix	Default
ModelInfo	Model information	Default
ResidualSummary	Residual summary from cross validation	CV=

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The MVPMODEL procedure assigns a name to each graph that it creates using ODS Graphics. You can use these names to refer to the graphs when you use ODS. The ODS graph names are listed in Table 12.5.

Table 12.5 ODS Graphics Produced by PROC MVPMODEL

ODS Graph Name	Plot Description	Statement
CVPlot	Cross validation and R^2 analysis	CV=
LoadingMatrix	Scatter plot matrix of variable loadings	PLOTS=LOADINGS
LoadingPlot	Scatter plot of variable loadings	PLOTS=LOADINGS(UNPACK)
ScoreMatrix	Scatter plot of scores	PLOTS=SCORE
ScorePlot	Scatter plot of scores	PLOTS=SCORE(UNPACK)
ScreePlot	Scree and variance-explained plots	Default
VariancePlot	Variance-explained plot	PLOTS=SCREE(UNPACK)

Examples: MVPMODEL Procedure

Example 12.1: Using Cross Validation to Select the Number of Principal Components

This example uses cross validation to select the number of principal components in a model. It uses the chromatography data from McReynolds (1970), which is also used in Wold (1978) and Eastment and Krzanowski (1982). The following statements create the chromatography data set:

```
data mcreynolds;
   input x1 - x10;
   datalines;
 653    590    627    652    699    690    818    841    654    1006
 654    591    628    654    701    691    818    842    655    1006
 665    592    624    653    710    690    828    843    659    1014
 662    595    629    658    710    692    827    843    660    1012
 663    595    630    659    712    693    829    843    663    1013
 664    596    629    659    712    692    830    843    663    1015
 667    604    635    669    720    700    833    846    668    1016
 684    612    642    682    739    702    850    851    682    1035
 685    612    642    684    741    703    853    852    685    1039

   ... more lines ...

1247   1447   1386   1683   1616   1370   1327   1220   1508    1275
1300   1509   1424   1695   1675   1403   1362   1229   1571    1305
1343   1581   1480   1762   1699   1463   1375   1212   1618    1285
;
```

The observations are liquid phases, and the variables are compounds. The (i, j) value is the retention index for liquid phase i in compound j. The retention index values in the original article had the value of squalane subtracted from them. In this data set, the values have been corrected by adding the retention indices for squalane to all observations.

The following statements use the MVPMODEL procedure to select the number of principal components by using one-at-a-time cross validation:

```
proc mvpmodel data=mcreynolds plots=(scree cvplot) noscale cv=one;
run;
```

The CV= option specifies which method of cross validation to use to produce model diagnostics; in this case one-at-a-time cross validation is used. The PLOTS= option produces only the combination scree plot and variance-explained plot in addition to the cross validation plots.

Output 12.1.1 shows the model and data set information.

Output 12.1.1 Summary of Model and Data Set Information

The MVPMODEL Procedure

Data Set	WORK.MCREYNOLDS
Number of Variables	10
Missing Value Handling	Exclude
Number of Observations Read	226
Number of Observations Used	225
Maximum Number of Principal Components	9
Validation Method	Leave-one-out Cross Validation

Output 12.1.1 shows that one observation, liquid phase 69 (Triton X-400), was omitted because of a missing value. Also, notice that the maximum number of principal components is $\min(15, nvar, nobs) - 1 = 9$, which is less than the number of variables; this is described in detail in Eastment and Krzanowski (1982).

The root mean PRESS values and the *W* statistic are shown in Output 12.1.2.

