

**SAS/QC<sup>®</sup> 14.3**  
**User's Guide**  
**The MVPDIAGNOSE**  
**Procedure**

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

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

### **SAS/QC® 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 12

## The MVPDIAGNOSE Procedure

### Contents

---

Overview: MVPDIAGNOSE Procedure . . . . .	<b>899</b>
Getting Started: MVPDIAGNOSE Procedure . . . . .	<b>900</b>
Syntax: MVPDIAGNOSE Procedure . . . . .	<b>904</b>
PROC MVPDIAGNOSE Statement . . . . .	904
BY Statement . . . . .	905
CONTRIBUTIONPANEL Statement . . . . .	906
CONTRIBUTIONPLOT Statement . . . . .	907
ID Statement . . . . .	908
SCOREMATRIX Statement . . . . .	908
SCOREPLOT Statement . . . . .	909
TIME Statement . . . . .	910
Common Plot Statement Options . . . . .	911
Details: MVPDIAGNOSE Procedure . . . . .	<b>912</b>
Contribution Plots . . . . .	912
Paneled Contribution Plot Layouts . . . . .	912
Input Data Sets . . . . .	913
ODS Graphics . . . . .	915
Examples: MVPDIAGNOSE Procedure . . . . .	<b>915</b>
Example 12.1: Phase II Analysis with MVPDIAGNOSE . . . . .	915
References . . . . .	<b>920</b>

---

---

## Overview: MVPDIAGNOSE Procedure

The MVPDIAGNOSE procedure is used in conjunction with the `MVPMODEL` and `MVPMONITOR` 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 11, “Introduction to Multivariate Process Monitoring Procedures,” for a description of how the MVP procedures work together, and Chapter 13, “The MVPMODEL Procedure,” and Chapter 14, “The MVPMONITOR Procedure,” for details about the other MVP procedures.

The MVPDIAGNOSE procedure produces the following graphs that can provide insight into the variation in a process:

- score plots for pairs of principal components

- score plot matrices containing pairwise plots for multiple pairs of principal components
- contribution plots for individual observations
- paneled contribution plots for multiple observations

Each point in a score plot corresponds to a single observation from the input data set. A contribution plot displays the process variable contributions to a squared prediction error (SPE) or  $T^2$  statistic from a single observation in the input data set. Therefore, each observation in the input data is independent in how PROC MVPDIAGNOSE handles it. This enables you to preprocess the input data flexibly by using the DATA step, WHERE expressions, and other SAS language elements to select the data to plot.

**NOTE:** ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement before invoking the procedure) in order for the MVPDIAGNOSE procedure to produce graphical output.

---

## Getting Started: MVPDIAGNOSE Procedure

This example illustrates the basic features of the MVPDIAGNOSE procedure by using airline flight delay data available from the U.S. Bureau of Transportation Statistics at <http://www.transtats.bts.gov>. Suppose you want to compare process variable contributions for an out-of-control  $T^2$  statistic with contributions for adjacent observations. This kind of comparison can help you understand the underlying causes of unusual variation in the process.

The following statements create a SAS data set named MWflightDelays that provides the delays for flights that originated in the midwestern United States. 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
02/07/07 14.2 3.0 2.1 -0.9 -0.6 7.8 19.9 8.6 6.4
02/08/07 6.5 6.8 1.8 7.7 1.3 6.9 6.1 9.2 5.4
02/09/07 12.8 9.4 5.5 9.3 -0.2 4.6 7.6 7.8 7.5
02/10/07 9.4 3.5 1.5 -0.2 2.2 9.9 3.1 12.5 3.0
02/11/07 12.9 5.4 0.9 6.8 2.1 7.9 3.7 10.7 5.6
02/12/07 34.6 15.9 1.8 1.0 4.5 10.2 14.0 19.1 4.9
02/13/07 34.0 16.0 4.4 6.1 18.3 9.1 30.2 46.3 50.6
02/14/07 21.2 45.9 16.6 12.5 35.1 23.8 40.4 43.6 35.2
02/15/07 46.6 36.3 23.9 20.8 30.4 24.3 30.3 59.9 25.6
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 delays in minutes for flights that depart from the Midwest. For example, on February 2, 2007, F9 (Frontier Airlines) flights departed an average of 6.2 minutes late.

The first step in multivariate process monitoring of the data is to build a principal component model of the process variation. The following statements use `PROC MVPMODEL` to create a model with three principal components. (See Chapter 13, “The `MVPMODEL` Procedure,” for details.)

```
proc mvpmmodel data=MWflightDelays ncomp=3 noprint
    out=mvpair outloadings=mvpairloadings;
    var AA CO DL F9 FL NW UA US WN;
run;
```

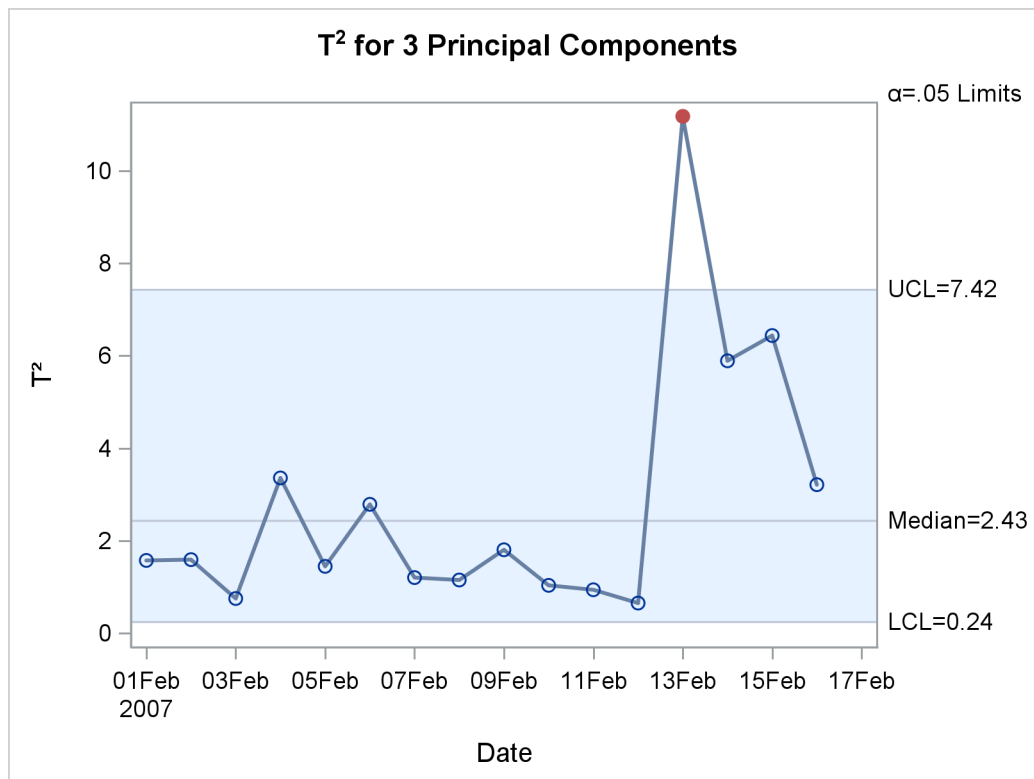
The `mvpair` data set contains the process data and associated principal component scores. The `mvpairloadings` data set contains the principal component loadings for the process variables and other data that describe the model.

