Paper 252-26

Multivariate Methods for Process Knowledge Discovery: The Power to Know Your Process

Robert N. Rodriguez and Randall D. Tobias, SASΔnstitute Inc., Cary, NC

Abstract

Knowing the relationship between the input and output variables of a process is critical to process monitoring, prediction, and optimization. Powerful methods for multivariate process modeling, which address the large volumes of data available in modern manufacturing environments, have been developed by chemometricians and are being applied within the chemical process industries. This paper provides an introduction to these methods, which are based on projection to latent structures. It discusses the implementation of these methods with the SAS® System and illustrates them with financial indicators used to monitor firms in a regulatory system.

Introduction

In manufacturing the keys to successful process monitoring and optimization are identifying the critical input and output variables, and understanding their relationships. Because modern manufacturing operations have invested heavily in automated data acquisition and measurement technology, vast quantities of process measurements are potentially available for modeling these relationships. However, these data are seldom fully exploited. Typically, the variables for a particular process step are examined in isolation, without accounting for their impact on subsequent steps or the final outcome of the process.

There are two main barriers to analyzing data across multiple process steps and developing the knowledge needed to manage an entire process. First, because such data live in a diversity of transactional or legacy systems (such as LIMS, MRP, ERP, and SPC), they are unavailable in analysis-ready form. SAS Institute is addressing this problem with the development of analytical data warehouse solutions for quality and process knowledge management, as discussed in a recent white paper; refer to SAS Institute Inc. (1999).

This paper addresses the second barrier, which is the complexity of the data once it is ready for analysis. Successful process knowledge discovery and management requires the power to analyze hundreds or even thousands of variables that are correlated and reflect multiple sources of variation, including changes over time.

During the past 15 years, powerful methods for multivariate process modeling and monitoring have been developed by chemometricians, and these methods have been applied successfully within the chemical process industries and in the field of analytical chemistry. The applications range from modeling the quality of paper pulp from digester process variables to development of structure-activity relationships for predicting the biological activity of new drug compounds from their chemical properties; refer to Dayal et al. (1994) and Wold (1995), respectively.

The first goal of this paper is to expose SAS users to the basic concepts and techniques in this area by providing simple examples along with the relevant SAS code. The second goal is to show that these methods are applicable to business processes, as well as industrial processes, by illustrating how they can be used to monitor financial indicators of firms in a regulatory system.

The field of chemometrics has grown so rapidly that a comprehensive survey of the literature is outside the scope of this presentation. Two papers by Kourti and MacGregor (1995, 1996) are strongly recommended as a starting point for readers who wish to learn more about multivariate process monitoring. The book by Beebe, Pell, and Seasholtz (1998) provides a good introduction to chemometrics.

Principal Components: A Brief Review

The key idea behind the methods illustrated in this paper is the use of projection to examine and model high-dimensional data in a low-dimensional “latent variable” subspace that describes most of the variability in the data. This section discusses how to determine this subspace using principal components analysis (PCA), a well-known technique in psychometrics, econometrics, market research, and many other areas. Refer to Jackson (1991) for a comprehensive reference on PCA.

Principal components analysis is particularly appropriate for solving problems in analytical chemistry and chemical engineering, where the number of measured variables or sensors is often greater than the number of samples, and where it is believed that a relatively small number of independent, unobservable factors or events dictate the behavior of the system. PCA is also appropriate in this setting because it handles highly correlated variables, each of which contributes a small amount of information about the latent factors.

To understand the concept of a principal component, consider Figure 1, which represents a cloud of points in a high-dimensional space. Most of the variation lies along a line, which is not parallel to any of the variable axes. This line, referred to as the *first principal component*, passes through the average of the points, and it is chosen so that the projections of the points onto the line minimize their distances to the data in a least squares sense.

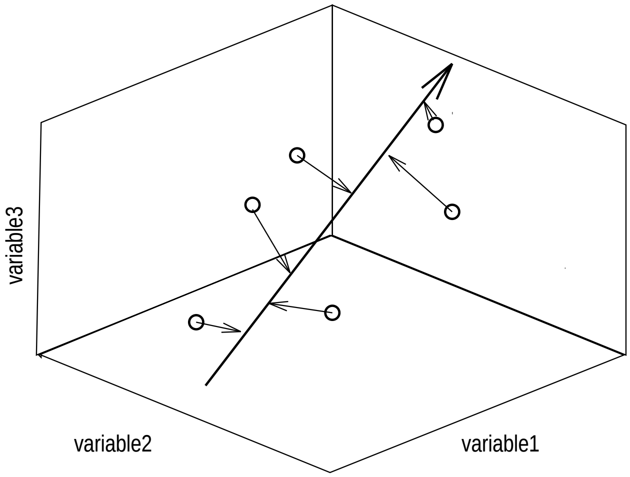

Figure 1. Data Cloud Projected to a Line

The *second principal component* is the line that passes through the average and minimizes the projection distances in a direction that is orthogonal to the first principal component. The first and second principal components define a plane, as illustrated in Figure 2.

In matrix notation, PCA is described as follows: Denote the *i*th measurement on the *j*th variable as X_{ij} for $i = 1, 2, \ldots, n$, where n is the number of measurements, and $j = 1, 2, \ldots, k$, where k is the number of variables. Then the *i*th sample can be represented as a vector $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{ik}]$, and the average of

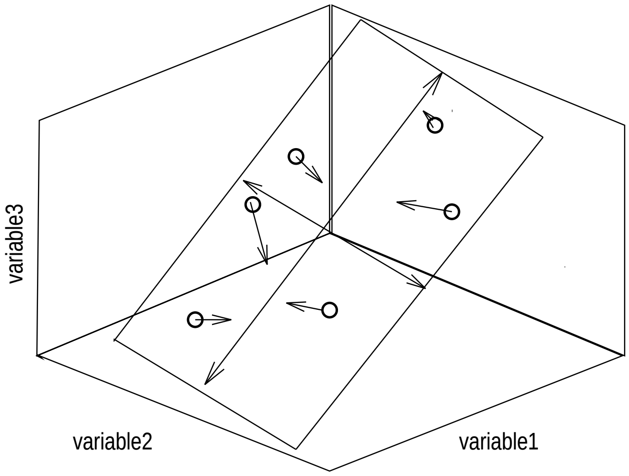

Figure 2. Data Cloud Projected to a Plane

the sample vectors is $\bar{\mathbf{X}}_n = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k]$, where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$. PCA decomposes the data matrix $\mathbf{X} = [\ X_{ij}\ ]$ as

$$\mathbf{X}_{n\times k} = \mathbf{1}_{n\times 1}\ \bar{\mathbf{X}}_{1\times k} + \mathbf{T}_{n\times p}\ \mathbf{P'}_{p\times k} + \mathbf{E}_{n\times k}$$

Here p is the dimension of the projection space ($p < k$), $\mathbf{T}$ is referred to as the *score matrix*, $\mathbf{P}$ is referred to as the *loading matrix*, and $\mathbf{E}$ is the *residual matrix*.