Output 12.1.2 Residual Summary

Cross Validation for the Number of Components

Number of Components	Root Mean PRESS	W
0	974.3136	.
1	30.77631	9586.179
2	26.85973	2.707278
3	26.49878	0.211824
4	22.94873	2.261922
5	21.50501	0.810642
6	20.91568	0.279385
7	20.53967	0.14514
8	20.25766	0.082967
9	20.03932	0.04342

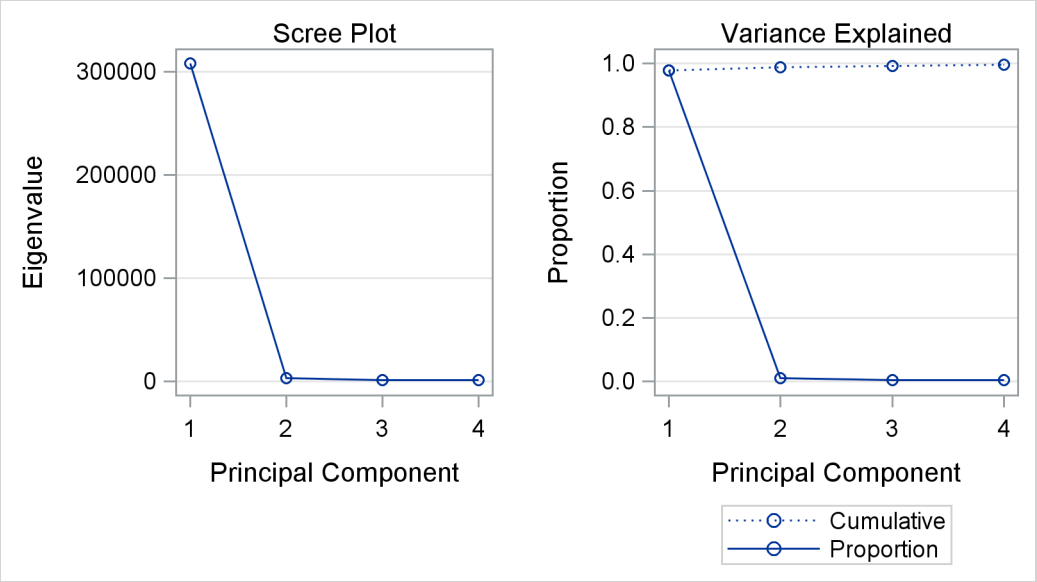
In this case the index of the last W statistics greater than one is $W[4]$, suggesting a model with four components as shown in Output 12.1.3.

Output 12.1.3 Cross Validation Results

Number of Components Suggested by W Statistic	4

You can also use scree and variance-explained plots to select the number of principal components, as shown in Output 12.1.4.

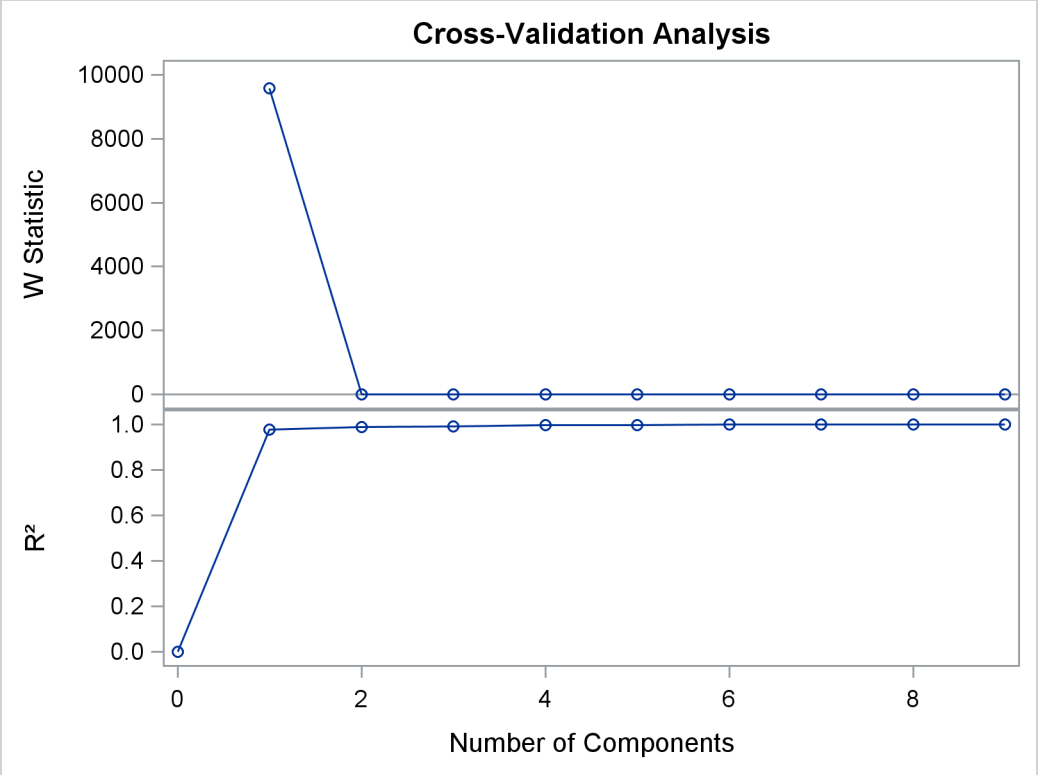
Output 12.1.4 Scree and Variance-Explained Plots