The following statements create a  $T^2$  control chart by using the principal components. (See Chapter 14, “The `MVPMONITOR` Procedure,” for details.)

```
ods graphics on;
proc mvpmmonitor history=mvpair loadings=mvpairloadings;
    time flightDate;
    tsquarechart / contributions;
run;
```

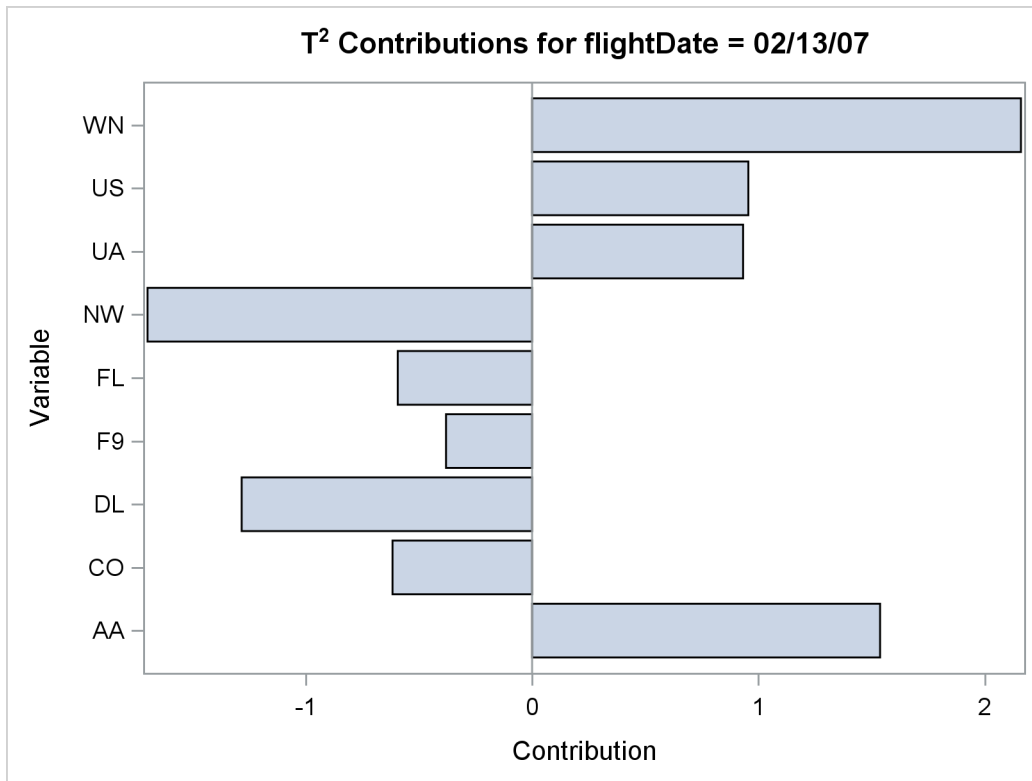
The `CONTRIBUTIONS` option produces contribution plots for any out-of-control points in the  $T^2$  chart. Figure 12.1 shows the  $T^2$  chart.

**Figure 12.1** Multivariate Control Chart for  $T^2$  Statistics



The  $T^2$  chart shows an out-of-control point on February 13, 2007. Figure 12.2 shows the contribution plot for this date that was produced by the CONTRIBUTIONS option.

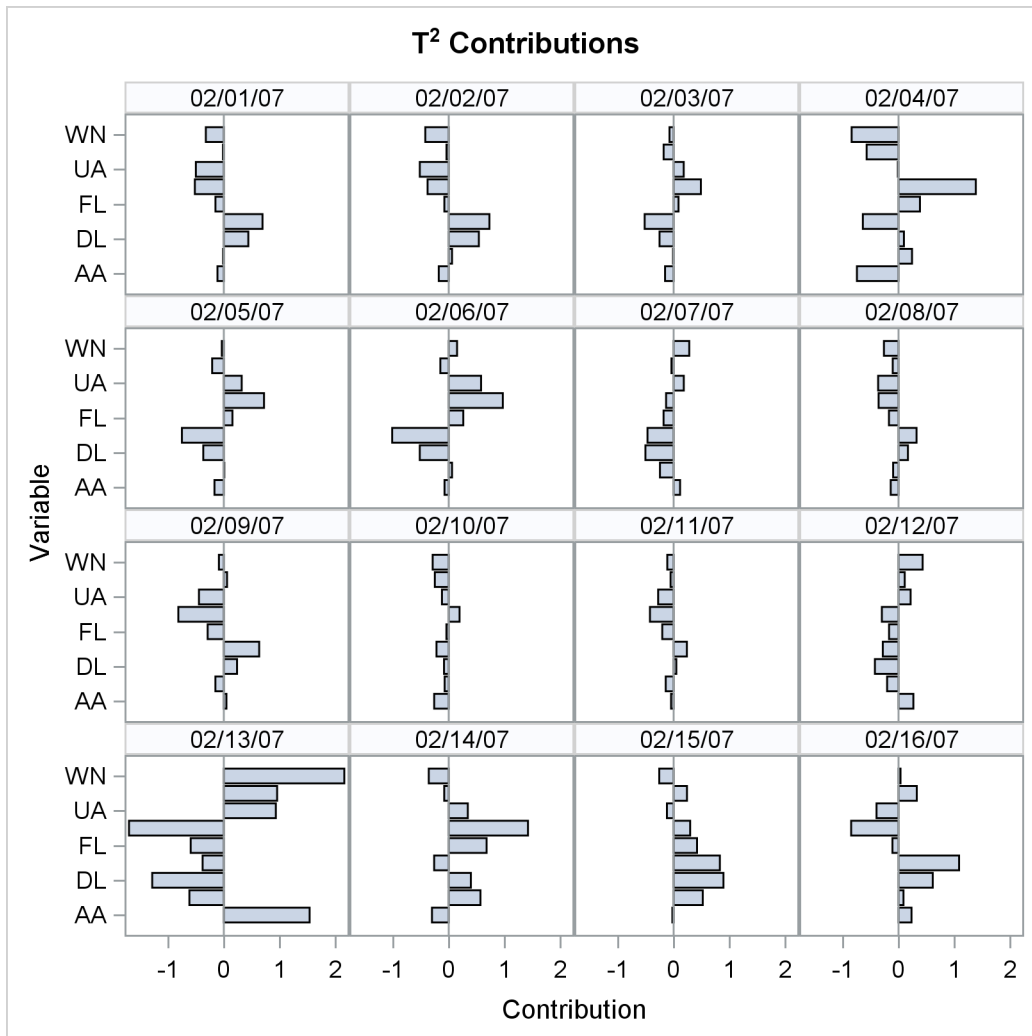
**Figure 12.2** Contribution Plot for Out-of-Control Point



The contribution plot shows that the delays for airlines AA, DL, NW, and WN are the major contributors to the out-of-control point. You can use PROC MVPDIAGNOSE to compare the contributions for this point to those for adjacent points. The following statements produce paneled contribution plots of all the observations in mvpair:

```
proc mvpdiagnose history=mvpair loadings=mvpairloadings;
  time flightDate;
  contributionpanel / type=tsquare;
run;
```

Figure 12.3 shows the paneled contribution plots.

Figure 12.3  $T^2$  Contributions for Flight Delays

The contribution plot for February 13 is in the lower-left corner of the plot. Notice that the magnitudes of all the process variable contributions are quite large for this date compared to those for the other dates. All the process variables contributed strongly to the out-of-control  $T^2$  statistic. This implies that something unusual occurred on February 13 that affected the flight delays for all the airlines.

In fact, on this day a strong winter storm battered the Midwest. This is an example of variation due to a *special cause*. Special causes, also referred to as *assignable causes*, are local, sporadic, or transient problems in a process. They are distinguished from *common causes* of variation, which are inherent in a system. Control charts are used to monitor the process for the occurrence of special causes and to measure and potentially to reduce the effects of common causes.