An important practical consideration in PCA is how the measurements are scaled prior to the decomposition. You should scale the data so that the variances of the variables reflect their importance. You can use uniform scaling (sometimes called autoscaling) when the variables are believed to be equally important. A second consideration is the choice of p, which you can make by examining the cumulative percent of variation explained by the decomposition or with statistical methods such as cross-validation. Refer to Jackson (1991) for further discussion.

The principal components are the column vectors of $\mathbf{T}$, and they are p new variables, each of which is a linear combination of the original k variables. These vectors, $\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_p}$, are customarily listed in decreasing order of importance with respect to their ability to describe variation in the data. Figure 1 illustrates the first principal component, $\mathbf{t_1} = \mathbf{X}\mathbf{p_1}$. The elements of $\mathbf{T}$ provide the coordinates of the projected observations along the p principal component axes and are referred to as scores. The k rows of $\mathbf{P}$ provide the loadings or weights of the original variables in the principal components.

A simple example drawn from geochemistry is helpful for understanding the roles of these statistics. One of the most extensive volcanic fields in the United States is the Zuni-Bandera field, which can be viewed at El Malpais National Monument in New Mexico. This landscape was formed three million years ago by lava flows from over thirty volcanoes. A geological field guide by Laughlin et al. (1993) provides the chemical compositions of basalts sampled from eight lava flows at El Malpais. Figure 3 tabulates these measurements. The columns, which you can think of as the “process variables” in this example, correspond to 10 oxides, and the rows correspond to the samples.

```
      Chemical Composition of Lava Flows

Flow           SiO2  TiO2  Al2O3  Fe2O3   MnO

Bluewater     51.62  1.25  15.13  11.49  0.16
Laguna        50.23  1.53  14.50   1.82  0.17
McCartys      51.48  1.41  15.18  11.87  0.16
TwinCraters   48.86  1.44  14.84  12.48  0.17
Bandera       44.47  3.04  15.22   4.39  0.15
RamahNavajo   50.70  1.17  15.05  11.66  0.16
FenceLake     50.03  1.38  14.92  12.24  0.17
NorthPlains   52.06  1.45  15.72  10.95  0.15


Flow           MgO   CaO  Na2O   K2O  P2O5

Bluewater     7.42  9.30  2.60  0.42  0.15
Laguna        9.45  8.83  2.91  0.77  0.22
McCartys      8.29  9.11  2.78  0.69  0.19
TwinCraters   9.15  8.87  2.81  0.74  0.22
Bandera       9.30  8.80  3.38  1.60  0.58
RamahNavajo   8.34  9.57  2.44  0.36  0.14
FenceLake     9.00  9.16  2.74  0.64  0.19
NorthPlains   6.34  9.99  2.79  0.66  0.22
```

Figure 3. Chemical Compositions of Basalts

Although this is a small table, it is not easy to compare and classify the samples because the oxides are correlated, and the samples represent points in 10-dimensional space. You can use the PRINCOMP procedure to compute a PCA for the data, which have been saved in a SAS data set named BASALT.

```
ods output eigenvectors = eigen;
proc princomp data=basalt out=prin std;
   var SiO2 TiO2 Al2O3 Fe2O3 MnO
       MgO CaO Na2O K2O P2O5;
run;
```

The output, shown in Figure 4, indicates that 59% of the variation is explained by the first principal component, and a further 31% is explained by the second principal component, leaving only 10% to be explained by the remaining principal components. Consequently, you can use the first two components to interpret the data.

The output data set PRIN contains the original data together with variables named Prin1 to Prin10, which correspond to $\mathbf{t}_1, \ldots, \mathbf{t}_{10}$ and provide the scores for the samples. The following statements plot the first two scores.

```
         Eigenvalues of the Correlation Matrix

     Eigenvalue    Difference    Proportion    Cumulative

 1   5.91766199    2.80458742        0.5918        0.5918
 2   3.11307457    2.49141767        0.3113        0.9031
 3   0.62165690    0.41853630        0.0622        0.9652
 4   0.20312060    0.10349710        0.0203        0.9856
 5   0.09962351    0.05921690        0.0100        0.9955
 6   0.04040661    0.03595079        0.0040        0.9996
 7   0.00445582    0.00445582        0.0004        1.0000
 8   0.00000000    0.00000000        0.0000        1.0000
 9   0.00000000    0.00000000        0.0000        1.0000
10   0.00000000                      0.0000        1.0000
```

Figure 4. Eigenvalues for Basalt Data

```
data pltanno; set prin;
   length text $ 11;
   retain function 'label' position '1'
          hsys '3' xsys '2' ysys '2'
          color 'black' style 'swiss';
   text = flow; x = prin1; y = prin2;
   run;

title "Score Plot for Basalt Samples";
symbol v=dot i=none;
proc gplot data=prin;
   plot prin2 * prin1 / anno=pltanno
        vaxis=axis1 haxis=axis2 frame;
axis1 label=(a=90 r=0
        "Score on Second Component")
        minor=none;
axis2 label=("Score on First Component")
        minor=none;
run;
```

The display in Figure 5 reveals that the samples from Bandera and Laguna are strongly differentiated from the other samples by their scores for the first principal component. The samples from Fence Lake and North Plains have nearly identical scores on this axis. In fact, Laughlin et al. (1991) point out that these flows are chemically very similar (although they do not indicate how they arrived at this conclusion.)