The plots in Output 12.1.4 indicate that one or two principal components explain almost all the variation.

The W statistic and R^2 plots are shown in Output 12.1.5.

Output 12.1.5 Cross Validation Analysis



The cross validation plot is produced only when you specify both the CV= option and PLOTS=ALL or PLOTS=CVPLOT.

It is interesting that the cross validation methods of Wold (1978) and Eastment and Krzanowski (1982) choose five and four components, respectively, for this model, whereas a visual examination of the knee in the scree plot might suggest using only one or two components.

Example 12.2: Computing the Classical T^2 Statistic

NOTE: The CV= option is experimental in this release.

This example uses the MVPMODEL procedure to produce a classical T^2 statistic and then compares it to the T^2 statistic produced by the principal component model with the NCOMP=ALL option. The two statistics are discussed in the section "Details: MVPMODEL Procedure" on page 930, and this example demonstrates that when the data set is centered and scaled correctly, the statistics are equal. The classical T^2 statistic is computed using the common quadratic form, which is implemented in SAS/IML. This example highlights the standardization that occurs by default in the MVPMODEL procedure. The example uses more of the airline delay data set that is first described in the section "Getting Started: MVPMODEL Procedure" on page 915. This data set covers the New England region of the continental United States. As before, the variables are airlines and the observations are mean daily delays during February 2007. The following statements create a SAS data set that contains these airline flight delays:

```
data flightDelaysNE;
   input  AA CO DL F9 FL NW UA US WN;
   datalines;
15.7  7.1  8.6  6.3 14.6  6.2  7.0 11.0  6.4
16.0 19.4 10.7  6.4 19.0  6.1  8.3 14.4 14.2
14.5  1.5  5.4 13.3 13.6  9.7 16.6  7.5  9.9
12.4 14.3  5.8  0.7 11.8 20.1 11.2  8.6  8.1
19.8 27.6  7.3 16.1 13.3 14.8 39.9 16.4  9.7
20.5 12.2  0.2 -4.8  3.7 14.2 41.7  4.9  9.2
 8.3  4.1  3.4  4.2 -2.3  6.3 24.9  8.7  4.4
 4.7 14.1  1.8 18.1 -1.9 10.2  5.4  5.8  3.7
16.7 15.0  3.5 11.8  0.8  7.3 11.1  7.2  5.1

   ... more lines ...

21.2 10.2  5.6  1.1 18.7  9.2 35.0 49.7 35.9
22.5 30.0 26.1 14.2 41.5 46.2 43.6 75.5 34.1
62.7 60.4 39.5 27.6 44.9 27.9 51.5 64.7 38.2
31.3 41.4 23.1 40.2 19.3 19.7 28.3 40.4 17.3
```

The following statements use the MVPMODEL procedure to create classical T^2 statistics:

```
proc mvpmodel data=flightDelaysNE ncomp=all plots=none out=mvpout;
   var AA CO DL F9 FL NW UA US WN;
run;
```

Specifying NCOMP=ALL sets the number of principal components to be used in the model equal to the number of process variables. Therefore, as discussed in the section "Details: MVPMODEL Procedure" on page 930, the mvpout data set contains the classical T^2 statistic for each observation, $T_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$.