---

## Syntax: MVPDIAGNOSE Procedure

```

PROC MVPDIAGNOSE < options > ;
  BY variables ;
  CONTRIBUTIONPANEL < / options > ;
  CONTRIBUTIONPLOT < / options > ;
  ID variables ;
  SCOREMATRIX < / options > ;
  SCOREPLOT < / options > ;
  TIME variable ;

```

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

---

### PROC MVPDIAGNOSE Statement

```
PROC MVPDIAGNOSE < options > ;
```

The PROC MVPDIAGNOSE statement invokes the MVPDIAGNOSE procedure and specifies input data sets. You can specify the following *options*:

#### **DATA=SAS-data-set**

specifies an input SAS data set that contains process measurement data for a Phase II analysis. If you specify a DATA= data set, you must also specify a **LOADINGS=** data set. You cannot specify both the **HISTORY=** option and the DATA= option. See the section “**DATA= Data Set**” on page 913 for details about DATA= data sets.

#### **HISTORY=SAS-data-set**

specifies an input SAS data set that contains process variable data that are augmented with principal component scores, multivariate summary statistics, and other calculated values. Usually you create a HISTORY= data set by using the **OUT=** option in the PROC MVPMODEL statement or the **OUTHISTORY=** option in the PROC MVPMONITOR statement. You cannot specify both the DATA= option and the HISTORY= option. See the section “**HISTORY= Data Set**” on page 913 for details about HISTORY= data sets.

#### **LOADINGS=SAS-data-set**

specifies an input SAS data set that contains eigenvalues, principal component loadings, and process variable means and standard deviations that are used to compute principal component scores and multivariate summary statistics for a Phase II analysis. Usually you create a LOADINGS= data set by using the **OUTLOADINGS=** option in the PROC MVPMODEL statement. See the section “**LOADINGS= Data Set**” on page 914 for details about LOADINGS= data sets.

#### **MISSING=AVG | NONE**

specifies how observations that have missing process variable values in the DATA= data set are to be handled. The option MISSING=AVG specifies that missing values for a given variable be replaced by the average of the nonmissing values for that variable. The default is MISSING=NONE, which excludes from the analysis any observation that has missing values for any of the process variables.

**PREFIX=***name*

specifies the prefix that is used to identify variables that contain principal component scores in the **HISTORY=** data set. For example, if you specify **PREFIX=ABC**, PROC MVPDIAGNOSE attempts to read the score variables ABC1, ABC2, ABC3, and so on. The default prefix is Prin, which is also the default score variable prefix for data sets created by using the **OUT=** option in the PROC MVPMODEL statement. If you are using an **OUT=** data set from PROC MVPMODEL in the **HISTORY=** data set, the **PREFIX=** values must match. That is, the **PREFIX=** value that is specified in the PROC MVPDIAGNOSE statement must match the **PREFIX=** value in the data set that is specified in the **OUT=** option in the PROC MVPMODEL statement.

**NOTE:** The number of characters in the prefix plus the number of digits that are required to enumerate the principal components must not exceed the current name length defined by the **VALIDVARNAME=** system option.

**RPREFIX=***name*

specifies the prefix that is used to identify variables that contain residuals in the **HISTORY=** data set. A residual variable name is formed by appending a process variable name to the prefix. The default prefix is R\_, which is also the default residual variable prefix for data sets created by using the **OUT=** option in the PROC MVPMODEL statement. If you are using a data set produced with the **OUT=** option in the PROC MVPMODEL statement as a **HISTORY=** data set, the **RPREFIX=** value must match the **RPREFIX=** value specified when the **OUT=** data set was created by PROC MVPMODEL.

If the combined length of the residual prefix and a process variable name 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\_** (nine characters), the maximum variable name length is 32, and there is a process variable named PrimaryThermometerReading (25 characters), then two characters are dropped from the middle of the process variable name. The resulting residual variable name is Residual\_PrimaryThermometerReading.

---

## BY Statement

**BY** *variables* ;

You can specify a BY statement with PROC MVPDIAGNOSE 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 MVPDIAGNOSE 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 *SAS Visual Data Management and Utility Procedures Guide*.

## CONTRIBUTIONPANEL Statement

**CONTRIBUTIONPANEL** < / options > ;

The CONTRIBUTIONPANEL statement displays a paneled layout of the contribution plots for each observation in the input data set, up to a maximum specified by the **MAXNPLOTS=** option. Individual contribution plots are displayed in panels from left to right and top to bottom, and they are identified by **TIME** variable values or observation numbers. You can use the **CONTRIBUTIONPLOT** statement to display each contribution plot as a separate graph.

Table 12.1 summarizes the *options* available in the CONTRIBUTIONPANEL statement.

**Table 12.1** CONTRIBUTIONPANEL Statement Options

Option	Description
<b>MAXNPLOTS=</b>	Specifies the maximum number of contribution plots displayed
<b>MAXNVAR=</b>	Specifies the maximum number of process variable contributions displayed in each plot
<b>NCOLS=</b>	Specifies the number of columns in the panel layout
<b>NROWS=</b>	Specifies the number of rows in the panel layout
<b>ODSFOOTNOTE=</b>	Adds a footnote to the paneled contribution plots
<b>ODSFOOTNOTE2=</b>	Adds a secondary footnote to the paneled contribution plots
<b>ODSTITLE=</b>	Specifies a title for the paneled contribution plots
<b>ODSTITLE2=</b>	Specifies a secondary title for the paneled contribution plots
<b>TYPE=</b>	Specifies the type of contribution plots produced

You can specify the following *options* in the CONTRIBUTIONPANEL statement. The section “Common Plot Statement Options” on page 911 describes additional options that are available in all plot statements.

### **MAXNPLOTS=*n***

specifies the maximum number of contribution plots to be produced by the CONTRIBUTIONPANEL statement. The number of plots that are produced is the minimum of *n* and the number of observations in the input data set. When *n* is less than the number of observations, contribution plots are produced for the first *n* observations. The default is 50.

### **MAXNVAR=*n***

specifies the maximum number of process variable contributions to be displayed in the paneled layout. The magnitudes for each contribution are summed over the observations to be plotted, and the *n* contributions with the greatest total magnitudes are displayed. Therefore each plot displays contributions for the same process variables. By default, all contributions are displayed.

**NCOLS=*c***

specifies the number of columns in the panel layout. See the section “[Paneled Contribution Plot Layouts](#)” on page 912 for a description of how the default numbers of columns and rows are calculated.

**NROWS=*r***

specifies the number of rows in the panel layout. See the section “[Paneled Contribution Plot Layouts](#)” on page 912 for a description of how the default numbers of columns and rows are calculated.

**TYPE=SPE | TSQUARE**

specifies the type of contribution plots displayed. If you specify TYPE=SPE, the contribution plots are based on the SPE statistics; if you specify TYPE=TSQUARE, the contribution plots are based on the  $T^2$  statistics. You must specify a **LOADINGS=** data set to create  $T^2$  contribution plots. By default, TYPE=TSQUARE if a **LOADINGS=** data set is provided and TYPE=SPE otherwise.

---

## CONTRIBUTIONPLOT Statement

**CONTRIBUTIONPLOT** < / *options* > ;

The CONTRIBUTIONPLOT statement produces a contribution plot for each observation in the input data set, up to a maximum that is specified by the **MAXNPLOTS=** option. Each contribution plot is displayed as a separate graph. You can use the **CONTRIBUTIONPANEL** statement to display multiple contribution plots in a paneled layout.

Table 12.2 summarizes the *options* available in the CONTRIBUTIONPLOT statement.