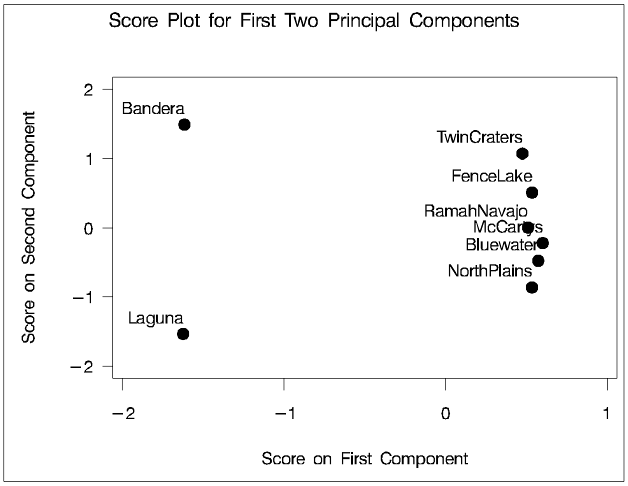

Figure 5. Score Plot for Basalt Data

You can use loading plots such as the one in Figure 6 to relate principal components back to the variables. These plots display the eigenvalues (saved in the data set EIGEN) corresponding to the variables for pairs of principal components. Figure 6 shows that the first two components are composed of contrasts between the oxides SiO2 and Fe2O3 and the average of the other oxides.

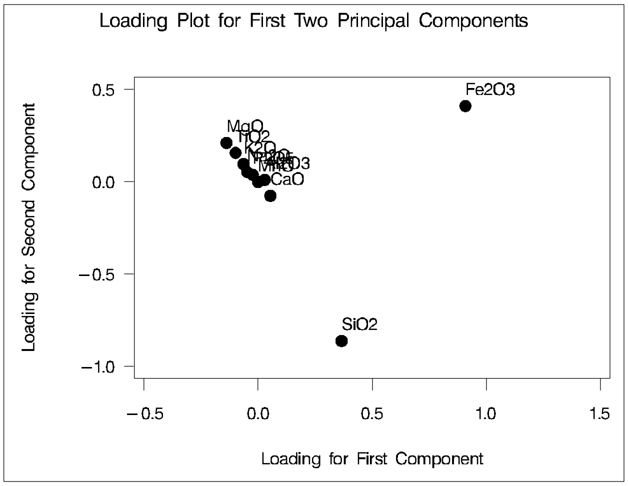

Figure 6. Loading Plot for Basalt Data

In applications to process monitoring discussed in the next section, score plots serve as "windows" for revealing relationships, outliers, and changes in multivariate process samples. Loading plots, combined with engineering knowledge or business rules, are helpful for determing causes for unusual behavior.

The SAS System offers a variety of facilities for computing principal components. For example:

1. If you are exploring your data interactively, you can request a PCA with the Multivariate Analysis task in SAS/INSIGHT® software. In this environment, you can create score plots that are dynamically linked to other graphical views of the data and to a data table.

2. You can also obtain a PCA with the PLS procedure (as demonstrated later in this paper) by specifying the variables on both sides of the MODEL statement:

```
proc pls data=basalt;
   model SiO2 TiO2 Al2O3 Fe2O3 MnO
         MgO CaO Na2O K2O P2O5 =
         SiO2 TiO2 Al2O3 Fe2O3 MnO
         MgO CaO Na2O K2O P2O5;
run;
```

The next section applies PCA to the problem of monitoring a process over time.

Multivariate Process Monitoring Based on Principal Components

Although the term *process* is most often associated with manufacturing, a process is any sequence of interconnected activities with inputs and outputs. Whether the final output is a manufactured product or a service, variation is present in all aspects of a process. Knowledge of the natural or "common cause" variation in the process is the basis for identifying unusual or "special" causes of variation, and statistical control of a process is brought about by eliminating such causes. Once the process is stable, its quality costs are predictable, and quality can be improved by reducing common cause variability in the system. Likewise, knowledge of the relationships between process and product variables can be leveraged to optimize the process.

The Shewhart chart is the most widely used statistical tool for establishing control and monitoring a process. The control limits indicate the expected common cause variation, and a point outside the limits signals a special cause; refer to Montgomery (1996). Extensions of the Shewhart chart for multivariate process measurements based on Hotelling's T^2 statistic were first proposed in the 1940s, and they continue to be developed; refer to Alt (1985) and Lowry and Montgomery (1995).

The T^2 chart is appropriate for situations where the process measurements are continuous, and where it is important to detect changes in their linear relationships as well as in their means and variances. Denote the measurement on the jth variable at time t as X_{tj} for $t = 1, 2, \ldots, n$, where n is the number of time periods, and $j = 1, 2, \ldots, k$, where k is the number of variables. A T^2 chart plots the statistic

$$T_t^2 = (\mathbf{X}_t - \bar{\mathbf{X}}_n)\mathbf{S}_n^{-1}(\mathbf{X}_t - \bar{\mathbf{X}}_n)'$$

at time t, where $\mathbf{X}_t = [X_{t1} \ldots X_{tk}]$, $\bar{\mathbf{X}}_n = [\bar{X}_1 \ldots \bar{X}_p]$, $\bar{X}_j = \frac{1}{n}\sum_{t=1}^{n} X_{ij}$, and

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{t=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_n)'(\mathbf{X}_i - \bar{\mathbf{X}}_n)$$

Control limits for T_t^2 are derived by assuming that $\mathbf{X}_t$ has a k-dimensional multivariate normal distribution. The control limits for T_i^2 are computed as percentiles of a χ^2, F, or beta distribution, depending on whether

- $\bar{\mathbf{X}}_n$ and $\mathbf{S_n}$ represent estimates or known values of the mean and covariance of the distribution
- X_{tj} is an individual measurement or a sample mean

You can construct T^2 charts with the SHEWHART procedure in SAS/QC® software; refer to the chapter on “Specialized Control Charts” in the *SAS/QC User’s Guide, Version 8* (1999).

In practice, the T^2 chart has two major drawbacks. First, it does not scale well to large numbers of process variables ($k > 20$), particularly when the variables are collinear. Second, it is difficult to interpret a point outside the control limits. Several useful approaches have been proposed for interpreting out-of-control points; refer to Doganaksoy et al. (1991), Hawkins (1991, 1993), Mason, Tracy, and Young (1995), and Mason and Young (2001).

The approach described here was proposed by Kourti and MacGregor (1995, 1996) to deal with both limitations. Because it is based on a PCA model, it can handle hundreds or even thousands of process variables. The model can be extended to deal with time series effects and to incorporate the relationship between process and product (quality) variables (see the next section). You can quantify the common cause variability of the process with an ellipsoid in a low-dimensional score space, as illustrated in Figure 7. A point outside the ellipse, like a point outside the control limits on a T^2 chart, signals a change that you should investigate. This approach also offers a number of diagnostic tools for interpreting the point.

