To the Graduate Council:

I am submitting herewith a thesis written by Harold David Miller entitled "Application of Partial Least Squares and Principal Component Analysis to Control and Monitoring of a Reactive Distillation Column." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Chemical Engineering.

Dr. Charles F. Moore, Major Professor

We have read this thesis
and recommend its acceptance:

Accepted for the Council:

Associate Vice Chancellor
and Dean of the Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at The University of Tennessee, Knoxville, I agree that the Library shall make it available to borrowers under the rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made.

Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in his absence, by the Head of Interlibrary Services when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature 

Date OCTOBER 21, 1994

APPLICATION OF PARTIAL LEAST SQUARES AND PRINCIPAL COMPONENT ANALYSIS TO CONTROL AND MONITORING OF A REACTIVE DISTILLATION COLUMN

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Harold David Miller
December 1994

ACKNOWLEDGMENTS

I would like to express my appreciation to a number of individuals and groups for their support during my graduate education.

Thanks to Dr. Charlie Moore, my committee chairman, for his willingness to help, his patience, and his good nature. Over the past six years he has fielded countless calls and e-mail messages as I solicited his advice, and he has always been encouraging and helpful. Thanks also to Dr. Peter Cummings and Dr. Mary Leitnaker for being willing to serve on my committee and taking the time to work with me over the past several months.

Thanks to Mr. Darrell Corpening, Mr. Bill Osborne, and Dr. Jim Downs for their support as supervisors during my graduate work. They have always provided encouragement and "gone to bat" for me in clearing administrative hurdles necessary for me to complete early portions of this work. Thanks also to Jim for his interest, sound advice, countless brilliant ideas, and enthusiastic "cheerleading" over the past five months as this thesis was being assembled.

Thanks to Dr. Steve Miller, Dr. Mike Paulonis, Dr. Ernie Vogel, Dr. Vera Williams, and all the members of the Eastman Advanced Controls Technology group in Kingsport. They have provided a nurturing and fun environment in which to work, and they have been especially helpful during the preparation of this thesis by sharing their own graduate experiences along with constructive technical and literary criticism which have greatly improved the quality of this document.

This work has been facilitated through the resources of Eastman Chemical Company, Tennessee Eastman Division, and I would like to

express my appreciation to the management of the Engineering & Construction Division for supporting this work.

Thanks to my parents, Harold and Linda Miller, for providing me with a wonderful childhood, spending time with me, and always expressing to me their belief that I could do anything I wanted, and that all they wanted for me in life was happiness.

And finally, thanks to Kellie, my wife. Over the past six years she has endured a sometimes harried, sometimes ecstatic, and sometimes frustrated husband with nothing less than total support and love. I think it is fair to say that the only person more glad than I am to see this work finished is Kellie.

DEDICATION

This work is dedicated to my mother, Linda Morton Miller. Her life is the greatest example I've ever known of strength and faith.

ABSTRACT

Today's typical chemical process has attached to it a vast, complex "nervous system" of sorts -- an array of sensors, valves, wires, and control computers which frequently monitor and store information regarding hundreds or thousands of measurements. In addition to these continuous measurements, key process and product streams are periodically analyzed to measure quality parameters such as composition or viscosity. The sheer volume of data available often results in "data overload" for those trying to improve or troubleshoot a process, since a human without data analysis tools is unable to interpret large sets of raw data and make analytical judgements regarding the state of the process. A number of multivariate analysis techniques, designed to use the data set as a collection of related information instead of as a group of individual measurements, have been developed to address this problem.

Over the past 15 years, a data analysis technique known as principal component analysis (PCA) has been adopted and applied to process data to perform multivariate process analysis. This method results in measurements which indicate variability in the process that is consistent with a base data set and variability that is not consistent with a base data set. These measurements can be collected and combined to make efficient use of the whole data set for diagnosis of process conditions and fault detection.

Another multivariate technique which has received recent attention in the area of process data analysis is partial least squares (PLS). PLS can be used to predict variables which might be hard to acquire on a continuous basis using available process measurements. Such

inferred quality measurements can be used for process monitoring and, in some cases, for closed loop control.

In this work, the usefulness of multivariate statistical monitoring tools such as PCA and PLS is assessed through their application to a challenging process control problem. The problem examined is control of a reactive distillation column in which the use of internal composition information for closed loop control has been identified as a key need. A control strategy for the column is designed using control analysis tools. A PLS-based estimate of the internal composition is developed and used for closed loop control on a rigorous simulation of the column. PCA-based process monitoring is then applied to the column model to help identify excursions from normal operation which suggest degradation of the inferred composition measurement.