**Table 12.2** CONTRIBUTIONPLOT Statement Options

Option	Description
<b>MAXNPLOTS=</b>	Specifies the maximum number of contribution plots displayed
<b>MAXNVAR=</b>	Specifies the maximum number of process variable contributions displayed in each plot
<b>ODSFOOTNOTE=</b>	Adds a footnote to the contribution plots
<b>ODSFOOTNOTE2=</b>	Adds a secondary footnote to the contribution plots
<b>ODSTITLE=</b>	Specifies a title for the contribution plots
<b>ODSTITLE2=</b>	Specifies a secondary title for the contribution plots
<b>TYPE=</b>	Specifies the type of contribution plots produced

You can specify the following *options* in the CONTRIBUTIONPLOT statement. The section “[Common Plot Statement Options](#)” on page 911 describes additional options that are available in all plot statements.

**MAXNPLOTS=*n***

specifies the maximum number of contribution plots to be produced by the CONTRIBUTIONPLOT statement. The number of plots that are produced is the minimum of *n* and the number of observations in the input data set. When *n* is less than the number of observations, contribution plots are produced for the first *n* observations. The default is 50.

**MAXNVAR=*n***

specifies that only the *n* contributions that have the greatest magnitudes be displayed in each plot. The contributions are ranked independently for each plot, so different process variable contributions might be displayed in different plots. By default, all contributions are displayed.

**TYPE=SPE | TSQUARE**

specifies the type of contribution plot to be created. The option TYPE=TSQUARE specifies that the contribution plots be based on the  $T^2$  statistics. The option TYPE=SPE specifies that the contribution plots be based on the SPE statistics. By default, TYPE=TSQUARE if a **LOADINGS=** data set is provided and TYPE=SPE otherwise. You can use more than one CONTRIBUTIONPLOT statement. You must specify a **LOADINGS=** data set to create  $T^2$  contribution plots.

**ID Statement**

**ID** *variables* ;

The first *ID variable* that is specified provides the labels for points in score plots. The values of all the *ID* variables are displayed in tooltips associated with points in a score plot 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.

**SCOREMATRIX Statement**

**SCOREMATRIX** < / *options* > ;

The **SCOREMATRIX** statement produces a matrix of score plots, each of which is a scatter plot of scores for a pair of principal components. You can use the **SCOREPLOT** statement to display a single score plot in a graph by itself.

Table 12.3 summarizes the *options* available in the **SCOREMATRIX** statement.

**Table 12.3** SCOREMATRIX Statement Options

Option	Description
ALPHA=	Specifies the $\alpha$ value for prediction ellipses
ELLIPSE	Requests prediction ellipses to be overlaid on score plots
GROUP=	Specifies a variable for grouping points in score plots
LABELS=	Specifies whether points in the score plots are labeled
NCOMP=	Specifies the number of principal components whose scores are plotted
ODSFOOTNOTE=	Adds a footnote to the score matrix
ODSFOOTNOTE2=	Adds a secondary footnote to the score matrix
ODSTITLE=	Specifies a title for the score matrix
ODSTITLE2=	Specifies a secondary title for the score matrix

You can specify the following *options* in the **SCOREMATRIX** statement. The section “Common Plot

Statement Options” on page 911 describes additional options that are available in all plot statements.

**ALPHA= $\alpha$** 

specifies the  $\alpha$  value for prediction ellipses that are overlaid on the score plots. The probability that a new observation falls outside the ellipse is  $\alpha$ . The default is 0.05. If you specify the ALPHA= option, you do not need to specify the ELLIPSE option.

**ELLIPSE**

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

**GROUP=*variable***

specifies a *variable* in the input data set that is used to group the points in the score plots. Points that have different GROUP= values are plotted using different markers or colors (or both) to distinguish the groups.

**LABELS=ON | OFF**

specifies whether points in the score plots are labeled. 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. The default is LABELS=OFF.

**NCOMP=*n***

specifies the number of principal components whose scores are plotted in the matrix. The principal components that are plotted are always 1 through *n*. By default, the matrix contains score plots for all principal components.

---

## SCOREPLOT Statement

**SCOREPLOT** < / *options* > ;

The SCOREPLOT statement produces a single score plot, which is a scatter plot of the scores that are associated with a pair of principal components. You can use the SCOREMATRIX statement to display a matrix of score plots for more than two principal components.

Table 12.4 summarizes the *options* available in the SCOREPLOT statement.

**Table 12.4** SCOREPLOT Statement Options

Option	Description
ALPHA=	Specifies the $\alpha$ value for the prediction ellipse
ELLIPSE	Requests a prediction ellipse to be overlaid on the score plot
GROUP=	Specifies a variable for grouping points in the score plot
LABELS=	Specifies which points in the score plot are to be labeled
ODSFOOTNOTE=	Adds a footnote to the score matrix
ODSFOOTNOTE2=	Adds a secondary footnote to the score matrix
ODSTITLE=	Specifies a title for the score matrix
ODSTITLE2=	Specifies a secondary title for the score matrix
XCOMP=	Specifies the principal component whose scores are plotted on the horizontal axis

Table 12.4 (continued)

Option	Description
YCOMP=	Specifies the number of principal components whose scores are plotted on the vertical axis

You can specify the following *options* in the SCOREPLOT statement. The section “Common Plot Statement Options” on page 911 describes additional options that are available in all plot statements.

**ALPHA= $\alpha$** 

specifies the  $\alpha$  value for a prediction ellipse that is overlaid on the score plot. The probability that a new observation falls outside the ellipse is  $\alpha$ . 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 plot. The probability that a new observation falls outside the ellipse is specified by the ALPHA= option.

**GROUP=*variable***

specifies a *variable* in the input data set that is used to group the points in the score plot. Points that have different GROUP= values are plotted with different markers or colors (or both) to distinguish the groups.

**LABELS=ON | OFF | OUTSIDE**

specifies which points in the score plot to label. 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. By default, LABELS=ON and all points are labeled. You can specify LABELS=OFF to suppress all point labels.

If you overlay a prediction ellipse on the score plot by specifying the ELLIPSE or ALPHA= option, you can specify LABELS=OUTSIDE to label only the points outside the prediction ellipse.

**XCOMP= $x$** 

specifies an integer  $x$  that identifies the principal component whose scores are plotted on the horizontal axis of the score plot. The default is 1. You cannot specify the same principal component number in both the XCOMP= and YCOMP= options.

**YCOMP= $y$** 

specifies an integer  $y$  that identifies the principal component whose scores are plotted on the vertical axis of the score plot. The default is  $\text{mod}(x, j) + 1$ , where  $x$  is the value that is specified by the XCOMP= option and  $j$  is the number of principal components in the model. You cannot specify the same principal component number in both the XCOMP= and YCOMP= options.

---

## TIME Statement

**TIME** *variable* ;

The TIME *variable* is a numeric variable that provides the chronological order or time values for measurements in the input data set. The *variable* name and value are incorporated into contribution plot titles to identify the

observations represented. If you do not specify a `TIME` variable, the observation number from the input data set is used instead.

---

## Common Plot Statement Options

You can specify the following *options* after a slash (/) in the `CONTRIBUTIONPANEL`, `CONTRIBUTIONPLOT`, `SCOREMATRIX`, and `SCOREPLOT` statements.

### **ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1** | *'string'*

adds a footnote to the plot. If you specify the `FOOTNOTE` (or `FOOTNOTE1`) keyword, the value of the SAS `FOOTNOTE` statement is used as the plot footnote. If you specify a quoted string, that string is used as the footnote. The quoted string can contain the following escaped characters, which are replaced by the values indicated:

`\n` is replaced by the `TIME` variable name.  
`\l` is replaced by the `TIME` variable label (or name if the analysis `TIME` has no label).