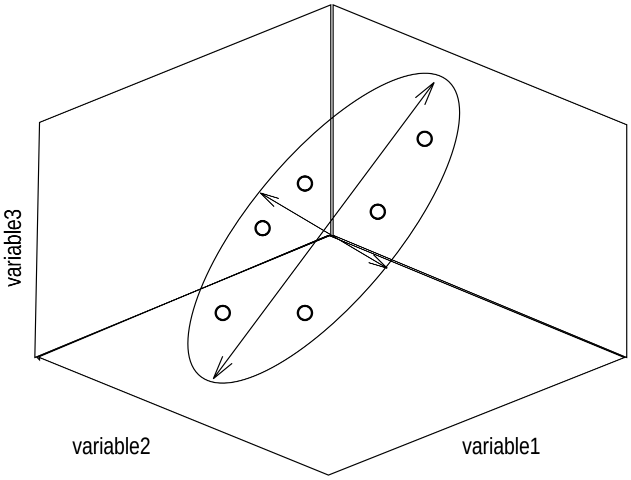

Figure 7. Control Ellipse in Projection Plane

To illustrate this approach, consider the problem faced by NASD Regulation (NASD-R), an independent subsidiary of the National Association of Security Dealers (NASD) charged with regulating the securities market and Nasdaq. NASD-R oversees 6,000 securities firms, each of which reports more than 100 variables at regular time intervals. The firm variables range from financial indicators to the frequency of customer complaints. NASD-R examiners use the data to select firms for the investigation of customer complaints, financial problems, and questionable sales practices. To decide how to allocate regulatory resources, statistical process control methods are used to find unusual patterns and sources of variation.

The illustrative data set PEERFIRMS used here contains seven variables, a small subset of the variables reported by a peer family of 54 firms for 12 quarters. A partial listing is shown in Figure 8.

```
                                           CompPer
   firmkey     peerkey    timekey          Sales     NReps

      7870          42         92         0.6135       904
      7870          42        183         0.2724      1049
      7870          42        275         0.4384      1117
      7870          42        367         0.4624      1117
       ...         ...        ...            ...       ...

MeritSales                  CompPer                  Excess
      Comp  MeritTerms          Rep  RevPerRep      Capital

        27           0       2.9867   68596.86     43387669
         9           0       0.8580   46791.96     25965370
        17           0       1.5219   49815.30     32599669
        23           0       2.0591       0.00      6053815
       ...         ...      ...            ...         . ..
```

Figure 8. Partial Listing of Firm Variables

You can build a PCA model for these variables using the PLS procedure as follows:

```
proc pls data=peerfirms nfac=3;
   model CompPerSales NReps MeritSalesComp
         MeritTerms CompPerRep RevPerRep
         ExcessCapital =
         CompPerSales NReps MeritSalesComp
         MeritTerms CompPerRep RevPerRep
         ExcessCapital;
   output out=pcastats tsquare=t2
          stdysse=sse yscore=yscr;
run;
```

The output, shown in Figure 9, shows that three factors (principal components) account for 82% of the variation.

```
                    The PLS Procedure

              Percent Variation Accounted for
              by Partial Least Squares Factors

Number of
Extracted        Model Effects          Dependent Variables
  Factors      Current       Total      Current       Total

        1      41.2846     41.2846      41.2846     41.2846
        2      25.5462     66.8308      25.5462     66.8308
        3      14.8343     81.6651      14.8343     81.6651
```

Figure 9. PCA Model for All Firms in Peer Family

The output data set PCASTATS contains a variable named T2, which is the sum of squares of the scores for the first three principal components, scaled by their variances:

$$T_t^2 = \frac{t_1^2}{s_{t_1}^2} + \frac{t_2^2}{s_{t_2}^2} + \frac{t_3^2}{s_{t_3}^2}$$

When all k terms are included in the sum, this statistic is equivalent to T_t^2 plotted on a traditional T^2 chart.

Now, suppose you are interested in monitoring Firm 7870 by making a control chart for T2. Begin by augmenting PCASTATS with the special variables required by the TABLE= input data set for PROC SHEWHART; refer to "Input Data Sets" in the chapter on "Specialized Control Charts" the *SAS/QC User's Guide, Version 8* (1999).

```
proc means data=pcastats noprint;
   var CompPerSales NReps MeritSalesComp
       MeritTerms CompPerRep RevPerRep
       ExcessCapital;
   output out=count (keep=_freq_) n=n1-n7;
run;

data pcastats (rename=(t2=_subx_));
   length _var_ $ 8;
   if _n_ = 1 then set count;
   rename _freq_ = n;
   set pcastats;
   _var_    = 'tsquare';
   _alpha_  = 0.05;
   _subn_   = 1;
   _limitn_ = 1;
   p        = 3;
run;
```

Next, add control limit variables computed with the beta distribution, which is appropriate since the principal components are estimates, and X_ij is an individual measurement; refer to Tracy, Young, and Mason (1992).

```
data pcastats;
   set pcastats;
   _lclx_   = ((n-1)*(n-1)/n)*
              betainv(_alpha_/2, p/2, (n-p-1)/2);
   _mean_   = ((n-1)*(n-1)/n)*
              betainv(0.5, p/2, (n-p-1)/2);
   _uclx_   = ((n-1)*(n-1)/n)*
              betainv(1-_alpha_/2, p/2, (n-p-1)/2);
run;
```

Now you can read PCASTATS with PROC SHEWHART to create the chart for T2, which is shown in Figure 10.

```
symbol value=dot;
title 'T' m=(+0, +0.6) '2'
          m=(+0, -0.6) ' Chart for Firm 7870';
proc shewhart table=pcastats;
   where firmkey=7870;
   xchart tsquare * timekey /
      xsymbol = 'Median'
      npanel  = 1300
      nolegend;
   label _subx_   = 'T-squared'
         timekey = 'Time in Days';
run;
```

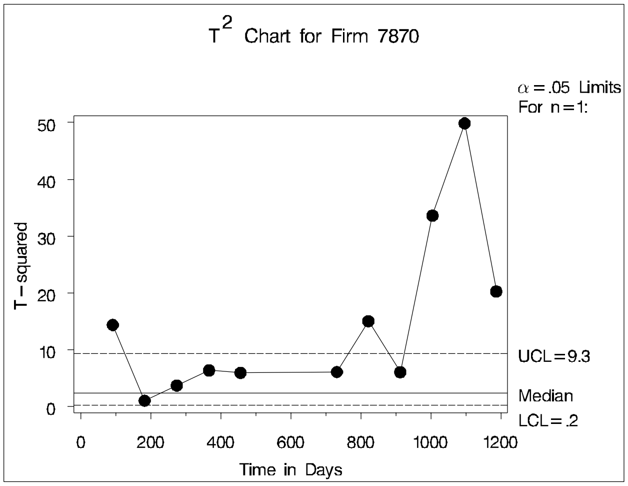

Figure 10. T^2 Chart for Firm 7870

The chart reveals unusual variation for Firm 7870 in the three most recent quarters. You can visualize how this variation departs from the normal pattern for the other firms in the family by making a scatter plot of the scores t_1 and t_2 for all the firms and adding a 95% confidence ellipse. These are saved in PCASTATS as the variables YSCR1 and YSCR2. In Figure 11, the points for Firm 7870 are connected and identified.