The PLS composition estimator accurately predicts the composition of interest when disturbances are applied that are contained in a base data set used for PLS regression. When disturbances are applied that are not included in the base data set, the estimator performance degrades. For those cases in which the estimate is reasonably accurate, closed loop control based on the estimate is successful in controlling product impurity levels. Closed loop control is considerably less successful in cases where the estimate is inaccurate.

The PCA process monitor is successful in differentiating between areas of process operations consistent with the base data set and areas which are not. This is accomplished through the use of the PCA residual statistic, or Q. This measure can be used as an indicator to the user of regions of operation where the estimator has the potential for gross error due to input data behavior not modeled in the PLS regression.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the 1990s, the typical chemical process has attached to it a vast, complex "nervous system" of sorts -- an array of sensors, valves, wires, and control computers which delivers information regarding hundreds or thousands of measurements every second or so to the operator, process improvement engineer, operations supervisor, and manager. In addition to these continuous measurements, key process and product streams are periodically analyzed to measure quality parameters such as composition or viscosity. It is becoming more common to have data environments in which these data can be viewed as tables, trends, and statistics. However, the sheer volume of data available often results in "data overload", since a human without data analysis tools sitting at a computer terminal is unable to interpret the raw data and make analytical judgements regarding the state of the process. More often than not, operators, engineers, and managers rely on a few key variables and ignore the bulk of the data available to them.

The "data overload" problem has two faces. First, the process data being collected exhibit a high degree of correlation -- the number of process measurements being made is larger than the actual number of process states or degrees of freedom of the data. *Compression* of these data into fewer variables would preserve the real process information while reducing the number of measurements to be considered.

The second face is also related to correlation. It is possible to draw information from how measurements covary as well as how they vary independently. In the case of correlated data, it is not only changes in the actual values that are of interest, but also changes

in how the collection of values relate to each other. This data *extraction* can lead to a deeper understanding of the process, and can be used in some cases to infer difficult to measure but valuable quality parameters by observing how these parameters covary with the process measurement data.

The redundancy of process data commonly found in today's plants can be exploited by checking incoming data to sense changes in the process or its sensors -- an approach known as process fault detection. In this approach, a process model is developed (empirically or theoretically) and used to predict process measurements. The predicted and actual measurements are compared, and the residuals between the two are tracked to monitor plant-model match.

Currently, many plants utilize an established quality control system based on univariate statistical process control monitoring. This approach is satisfactory for independent, uncorrelated variables, but misses any important information contained in the correlation between variables. The objectives of multivariate process monitoring techniques are to monitor the process using all the data in concert to detect process upsets, malfunctions, or other special causes, and to infer quality parameters when economic or technical considerations make continuous measurement of these parameters infeasible.

Over the past 15 years, a data analysis technique known as principal component analysis (PCA) has been adopted and applied to process data to perform multivariate process analysis. PCA was first popularized in the field of psychometry, the measurement of mental or psychological data, due to the need to analyze data sets with many correlated variables and large amounts of noise and uncertainty.

PCA takes a body of data termed a base data set -- multiple observations of a collection of data points -- and characterizes it in terms of its "principal components" or "latent vectors" -- the axes in the data space along which the data most widely vary. This decomposition, closely related to singular value decomposition, results in measurements which indicate variability in the process that is consistent with the base data set and variability that is not consistent with the base data set. These measurements can be monitored and used to make efficient use of the whole data set for diagnosis of process conditions and fault detection. This is typically done through application of the latent vectors to future observations of the data points, testing consistency of the new observations with the base data set.

Another multivariate technique which has received recent attention in the area of process data analysis is partial least squares (PLS), also known as projection to latent structures. PLS is similar to PCA in that it characterizes multiple observations of a given set of data points, or "blocks" of data, using an alternative set of orthogonal axes; however using PLS the determination of these axes is based not only on how the data in a single block vary, but also on how two blocks of data covary. PLS can be used to predict variables which might be hard to acquire on a continuous basis using available process measurements. Such inferred quality measurements can be used for process monitoring and, in some cases, for closed loop control.

The development of PCA and PLS for process data analysis is an important area of activity in the growing field of chemometrics, the science of relating measurements made on a chemical system to the state of the system via application of mathematical or statistical methods.