The following SAS/IML statements generate the Hotelling T^2 statistic for the data set by using the traditional quadratic form. However, the data must first be standardized as done by the MVPMODEL procedure.

NOTE: If you do not want PROC MVPMODEL to center or scale the data, specify the NOCENTER or NOSCALE option, respectively.

```
proc iml;
   use flightDelaysNE;
   read all into x;
   n = nrow(x);
   p = ncol(x);
   xc = x-x[:,]; /* Create a centered data set*/
   ss = xc[##,];  /* Compute sum of squares */
   std=sqrt(ss/(n-1));  /* Compute standard deviations */
   std_x = xc/std;  /* Create a standardized data set*/
   S= cov(std_x);  /* Compute covariance of standardized data*/
   tsq = J(n,1,.);
   do i = 1 to n;
   /* Compute the classical T2 statistic using quadratic form */
      tsq[i] = std_x[i,]*inv(S)*std_x[i,]`;
   end;
   varnames = "tsq";
   create classicTsq from tsq [colname = varnames];
   append from tsq;
quit;
```

To compare the output from the MVPMODEL procedure with the output from SAS/IML, a new data set, mvpTsq, which contains the T^2 statistics computed by using the quadratic form in SAS/IML, is created:

```
data mvpTsq;
   set mvpOut(rename=(_TSQUARE_=tsq));
   keep tsq;
run;
```

Finally, you can verify that the two statistics are equivalent within machine precision by using the COMPARE procedure:

```
proc compare base=classicTsq compare=mvpTsq
             method=relative briefsummary;
run;
```

Output 12.2.1 Comparison of T^2 Statistics

```
                    The COMPARE Procedure
          Comparison of WORK.CLASSICTSQ with WORK.MVPTSQ
               (Method=RELATIVE, Criterion=0.00001)

NOTE: All values compared are within the equality criterion used. However, 16
      of the values compared are not exactly equal.
```

References

Alt, F. (1985), "Multivariate Quality Control," in S. Kotz, N. L. Johnson, and C. B. Read, eds., *Encyclopedia of Statistical Sciences*, volume 6, 110–122, New York: John Wiley & Sons.

Cooley, W. W. and Lohnes, P. R. (1971), *Multivariate Data Analysis*, New York: John Wiley & Sons.

Eastment, H. T. and Krzanowski, W. J. (1982), "Cross-Validatory Choice of the Number of Components from a Principal Component Analysis," *Technometrics*, 24, 73–77.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley & Sons.

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–441, 498–520.

Jackson, J. E. (1991), *A User's Guide to Principal Components*, New York: John Wiley & Sons.

Kourti, T. and MacGregor, J. F. (1995), "Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods," *Chemometrics and Intelligent Laboratory Systems*, 28, 3–21.

Kourti, T. and MacGregor, J. F. (1996), "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, 28, 409–428.

Kshirsagar, A. M. (1972), *Multivariate Analysis*, New York: Marcel Dekker.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

McReynolds, W. O. (1970), "Characterization of Some Liquid Phases," *Journal of Chromatographic Science*, 8, 685–691.

Miller, P., Swanson, R. E., and Heckler, C. H. E. (1998), "Contribution Plots: A Missing Link in Multivariate Quality Control," *Applied Mathematics and Computer Science*, 8, 775–792.

Morrison, D. F. (1976), *Multivariate Statistical Methods*, 2nd Edition, New York: McGraw-Hill.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 6, 559–572.

Rao, C. R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhyā, Series A*, 26, 329–358.

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, 20, 397–405.

Subject Index

Syntax Index