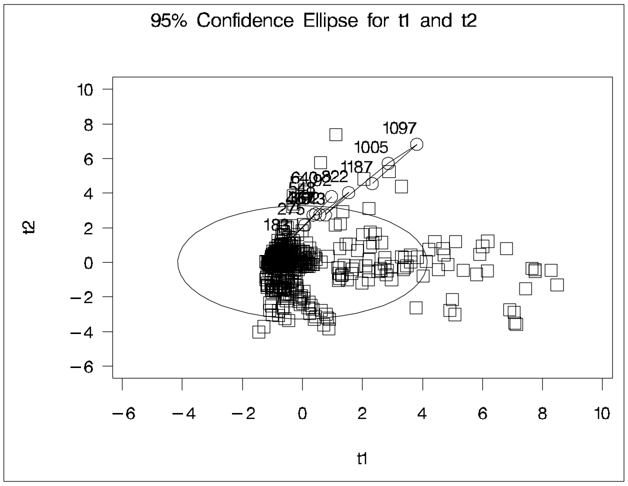

Figure 11. Score Plot for Firm 7870

Figure 12 shows only the points for Firm 7870 and reveals more clearly how its process is wandering away from the normal region of variability for its peers. Of course, you should also examine plots of t_1 versus t_3 and t_2 versus t_3, which are not shown here.

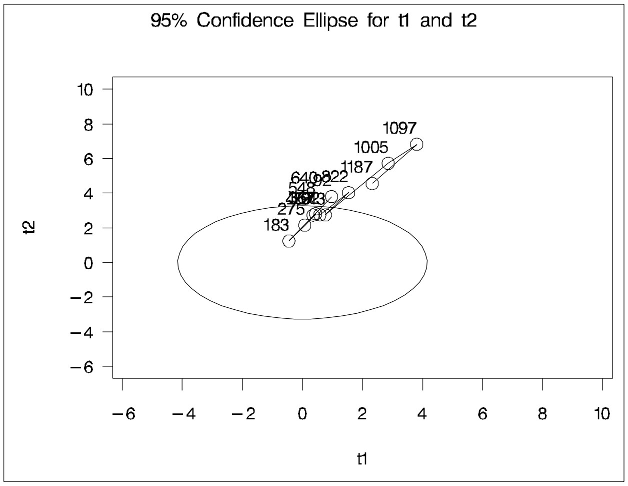

Figure 12. Score Plot for Firm 7870

There are two explanations to consider for this behavior. The first is that the process has moved outside the control ellipse but is still within the hyperplane defined by the PCA model. The second is that the process has moved off the hyperplane and has changed in a way that is not captured by the model.

You can check for the second possibility by plotting the distance of each time point to the model hyperplane, or equivalently, the squared prediction error (SPE). This statistic is saved as the variable SSE in the data set PCASTATS, and you can create an SPE chart as follows:

```
data sse; set pcastats;
   keep timekey firmkey sse;
proc sort data=sse;
   by timekey;

proc means noprint data=sse;
   var sse;
   output out=spelimits (drop=_freq_ _type_ )
          n=n mean=mean var=var;
   by timekey;

data spelimits;
   set spelimits;
   _var_     = 'spe'
   _alpha_   = 0.05;
   _limitn_  = 2;
   _subn_    = 2;
   _lcls_    = ( var / ( 2.0 * mean ) ) *
               cinv(_alpha_/2,2*mean*mean/var);
   _s_       = ( var / ( 2.0 * mean ) ) *
               cinv(0.5,2*mean*mean/var);
   _ucls_    = ( var / ( 2.0 * mean ) ) *
               cinv(1.0-_alpha_/2,2*mean*mean/var);
```

```
data sse;
   rename sse = _subs_;
   merge sse spelimits;
   by timekey;
   if firmkey = . then delete;
run;

symbol value=dot;
title 'SPE Chart for Firm 7870';

proc shewhart table=sse;
   where firmkey=7870;
   schart spe * timekey /
      ssymbol = 'Median'
      llimits = 1
      npanel  = 1300
        nolegend;
   label _subs_  = 'SPE'
         timekey = 'Time in Days';
run;
```

Note that the control limits for SSE are computed using all the firms in the peer family as a reference normal data set; refer to Nomikos and MacGregor (1995). Alternatively, you can compute the limits using the method of Jackson and Mudholkar (1979), which assumes a multivariate normal distribution for the measurements.

The SPE chart displayed in Figure 13 indicates that the process has moved off the model plane. Not only is this firm's behavior diverging from the control region, it is also diverging in a new way with variation not observed in the data that was used to develop the model. If this behavior persists and is stable, you should consider constructing a new model.

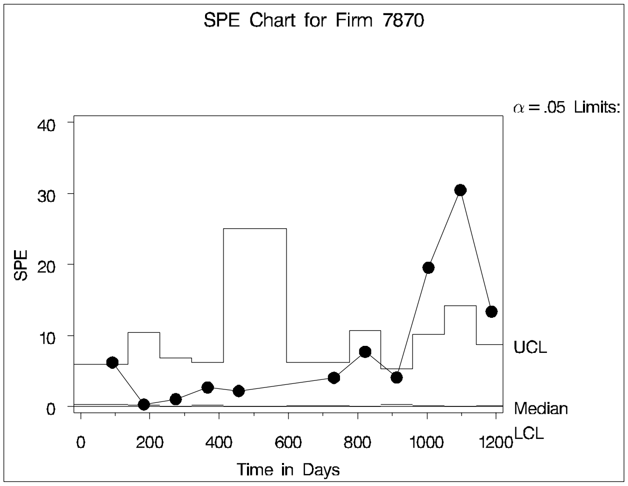

Figure 13. SPE Chart

Another way to diagnose the behavior in Figure 12 is with a contribution plot, which tells you which variables contribute to the "gap" between a point such as the one at TIMEKEY=1097 and the center of the ellipse. For each variable in the model, the contribution plot displays a root sum of squares of weighted residuals.