The objective of this work is to assess the usefulness of multivariate statistical monitoring tools such as PCA and PLS in a challenging process control problem. The problem examined is control of a reactive distillation column, in which the use of internal composition information for closed loop control has been identified as a key need. In the course of the work, existing analysis tools will be evaluated and needed tools for the pilot effort will be selected or developed. A control study of the reactive distillation column will be performed, and will identify potential strategy improvements for which PLS can provide an useful internal composition estimate. A PLS-based estimate of the internal composition will be developed and used for closed loop control on a rigorous simulation of the column. PCA-based process monitoring will be applied to the column model to help identify excursions from normal operation which suggest degradation of the inferred composition measurement. Results of these simulations will be addressed, and future extensions of the technology which would facilitate industrial implementation of such a system will be identified.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Principal Component Analysis

The goal of principal component analysis (PCA) is to analyze the variability in a data set by identifying the dominant directions in which it varies. This is done by developing a set of basis vectors for a matrix based on these dominant directions and viewing data observations not only as a set of individual or univariate measurements, but as a multivariate observation whose behavior can be observed in terms of the new basis vectors.

In the analysis of process data, PCA is typically performed using a matrix composed of multiple observations of a set of individual variables. The variables are stored as columns, and the individual observations as rows. In this work, the number of observations will be referred to as k, and the number of variables as n; hence, such an organization results in a k by n data matrix.

In PCA, this k by n data matrix $\mathbf{X}$ is decomposed into the sum of the products of n pairs of vectors (Wise 1991, Kresta *et al.* 1991). Each pair is made up of a vector of length n, $\mathbf{p}_i$, known as the loadings, and a vector of length k, $\mathbf{t}_i$, called the scores. As a result, $\mathbf{X}$ can be written as a sum of inner products, or

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{t}_i \mathbf{p}_i^T$$

The first loading vector $\mathbf{p}_1$ is calculated as the 'orthogonal regression' of a line in $\mathbf{X}$ space through the data, a line positioned such that the sum of the perpendicular distances from the points to

the line is minimized (Kresta 1992), and describes the direction of greatest variability in the data. Next, the second loading vector $\mathbf{p}_2$ is calculated by fitting a line through the remaining data after variability lying in the $\mathbf{p}_1$ direction is subtracted from $\mathbf{X}$. This loading vector describes the greatest amount of the remaining variability. This procedure can be followed through for n vectors until all variability is mapped using the loading vectors, forming a new orthogonal set of basis vectors for the space spanned by $\mathbf{X}$ which are "ranked" in order of the amount of variability explained by the vectors.

These loading vectors are often referred to as latent variables, latent vectors, or principal components, since they are linear combinations of the original variables which together explain large pieces of information contained in the data set. Sometimes, these principal components can be associated with specific process states or groups of process variables. The $\mathbf{t}_i$s can be found by projecting $\mathbf{X}$ onto the corresponding $\mathbf{p}_i$:

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i$$

This decomposition is similar to the singular value decomposition (SVD), in which a data matrix $\mathbf{Z}$ is decomposed as

$$\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{S}$ is diagonal.

SVD can be used as a tool for PCA using the following procedure (Wise 1994):

1. Calculate the scatter of $\mathbf{X}$ [scatter($\mathbf{X}$) = $\mathbf{X}^T\mathbf{X}/(k-1)$]

2. Perform a SVD of scatter($\mathbf{X}$)

3. The V matrix from the SVD corresponds to the loading vectors $\mathbf{p}_i$, and the **S** matrix contains the eigenvalues corresponding to the relative variance explained for each of the latent vectors.

If the original data matrix **X** is mean centered, scatter(**X**) becomes the covariance matrix of **X**. If the data matrix is mean centered and scaled to unit variance, scatter(**X**) becomes the correlation matrix of **X**.

Jackson (1980) offers two theorems associated with the unique properties of principal component decomposition.

1. No linear combination of the original variables will explain more variation than the first component (also, no linear combination with at least one nonzero coefficient will explain less than the last principal component)

2. No linear combinations of rank *A* will explain more variation than the first *A* principal components

These theorems assert that the principal component method yields the orthogonal set of vectors which are most efficient at representing the maximum amount of information in the smallest dimensional space. The low dimension hyperplanes defined by the dominant latent vectors provide a low dimension window on behavior of very high dimensional processes. However, the true "dimension" of the process measurement space, or number of directions it can move independently, is almost always much lower than the number of independent sensor measurements; thus, such an approach is not only efficient but makes intuitive sense.