### **ODSFOOTNOTE2=FOOTNOTE2** | *'string'*

adds a secondary footnote to the plot. If you specify the `FOOTNOTE2` keyword, the value of the SAS `FOOTNOTE2` statement is used as the secondary plot footnote. If you specify a quoted string, that string is used as the secondary footnote. The quoted string can contain the following escaped characters, which are replaced by the values indicated:

`\n` is replaced by the `TIME` variable name.  
`\l` is replaced by the `TIME` variable label (or name if the `TIME` variable has no label).

### **ODSTITLE=TITLE | TITLE1 | NONE | DEFAULT** | *'string'*

specifies a title for the plot. You can specify the following values:

`TITLE` (or `TITLE1`) uses the value of the SAS `TITLE` statement as the plot title.

`NONE` suppresses all titles from the plot.

`DEFAULT` uses the default title.

If you specify a quoted string, that string is used as the graph title. The quoted string can contain the following escaped characters, which are replaced by the values indicated:

`\n` is replaced by the `TIME` variable name.  
`\l` is replaced by the `TIME` variable label (or name if the analysis variable has no label).

### **ODSTITLE2=TITLE2** | *'string'*

specifies a secondary title for the plot. If you specify the `TITLE2` keyword, the value of the SAS `TITLE2` statement is used as the secondary plot title. If you specify a quoted string, that string is used as the secondary title. The quoted string can contain the following escaped characters, which are replaced by the values indicated:

\n	is replaced by the <b>TIME</b> variable name.
\l	is replaced by the <b>TIME</b> variable label (or name if the analysis variable has no label).

## Details: MVPDIAGNOSE Procedure

### Contribution Plots

One way to diagnose the behavior of out-of-control points in multivariate control charts is to use contribution plots (Miller, Swanson, and Heckler 1998). These plots tell you which variables contribute to the distance between the points in an SPE or  $T^2$  chart and the sample mean of the data.

A contribution plot is a bar chart of the contributions of the process variables to the statistic. For the  $i$ th SPE statistic, the contribution of the  $k$ th variable is the  $k$ th entry of the vector  $\mathbf{e}_i$ , which is computed as

$$\mathbf{e}_i = \mathbf{x}_i (\mathbf{I} - \mathbf{P}_j \mathbf{P}_j')$$

where  $\mathbf{e}_i$  is the vector of errors from the principal component model for observation  $i$  and  $\mathbf{x}_i$  is the  $i$ th observation. The contributions to the  $i$ th  $T^2$  statistic are computed in the same way as the entries of the vector

$$\mathbf{T}_i^2 = \mathbf{x}_i \mathbf{P}_j \mathbf{L}^{-1} \mathbf{P}_j'$$

where  $\mathbf{P}_j$  is the matrix of the first  $j$  eigenvectors and  $\mathbf{L}$  is the diagonal matrix of the first  $j$  eigenvalues.

### Paneled Contribution Plot Layouts

The **CONTRIBUTIONPANEL** statement produces paneled contribution plots. You can use the options **NCOLS= $c$**  and **NROWS= $r$**  to specify the number of columns and rows in the layout, respectively.

By default,  $c$  and  $r$  are determined by  $p$ , the number of contribution plots to be displayed. If  $p \leq 16$ , then  $c = \lceil \sqrt{p} \rceil$  and  $r = \lceil p/c \rceil$ . Otherwise, one of the following three layouts is used to minimize the number of empty panels in the last page of the graph:

- $c = 4, r = 4$
- $c = 4, r = 3$
- $c = 3, r = 3$

If you specify only **NCOLS= $c$** , then  $r = \lceil p/c \rceil$ . If you specify only **NROWS= $r$** , then  $c = \lceil p/r \rceil$ .

Although  $c \leq 4$  and  $r \leq 4$  by default, you can specify values greater than 4 in the **NCOLS=** and **NROWS=** options.

## Input Data Sets

The MVPDIAGNOSE procedure accepts a primary input data set that has one of the following two types:

- a **DATA=** data set that contains new process data to be analyzed by using an existing principal component model (Phase II analysis)
- a **HISTORY=** data set that contains process data and the accompanying scores, residuals, and statistics that are produced by using a principal component model. The process data can be the original data that were used to create the model (Phase I analysis) or subsequent data that were analyzed by using a previously created model (Phase II analysis)

These options are mutually exclusive. If you do not specify an option that identifies a primary input data set, PROC MVPDIAGNOSE uses the most recently created SAS data set as a **DATA=** data set.

When you specify a **DATA=** data set, you must also specify a **LOADINGS=** data set that contains principal component loadings and other information that describes the principal component model. When you specify a **HISTORY=** data set, you must also specify a **LOADINGS=** data set if you use a **CONTRIBUTIONPANEL** or **CONTRIBUTIONPLOT** statement and specify the **TYPE=TSQUARE** option.

### DATA= Data Set

A **DATA=** data set provides the process measurement data for a Phase II analysis. In addition to containing the process variables, a **DATA=** data set can contain the following:

- **BY** variables
- **ID** variables
- a **TIME** variable

When you specify a **DATA=** data set, you must also specify a **LOADINGS=** data set that contains the loadings for the principal component model that describes the variation of the process. These loadings are used to score the new data from the **DATA=** data set. The process variables in the **LOADINGS=** data set must have the same names as those in the **DATA=** data set.

### HISTORY= Data Set

A **HISTORY=** data set provides the input data set for a Phase I or Phase II analysis. In addition to containing the original process variables, a **HISTORY=** data set contains principal component scores, residuals, SPE and  $T^2$  statistics, and a count of the observations that are used to construct the principal component model. These variables are summarized in Table 12.5.

**Table 12.5** Variables in the **HISTORY=** Data Set

Variable	Description
Prin1–Prin <i>j</i>	Principal component scores
R_var1–R_var <i>p</i>	Residuals

**Table 12.5** (continued)

Variable	Description
<code>_NOBS_</code>	Number of observations used in the analysis
<code>_SPE_</code>	Squared prediction error (SPE) statistic
<code>_TSQUARE_</code>	$T^2$ statistic computed from principal component scores

The score variables names must consist of a common prefix followed by the numbers 1, 2, . . . ,  $j$ , where  $j$  is the number of principal components. By default, the common prefix is Prin. You can use the `PREFIX=` option to specify another prefix for score variables.

If the number of principal components is less than the total number of process variables, the `HISTORY=` data set should also contain residual variables. A residual variable name consists of a common prefix followed by the corresponding process variable name. The default residual variable prefix is `R_`. For example, if the process variables are A, B, and C, the default residual variable names are `R_A`, `R_B`, and `R_C`. You can use the `RPREFIX=` option to specify a different residual variable prefix.

**NOTE:** Usually you create a `HISTORY=` data set by specifying the `OUT=` option in the PROC MVPMODEL statement or the `OUTHISTORY=` option in the PROC MVPMONITOR statement. If the `PREFIX=` or `RPREFIX=` option is used when such an output data set is created, you must specify the same prefixes to identify the score and residual variables when you read it as a `HISTORY=` data set.

### LOADINGS= Data Set

A `LOADINGS=` data set contains the following information about the principal component model:

- eigenvalues of the correlation or covariance matrix used to construct the model
- principal component loadings
- process variable means used to center the variable values
- process variable standard deviations used to scale the variable values

You can produce a `LOADINGS=` data set by using the `OUTLOADINGS=` option in the PROC MVPMODEL statement. Table 12.6 lists the variables that are required in a `LOADINGS=` data set.