You can compute these quantities by applying scale factors, available in an ODS output data set from PROC PLS, to the PCA residuals as follows:

```
ods output YVariableCenScale=YCS;
ods listing close;
proc pls data=peerfirms nfac=3 censcale;
   model CompPerSales NReps MeritSalesComp
         MeritTerms CompPerRep RevPerRep
         ExcessCapital =
         CompPerSales NReps MeritSalesComp
         MeritTerms CompPerRep RevPerRep
         ExcessCapital;
   output out=stdres
          stdy=stdy stdysse=stdysse
          yresidual=yres1-yres7;
run;
```

You can do the rescaling with PROC IML:

```
proc iml;
   use stdres;
   read all var ("timekey")       into TIME;
   read all var ("yres1":"yres7") into YR;
   use YCS;
   read all var {Scale} into YS;

   YR = YR*diag(1/YS);

   data = TIME || YR;
   create Contribution
          var ( "timekey" || ("yres1":"yres7"));
   append from data;
quit;
```

The following statements arrange the contributions in a form suitable for plotting.

```
data ContribPlot; set Contribution;
   if ( timekey = 1097 );
proc transpose data=ContribPlot out=TContribPlot;
   var yres1-yres7;
run;
data TContribPlot;
   rename Col1 = Contribution;
   merge ycs TContribPlot;
run;
```

You can use the GCHART procedure to create the contribution plot, which is displayed in Figure 14.

```
title "Contribution Plot for Time 1097";
pattern1 value=solid color=ligr;

proc gchart data=TContribPlot;
   hbar variable /
      sumvar=Contribution
          nostats
          href  = -2 2
          lref  = 3
          clipref
          raxis = axis1
          maxis = axis2;

axis1 label=("Contribution")
      major=(number=5) minor=none;
axis2 label=none;
run;
```

The plot shows that two complaint variables and a variable measuring merit terminations contributed to the jump at TIMEKEY=1097. Events related to these variables should be examined to determine the causes for this variation.

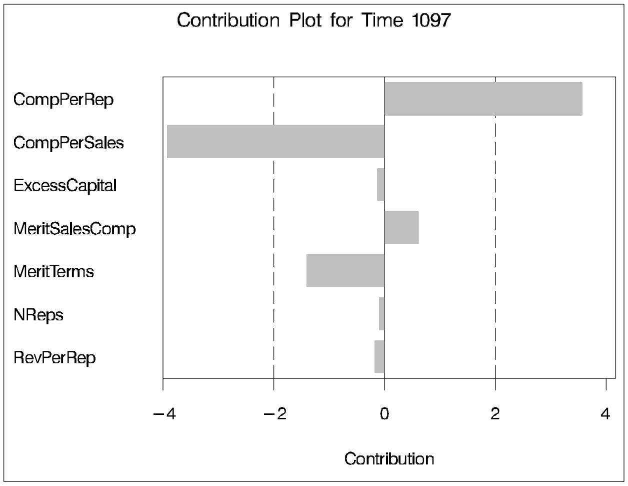

Figure 14. Contribution Plot for TIMEKEY=1097

PLS Modeling

As discussed in the preceding sections, PCA is a method for describing the variation among many variables with a few latent factors. Now, suppose that in addition to describing a block of variables $\mathbf{X}$, you want to use $\mathbf{X}$ to *predict* another block of variables $\mathbf{Y}$. For example, in a manufacturing process $\mathbf{X}$ might represent a set of process variables, and $\mathbf{Y}$ might represent a set of quality variables. Or, in the case of the NASD-R data, $\mathbf{X}$ might represent the financial variables for a set of firms, and $\mathbf{Y}$ might represent the complaint variables.

There are many ways to approach this problem when $\mathbf{X}$s and $\mathbf{Y}$s have relatively few variables, and multiple linear regression (MLR) is one of the most common techniques. However, when $\mathbf{X}$ and $\mathbf{Y}$ have many variables—perhaps more than you have observations—or when there are strong correlations among the variables, then most common techniques (including MLR) will fail.

Partial least squares (PLS)–also known as projection to latent structures–is a method for predicting $\mathbf{Y}$ variables from $\mathbf{X}$ variables that works when you have many correlated variables. Like PCA, PLS extracts latent factors that are functions of all the variables, but PLS extends PCA by extracting factors from *both* $\mathbf{X}$ and $\mathbf{Y}$. After centering and scaling, both $\mathbf{X}$ and $\mathbf{Y}$ are individually modeled with so-called outer relationships

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\,\mathbf{P}' + E\\ \mathbf{Y} &= \mathbf{U}\,\mathbf{Q}' + F\end{aligned}$$

and the relationship between $\mathbf{X}$ and $\mathbf{Y}$ is modeled through the so-called inner relationship

$$\mathbf{U} = \mathbf{bT}$$

The latent $\mathbf{X}$ factors are selected with the goal of explaining both $\mathbf{X}$ and $\mathbf{Y}$ variation, and the part of $\mathbf{Y}$ that a certain $\mathbf{X}$ factor predicts is given by the corresponding $\mathbf{Y}$ factor.

The PLS procedure implements partial least squares regression and related prediction techniques; refer to the chapter on the PLS procedure in the *SAS/STAT User's Guide, Version 8* (1999) for syntax and algorithms.

Historically, PLS emerged as an econometric and psychometric technique when Herman Wold developed the NIPALS algorithm in the 1960s as a general approach for relating any number of blocks of variables; refer to Wold (1966). In the 1980s, researchers in chemometrics (among them Herman Wold's son, Svante Wold) found two-block PLS useful for predicting chemical properties ($\mathbf{Y}$) from highly multivariate chemical measurements ($\mathbf{X}$), such as spectra; refer to Wold, Martens, and Wold (1983). More recently, chemical process engineers have applied PLS to predict process quality characteristics based on the hundreds or even thousands of process variables that are measured in a modern industrial plant.