Determination of which latent variables are "dominant" can be done in a number of ways. The goal is to choose a number of principal components which represent a significant majority of the deterministic process variation, while including negligible contribution from stochastic variation. Choosing too few components leads to "wasting" real process information, while including too many can cause corruption of deterministic information with noise.

The simplest way to judge which latent variables are "dominant" is by considering the amount of variation in the data set explained by each one. The eigenvalues are indicators of the data set variance associated with the corresponding eigenvector. The fraction of variance in a given latent vector can be expressed as

$$\text{frac var in } \mathbf{p}_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

It is common to choose some threshold value, either a target total fraction of variance explained or a minimum fraction of variance for a given vector, as a test in selecting the number of latent variables to include in the PCA model.

Once the number of dominant vectors, or factors, that describe the major trends in the data have been identified, the concept of residuals can be considered. The residuals are the data variability from a base data set which are not explained via projection onto the A dominant vectors of the model. The residual data can be expressed as a matrix $\mathbf{E}$, yielding the expression

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E}$$

Adding to an earlier point, the goal in PCA is to segment the data variability into two parts: a deterministic part represented by the loading vectors, and a stochastic part represented by the residuals.

Up to this point, the discussion has focused on analysis of a base data set, yielding a set of latent vectors which describes a significant portion of the deterministic process information. Once these components are identified using a base data set, new process information can be interpreted via this framework.

2.2 Introduction to Partial Least Squares

PCA is a useful technique for compressing and extracting process information from a data set, and providing a framework by which future observations of the data can be measured against the base data set for consistency and variance from mean. However, in chemical processes, there often exist two distinct groups of data: process measurements which are readily available and collected on a frequent basis, and quality measurements which are often much harder to obtain and, as a result, collected less frequently. Often these quality measurements are acquired off-line through analytical techniques: therefore, even though the quality measurement may be directly related to the associated process measurements taken simultaneously, connection between the two for monitoring and control is cumbersome and difficult. In this work, process measurement data will be referred to as **X**, and quality measurement data as **Y**. Each is stored in a form consistent with the PCA discussion above, with the individual variables stored in columns and observations in rows. In fact, PCA can be performed on both the **X** and **Y** data sets to obtain descriptions of the latent vectors along which the data most widely vary.

While the individual latent vectors of **X** and **Y** available via PCA might provide some useful information, a more fruitful task might be the analysis of the **X** and **Y** data sets in concert to obtain some information about how the two data sets covary. Specifically, this could result in a regression by which the more difficult to acquire **Y** data can be estimated using the more frequent and easily available **X** data, providing an inferential measurement of key quality parameters.

One regression technique which has been widely discussed in the chemometrics community for prediction of **Y** data using **X** data is partial least squares (PLS). PLS decomposes both the *k* by *n* **X** matrix and the *k* by *m* **Y** matrix into the sum of the products of *n* pairs of vectors (Wise 1991, Kresta 1992). The scores and loadings terminology still holds, though in this case two sets of each are produced:

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{t}_i \mathbf{p}_i^T$$

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{u}_i \mathbf{q}_i^T$$

The main difference in PCA and PLS is in how the loadings, in which the directions of the latent vectors are stored, are formed. Instead of attempting to describe the dominant directions in **X** or **Y** variability, the PLS algorithm results in latent vectors determined using a combination of the directions of greatest variance in the data set of interest and the directions having the highest correlation with the other data set. In other words, the latent vectors for **X** are formed considering both variance in the **X** data and correlation of **X** data with **Y** data. A useful mathematical analogy for understanding PLS in terms of PCA is thinking of PLS as performing PCA on the covariance of **X** and **Y**, $\mathbf{Y}^T\mathbf{X}$ (Kresta 1992.)

One efficient method for performing the PLS regression is the NIPALS algorithm, which forms the latent variables resulting in dimension reduction one PLS dimension at a time. This allows the user to determine after each latent vector is formed whether additional vectors are needed to explain the variance and extract information for predicting **Y** variables from **X** data, making it more computationally efficient. A brief outline of the algorithm is described as follows (Wold *et al.* 1987, Kresta 1992):