**Table 12.6** Variables in the `LOADINGS=` Data Set

Variable	Description
<code>_VALUE_</code>	The value contained in <i>process variables</i> for a given observation
<code>_NOBS_</code>	Number of observations used to build the principal component model
<code>_PC_</code>	The principal component number; 0 for the observation that contains eigenvalues
<i>process variables</i>	Values associated with the 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

The `LOADINGS=` data set contains one EIGEN observation and  $j$  LOADING observations, where  $j$  is the number of principal components in the model. The presence of a MEAN observation indicates that the process variables were centered when the principal component model was built, and the presence of a STD observation indicates that the process variables were scaled when the principal component model was built. The means and standard deviations are used to center and scale new data in a Phase II analysis.

---

## 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 MVPDIAGNOSE procedure assigns a name to each graph that it creates. You can use these names to refer to the graphs when you use ODS Graphics. The ODS graph names are listed in Table 12.7.

**Table 12.7** ODS Graphics Produced by PROC MVPDIAGNOSE

ODS Graph Name	Plot Description	Statement
ContributionPanel	Paneled contribution plots	CONTRIBUTIONPANEL
ContributionPlot	Contribution plots	CONTRIBUTIONPLOT
ScoreMatrix	Matrix of pairwise score plots	SCOREMATRIX
ScorePlot	Score plot	SCOREPLOT

---

## Examples: MVPDIAGNOSE Procedure

### Example 12.1: Phase II Analysis with MVPDIAGNOSE

The example in “Getting Started: MVPDIAGNOSE Procedure” on page 900 illustrates how you build a principal component model and apply the MVPMONITOR and MVPDIAGNOSE procedures to perform a Phase I analysis. In Phase I analysis you analyze the data that were used to build the principal component model. This example is a continuation of that example and illustrates how you can use PROC MVPDIAGNOSE to analyze process data that were not used to build the model. This is called a Phase II analysis. A Phase II analysis is usually performed on data that are collected after the data that are used to build the model.

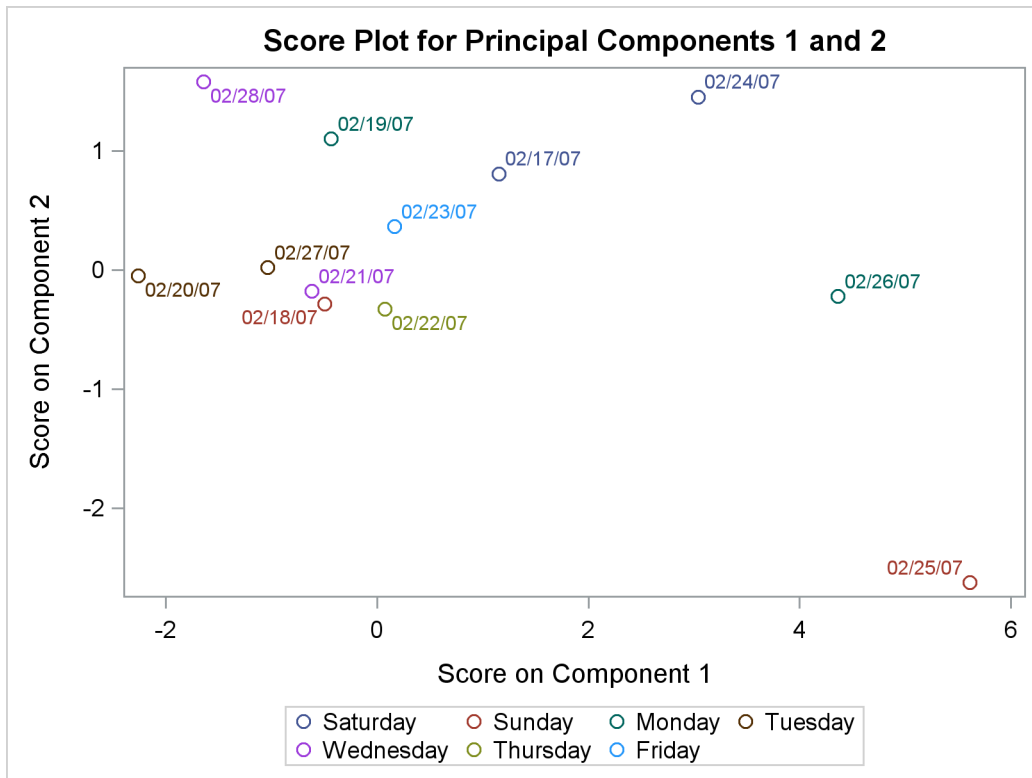
In the original example the principal component model was built using flight delay data from February 1–16, 2007. The following statements create a data set named `MWflightDelays2` that contains average delays for flights that originated in the midwestern United States on February 17–28, 2007:

```
data MWflightDelays2;
  label flightDate='Date';
  format flightDate MMDDYY8.;
  input flightDate :MMDDYY8. AA CO DL F9 FL NW UA US WN;
  dayofweek = put(flightDate,downname.);
  datalines;
02/17/07 25.6 7.8 15.5 13.4 16.1 16.2 23.0 24.2 8.2
02/18/07 5.4 16.0 9.9 1.1 11.5 17.0 15.6 15.5 5.1
02/19/07 13.2 16.3 10.0 10.6 5.4 10.3 9.5 16.8 9.3
02/20/07 4.2 6.9 1.4 0.1 7.2 6.6 7.4 10.4 2.9
02/21/07 5.4 -0.1 7.4 8.7 16.3 24.3 9.4 6.0 10.2
02/22/07 19.6 30.2 6.8 2.7 8.9 16.4 14.3 12.6 8.2
02/23/07 14.9 18.9 9.9 9.1 12.0 16.5 17.4 12.8 6.0
02/24/07 21.4 5.5 11.1 46.1 10.6 55.3 22.9 8.8 3.4
02/25/07 42.6 7.7 14.6 14.4 32.0 50.7 46.1 49.4 39.1
02/26/07 43.2 25.1 18.1 18.2 28.8 31.1 38.6 29.6 18.6
02/27/07 11.3 17.1 5.3 4.1 4.8 13.9 9.8 9.7 7.1
02/28/07 8.1 3.7 2.7 17.1 -0.8 5.5 11.0 14.3 3.1
;
```

The `dayofweek` variable contains the day of the week for each date in the input data set. The following statements apply the model that is saved in the `mvpairloadings` data set to the `flightDelays2` data and produce a score plot for the first two principal components:

```
proc mvpdiagnose data=MWflightDelays2 loadings=mvpairloadings;
  id flightDate;
  scoreplot / labels=on group=dayofweek;
  label dayofweek='';
run;
```