The following statements use PROC PLS to construct a PLS model with NFAC=4 latent factors for all the firms in a peer family for the current time period. The MODEL statement specifies a $\mathbf{Y}$ block of complaint variables and an $\mathbf{X}$ block of financial variables. The OUTPUT statement saves the extracted factors in a data set for subsequent processing and display.

```
proc pls data=firms nfac=4;
   model MeritSalesComp CompPerRep
         CompPerSales =
         NTransBERreports MeritTerms
         NReps NSelectNetOrders
         PerAlertsSelectNetOrders
         PerAlertsTransBEReports
         RevPerRep ExcessCapital ;
   output out=outpls xscore = xscr
                     yscore = yscr;
run;
```

The output shown in Figure 15 analyzes how much variation is individually and cumulatively explained by the first four PLS factors. Each factor explains successively less variation in both $\mathbf{X}$ and $\mathbf{Y}$. Cumulatively, the four factors account for most of the variation, so you can have some confidence in the predictive model.

```
                    The PLS Procedure

              Percent Variation Accounted for
             by Partial Least Squares Factors

Number of
Extracted         Model Effects           Dependent Variables
  Factors       Current       Total       Current       Total

        1       46.6854     46.6854       33.5840     33.5840
        2       19.2238     65.9092       11.5366     45.1206
        3       11.6540     77.5632        9.3138     54.4344
        4        6.4592     84.0224        4.1806     58.6150
```

Figure 15. Output from PROC PLS

Most of what PLS has to say about the data is best displayed graphically. Figure 16 and Figure 17 depict the first and second inner PLS relationships, with the two PLS scores for $\mathbf{X}$ plotted against the corresponding PLS scores for $\mathbf{Y}$. Given how much variation the first factor explains in both $\mathbf{X}$ and $\mathbf{Y}$ (see Figure 15), the strong linear trend in Figure 16 is expected. However, note the apparent influence of the two observations in the lower left of this plot. They are extreme for both axes, indicating that they are highly informative for the PLS analysis and should be confirmed.

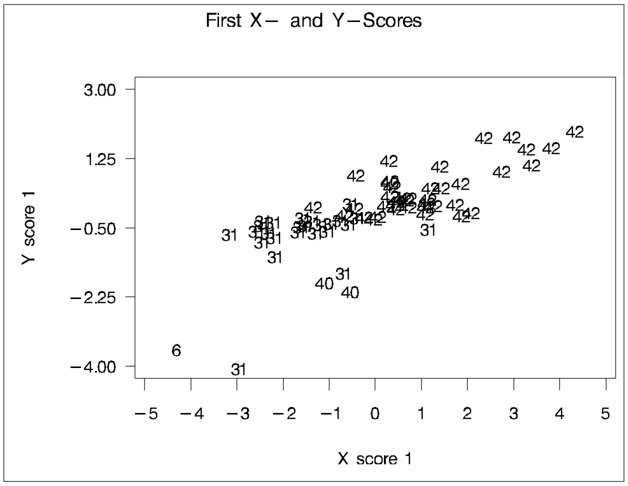

Figure 16. Score Plot for PLS Analysis

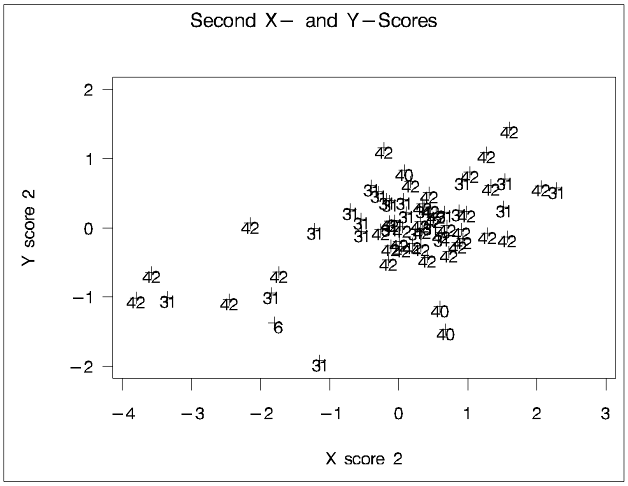

Figure 17. Score Plot for PLS Analysis

Another way to examine a PLS model is to plot the $\mathbf{X}$ scores against each other. Conceptually, this is much like PCA, except that the scores are selected not only because they account for $\mathbf{X}$ variation, but also because they predict $\mathbf{Y}$ well. Figure 18 displays this plot for the first two PLS factors, with the peer key again used as the plotting symbol.

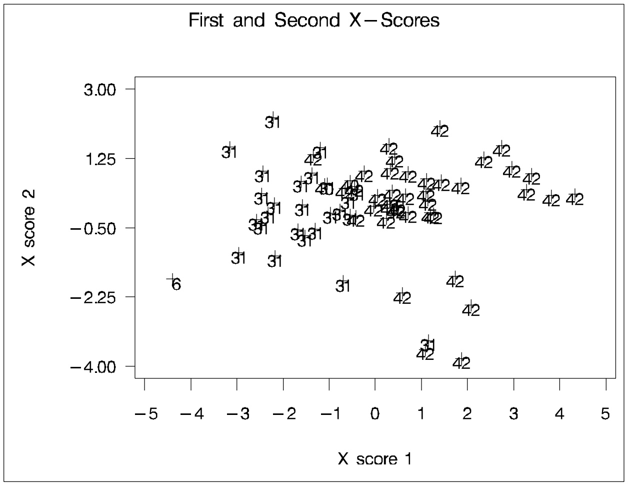

Figure 18. Score Plot for PLS Analysis

There is fairly good separation between firm peer groups in this plot. However, the PLS model distinguishes between two groups of firms with PEERKEY=42, namely, those that appear in the upper and lower right corners of the plot. This indicates that the PLS model is tracking the same features as those that went into composing the peer groups, but it has discovered additional features useful for predicting complaints, or perhaps for subdividing the peer group into two groups. As in PCA, the extracted factors in PLS are linear combinations of the centered and scaled $\mathbf{X}$ variables. Figure 19 displays the coefficients of this linear combination for the first factor, the one that explains the most $\mathbf{X}$ and $\mathbf{Y}$ variation.

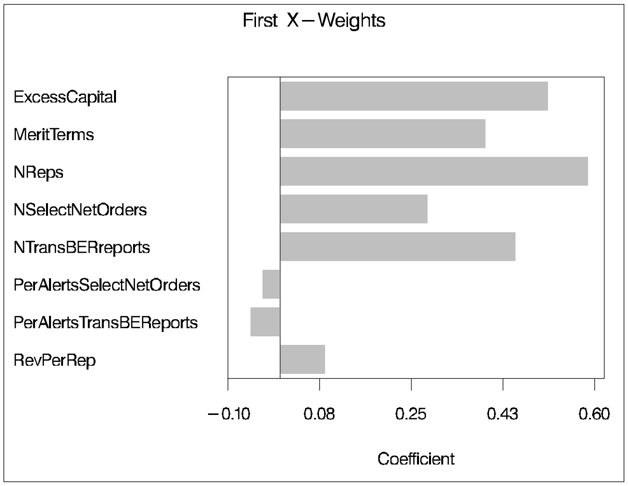

Figure 19. First X-weights for PLS Analysis

You can see that five variables contribute most strongly to this factor. However, you should not conclude that the model depends only on these terms, since PLS factors (like PCA factors) are functions of all the variables.