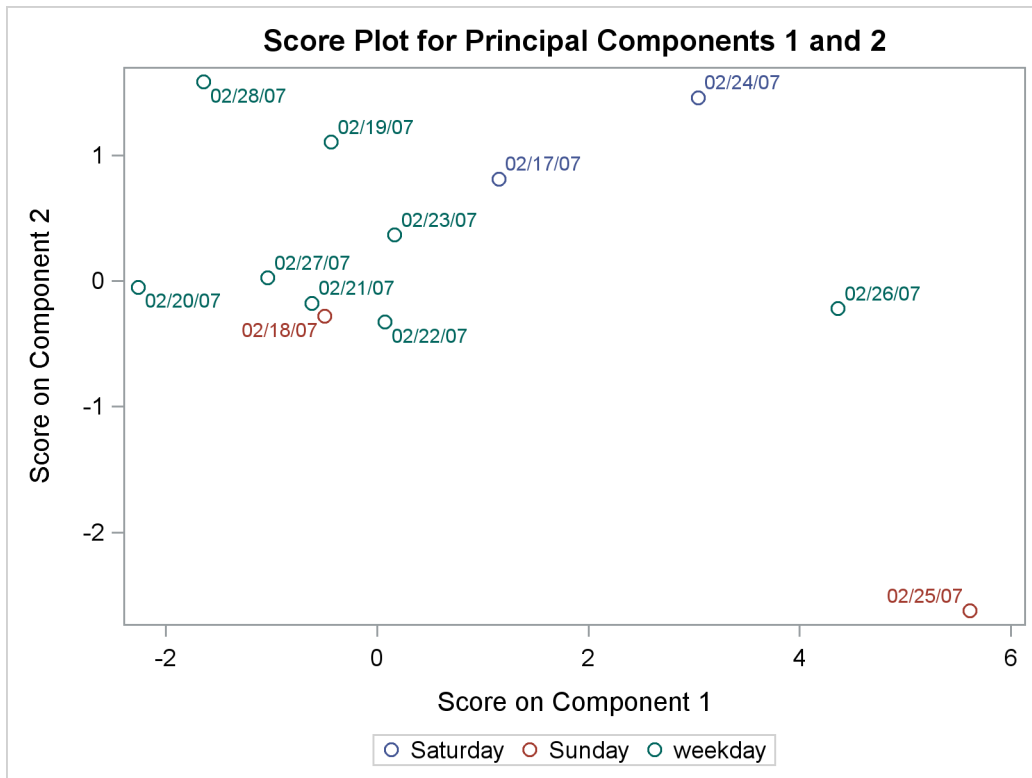
The `ID` statement labels the points in the score plot with `flightDate` variable values. The `GROUP=` option displays the observations grouped by day of the week. The `LABEL` statement suppresses the `GROUP=` legend label. Figure 12.1.1 shows the score plot.

**Output 12.1.1** Score Plot with Observations Grouped by Day of the Week

The Saturday and Sunday observations seem to be divided by scores for principal component 1. The following statements modify the dayofweek values to merge the observations for the other days into a “weekday” group:

```
data MWflightDelays2;
  set MWflightDelays2;
  weekday = put(flightDate,weekday.);
  if not ( weekday in (1 7) ) then dayofweek='weekday';
run;
```

Figure 12.1.2 shows the score plot that is produced by running PROC MVPDIAGNOSE with this new grouping. Merging the weekday observations into a single group emphasizes the Saturday and Sunday scores.

**Output 12.1.2** Score Plot with Alternate Grouping

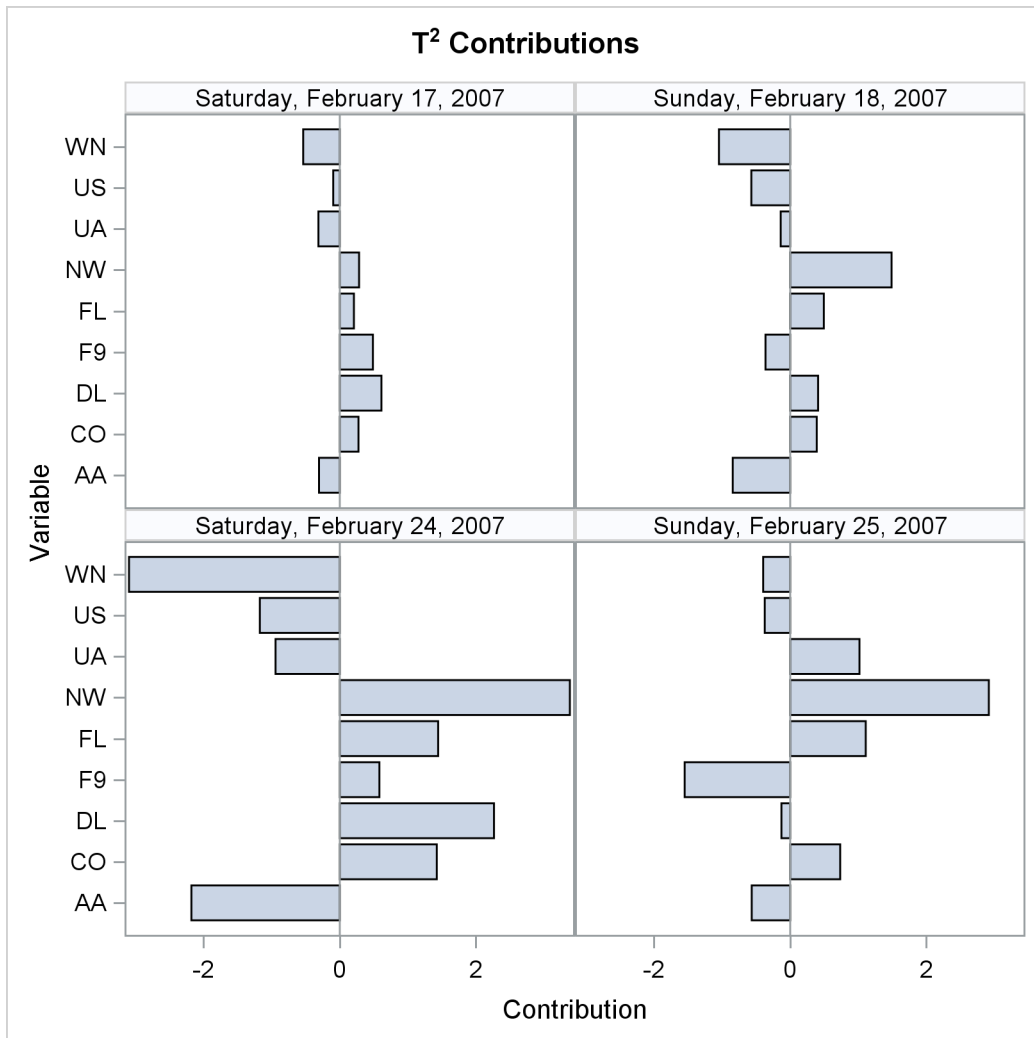
Because the score plot shows that something interesting might be happening on the weekends, you might want to examine contribution plots for those days. The following statements produce paneled  $T^2$  and SPE contribution plots for the weekend observations:

```
proc mvpdiagnose data=MWflightDelays2 loadings=mvpairloadings;
  where dayofweek ne 'weekday';
  time flightDate;
  contributionpanel;
  contributionpanel / type=spe;
  format flightDate weekdate.;
run;
```