A critical step in formulating a PLS regression model is choosing the proper number of factors. You can add more factors to fit the training data better, but if you add too many you risk *overfitting*—that is, tailoring the model too closely to the data at hand and making it unfit for predicting new data. You can choose the number of factors by validation—training the model on one set of data and testing how well it predicts another set—or cross-validation—where the training and test data sets are selected from the same set of original data. The CV= and CVTEST options in PROC PLS implement various kinds of validation.

Extensions of PLS Models

The PLS algorithm was originally devised to relate more than just two blocks of variables, and such *multi-block* methods have been used to analyze complex processes; refer to Wangen and Kowalski (1988). Another extension of PLS comes from the fact that measurements sometimes have more than two indices (“variables” and “observations”). For example, in the NASD-R data, measurements are indexed by financial indicator, firm, and time. In this case, *three-way* (or more generally, *multiway*) PLS may be appropriate for process monitoring. Multiway models analyze variation in all the dimensions that index the data, potentially providing greater intuition about interrelationships between measurements over each dimension; refer to Wold et al. (1987).

Summary

You can use PCA methods to describe and monitor large numbers of process variables over time. You can summarize the process behavior with a small number of displays, including a T^2 chart, score charts, and an SPE chart. Contribution plots are helpful for diagnosing points outside the control limits. You can use PLS methods to predict quality variables from process variables. As in PCA, you can interpret the results using a small number of displays. These techniques are highly developed within the field of chemometrics, but they deserve to be explored and applied to process data in other areas.

References

Alt, F. B. (1985), "Multivariate Quality Control," *Encyclopedia of Statistical Sciences*, Volume 6, edited by N. L. Johnson and S. Kotz, New York: Wiley & Sons.

Beebe, K. R., Pell, R. J., and Seasholtz, M. B. (1998), *Chemometrics: A Practical Guide*, New York: Wiley & Sons.

Dayal, B. S., MacGregor, J. F., Taylor, P. A., Kildaw, R., and Marcikic, S. (1994), "Application of Feedforward Neural Networks and Partial Least Squares Regression for Modelling Kappa Number in a Continuous Kamyr Digester," *Pulp & Paper Canada*, 95, T7–T13.

Doganaksoy, N., Faltin, F. W., and Tucker, W. T. (1991), "Identification of Out-of-Control Quality Characteristics in a Multivariate Manufacturing Environment," *Communications in Statistics–Theory and Methods*, 20, 2775–2790.

Hawkins, D. M. (1991), "Multivariate Quality Control Based on Regression-Adjusted Variables," *Technometrics*, 33, 61–75.

Hawkins, D. M. (1993), "Regression Adjustment for Variables in Multivariate Quality Control," *Journal of Quality Technology*, 25, 170–182.

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

Jackson, J. E. and Mudholkar, G. S. (1979), "Control Procedures for Residuals Associated With Principal Component Analysis," *Technometrics*, 21, 341–349.

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.

Laughlin, A. W., Charles, R.W., Reid, K., and White, C. (1993), *Field-trip Guide to the Geochronology of El Malpais National Monument and the Zuni-Bandera Volcanic Field, New Mexico*, Bulletin 149, Socorro, New Mexico: New Mexico Bureau of Mines & Mineral Resources.

Lowry, C. A. and Montgomery, D. C. (1995), "A Review of Multivariate Control Charts," *IIE Transactions*, 27, 800–810.

Mason, R. L., Tracy, N. D., and Young, J. C. (1995), "Decomposition of T^2 for Multivariate Control Chart Techniques," *Journal of Quality Technology*, 27, 99–108.

Mason, R. L. and Young, J. C. (2001), *Multivariate Statistical Process Control: With Industrial Applications*, Philadelphia, PA: ASA-SIAM (forthcoming).

Montgomery, D. C. (1996), *Introduction to Statistical Quality Control*, Third Edition, New York: Wiley & Sons.

Nomikos, P. and MacGregor, J. F. (1995), "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, 37, 41–59.

SAS Institute Inc. (1999), *SAS/QC User's Guide, Version 8,* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 8,* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999), "The Quality Data Warehouse: Serving the Analytical Needs of the Manufacturing Enterprise," *SAS Institute White Paper*, available at http://www.sas.com.

Tracy, N. D., Young, C., and Mason, R. L. (1992), "Multivariate Control Charts for Individual Observations," *Journal of Quality Technology*, 24, 88–95.

Wangen, L. and Kowalski, B. (1988), "A Multi-block Partial Least Squares Algorithm for Investigating Complex Chemical Systems," *Journal of Chemometrics*, 3, 3-20.

Wold, H. (1966), "Estimation of Principal Components and Related Models by Iterative Least Squares," in *Multivariate Analysis*, ed. P. R. Krishnaiah, New York: Academic Press, 391–420.

Wold, S. (1995), “PLS for Multivariate Linear Modelling,” in *QSAR: Chemometric Methods in Molecular Design*, edited by van deWaterbeemd. Methods and Principles in Medicinal Chemistry, Vol 2. Weinheim: Verlag Chemie.

Wold, S., Martens, H., and Wold, H. (1983), “The Multivariate Calibration Problem in Chemistry Solved by the PLS Method,” *Proc. Conf. Matrix Pencils* (A. Ruhe, B. Kagstrom, eds.), March 1982, *Lecture Notes in in Mathematics,* Heidelberg: Springer Verlag, 286–293.

Wold, S., Geladi, P., Esbensen, K. and Ohman, J. (1987), “Multi-Way Principal Components and PLS Analysis,” *Journal of Chemometrics*, 1, 41-56.

Acknowledgments

We would like to express our appreciation to David DeNardis, Ray Michie, and Jon Weisz of SAS Institute for assistance in the preparation of this paper.

Contact Information

Robert N. Rodriguez, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 531-7650, FAX (919) 677-4444, Email Bob.Rodriguez@sas.com.

Randall D. Tobias, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 531-7933, FAX (919) 677-4444, Email Randy.Tobias@sas.com.