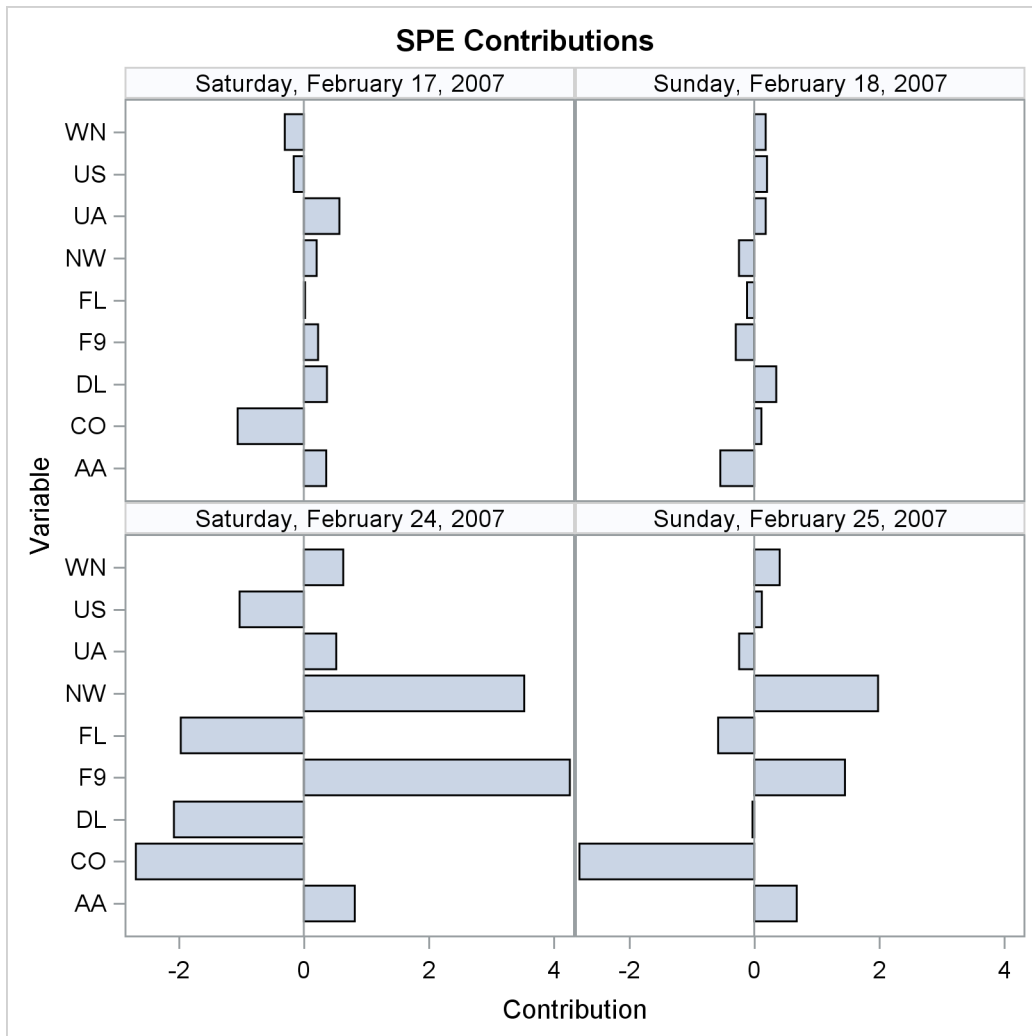
Because the **CONTRIBUTIONPANEL** statement produces contribution plots for a series of observations starting with the first observation in the input data set, it is convenient to use a **WHERE** statement to select observations of interest.

Figure 12.1.3 and Figure 12.1.4 show the paneled contribution plots.

**Output 12.1.3**  $T^2$  Contribution Plots for Weekends



**Output 12.1.4** SPE Contribution Plots for Weekends



## References

- Alt, F. (1985). "Multivariate Quality Control." In *Encyclopedia of Statistical Sciences*, vol. 6, edited by S. Kotz, N. L. Johnson, and C. B. Read. New York: John Wiley & Sons.
- Cooley, W. W., and Lohnes, P. R. (1971). *Multivariate Data Analysis*. New York: John Wiley & Sons.
- 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., and Mudholkar, G. S. (1979). "Control Procedures for Residuals Associated with Principal Component Analysis." *Technometrics* 21:341–349.

- Jensen, D. R., and Solomon, H. (1972). "A Gaussian Approximation to the Distribution of a Definite Quadratic Form." *Journal of the American Statistical Association* 67:898–902.
- 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.
- 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 ed. 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.
- Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons.

# Subject Index

contribution plots, 906, 907

extreme observations, 908

MVPDIAGNOSE procedure

    examples, 915

    extreme observations, 908

    missing values, 904

    ODS graph names, 915

score plots, 908, 909



# Syntax Index

- ALPHA= option
  - SCOREMATRIX statement, 909
  - SCOREPLOT statement, 910
- BY statement
  - MVPDIAGNOSE procedure, 905
- CONTRIBUTIONPANEL statement
  - MVPDIAGNOSE procedure, 906
- CONTRIBUTIONPLOT statement
  - MVPDIAGNOSE procedure, 907
- DATA= option
  - PROC MVPDIAGNOSE statement, 904, 913
- ELLIPSE option
  - SCOREMATRIX statement, 909
  - SCOREPLOT statement, 910
- GROUP= option
  - SCOREMATRIX statement, 909
  - SCOREPLOT statement, 910
- HISTORY= option
  - PROC MVPDIAGNOSE statement, 904, 913
- ID statement
  - MVPDIAGNOSE procedure, 908
- LABELS= option
  - SCOREMATRIX statement, 909
  - SCOREPLOT statement, 910
- LOADINGS= option
  - PROC MVPDIAGNOSE statement, 904, 914
- MAXNPLOTS= option
  - CONTRIBUTIONPANEL statement, 906
  - CONTRIBUTIONPLOT statement, 907
- MAXNVAR= option
  - CONTRIBUTIONPANEL statement, 906
  - CONTRIBUTIONPLOT statement, 908
- MISSING= option
  - PROC MVPDIAGNOSE statement, 904
- MVPDIAGNOSE procedure
  - syntax, 904
- MVPDIAGNOSE procedure, BY statement, 905
- MVPDIAGNOSE procedure,
  - CONTRIBUTIONPANEL statement, 906
  - MAXNPLOTS= option, 906
  - MAXNVAR= option, 906
  - NCOLS= option, 907
  - NROWS= option, 907
  - TYPE= option, 907
- MVPDIAGNOSE procedure, CONTRIBUTIONPLOT statement, 907
  - MAXNPLOTS= option, 907
  - MAXNVAR= option, 908
  - TYPE= option, 908
- MVPDIAGNOSE procedure, ID statement, 908
- MVPDIAGNOSE procedure, plot statement
  - ODSFOOTNOTE2= option, 911
  - ODSFOOTNOTE= option, 911
  - ODSTITLE2= option, 911
  - ODSTITLE= option, 911
- MVPDIAGNOSE procedure, plot statement options, 911
- MVPDIAGNOSE procedure, PROC MVPDIAGNOSE statement, 904
  - DATA= option, 904, 913
  - HISTORY= option, 904, 913
  - LOADINGS= option, 904, 914
  - MISSING= option, 904
  - PREFIX= option, 905
  - RPREFIX= option, 905
- MVPDIAGNOSE procedure, SCOREMATRIX statement, 908
  - ALPHA= option, 909
  - ELLIPSE option, 909
  - GROUP= option, 909
  - LABELS= option, 909
  - NCOMP= option, 909
- MVPDIAGNOSE procedure, SCOREPLOT statement, 909
  - ALPHA= option, 910
  - ELLIPSE option, 910
  - GROUP= option, 910
  - LABELS= option, 910
  - XCOMP= option, 910
  - YCOMP= option, 910
- MVPDIAGNOSE procedure, TIME statement, 910
- NCOLS= option
  - CONTRIBUTIONPANEL statement, 907
- NCOMP= option
  - SCOREMATRIX statement, 909
- NROWS= option
  - CONTRIBUTIONPANEL statement, 907
- ODSFOOTNOTE2= option

- plot statement, 911
- ODSFOOTNOTE= option
  - chart statement, 911
- ODSTITLE2= option
  - plot statement, 911
- ODSTITLE= option
  - plot statement, 911
- plot statement options
  - MVPDIAGNOSE procedure, 911
- PREFIX= option
  - PROC MVPDIAGNOSE statement, 905
- PROC MVPDIAGNOSE statement, 904, *see*
  - MVPDIAGNOSE procedure
- RPREFIX= option
  - PROC MVPDIAGNOSE statement, 905
- SCOREMATRIX statement
  - MVPDIAGNOSE procedure, 908
- SCOREPLOT statement
  - MVPDIAGNOSE procedure, 909
- TIME statement
  - MVPDIAGNOSE procedure, 910
- TYPE= option
  - CONTRIBUTIONPANEL statement, 907
  - CONTRIBUTIONPLOT statement, 908
- XCOMP= option
  - SCOREPLOT statement, 910
- YCOMP= option
  - SCOREPLOT statement, 910