0. Mean center and scale **X** and **Y**
1. Set **u** equal to a column of **Y**
2. $\mathbf{w}^T = \mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u}$ (regress columns of **X** on **u**)
3. Normalize **w** to unit length
4. $\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$ (calculate the scores)
5. $\mathbf{q}^T = \mathbf{t}^T\mathbf{Y}/\mathbf{t}^T\mathbf{t}$ (regress columns of **Y** on **t**)
6. Normalize **q** to unit length
7. $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}^T\mathbf{q}$ (calculate new **u** vector)
8. Check convergence on **u**:
   If converged go to 9, else go to 2
9. **X** loadings: $\mathbf{p} = \mathbf{X}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$
10. Regression: $b = \mathbf{u}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$ (where b is a scalar)
11. Calculate residual matrices: $\mathbf{E} = \mathbf{X}-\mathbf{t}\mathbf{p}^T$ and $\mathbf{F} = \mathbf{Y}-b\mathbf{t}\mathbf{q}^T$
12. To calculate the next set of latent vectors (if desired), replace **X** & **Y** by **E** & **F** and go to 1

The issues surrounding fraction of variance explained in PLS are similar to PCA, the main difference being that an amount of variance calculation can be done for both the **X** and **Y** matrices. Selection of the optimal number of latent vectors for PLS is also similar to PCA. It is common to give attention to the effectiveness of the PLS model at predicting **Y** values as the number of latent vectors is increased as an indicator of model "completeness".

PLS residuals are also similar to PCA residuals, with the extension to the **Y** matrix as well as the **X** matrix. The symbol **F** is used for the **Y** residual matrix, resulting in the expressions (for an optimum number of latent vectors *A*)

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E}$$

$$\mathbf{Y} = \sum_{i=1}^{A} \mathbf{u}_i \mathbf{q}_i^T + \mathbf{F}$$

While a residual matrix **E** exists for the PLS analysis of the **X** data, it is important to remember that this residual matrix is in nearly all cases different from the residual matrix remaining after PCA analysis. This is due to the fact that the PCA latent vectors are identified using only **X** information, while PLS latent vectors are developed with information from both **X** and **Y**. The only theoretical case in which the two sets of latent vectors will be the same (and, as a result, the residual matrices the same) is when the directions in which **X** most widely varies are exactly the same as the directions in **X** having the greatest covariance with **Y**.

2.3 Scaling of Data in PCA & PLS

Process data used in statistical analyses are commonly subject to scaling, and data used in PCA and PLS is no different. Kresta (1991) refers to the criticality of use of proper scaling and a good reference data set to the success of these techniques, and stresses that scaling should be performed such that the variance of each measurement reflects its relative importance. In an application of PCA to an extractive distillation column, he points out that if interrelationships between a particular group of variables is to be maintained, the same scaling factor should be used for the entire group.

Jackson (1991) refers to three standard scaling methods: no scaling of $\mathbf{X}$ (in which case the latent vectors are obtained from $\mathbf{X}^T\mathbf{X}/k$, the *product* or *second moment* matrix), scaling so that each variable has zero mean (the latent vectors being obtained from $\mathbf{X}^T\mathbf{X}/(k-1)$, the *covariance* matrix), and scaling the data to standard units -- zero mean and unit standard deviation (the latent vectors again being obtained from $\mathbf{X}^T\mathbf{X}/(k-1)$, which in this case is the *correlation* matrix.) Applications of PCA and PLS used with no scaling of the reference data set are most commonly found in chemistry applications (spectrophotometric interpretation, etc.) and will not be considered in detail here.

Jackson points out that people feel most comfortable when dealing with a covariance matrix, mainly because variables remain in their original units and procedures by which information can be inferred via PCA is considerably better developed for the covariance matrix than for any other scaling approach. However, there are circumstances in which using a covariance matrix will not work. When the original data variables are in different units, it can throw the covariance-based analysis off, because it is based on recognition of variability, and a measurement made in inches will "appear" to be more variable than a measurement made in yards. In some cases where the variables are in the same units, the variances may differ widely because they are based on their means (e.g. a measurement with a mean of one inch will most likely have a smaller variance than a measurement with a mean of 1000 inches.)

Wise (1991) discusses the relative merits of covariance versus correlation matrices as data sets after performing a PCA analysis of process data. He points out that similar latent vectors often, but not always, appear regardless of the scaling, although the apparent relative importance of the vector can shift based on the scaling. He observes based on his experience that covariance matrices have

yielded better results and recommends this method for initial investigations. Given variables of different units and no *a priori* knowledge of the variable importance, correlation matrices are recommended. In addition, he mentions user specified scaling when a fundamental understanding of the system under investigation exists, weighting the physically important variables higher than others.

Another type of scaling that is especially applicable to chemical engineering problems involves nonlinear transformations. Temperatures, compositions, pressures, and other process measurements often are related in nonlinear ways which can be approximated through exponentials, logarithms, and other nonlinear relationships. Mejdell (1990) uses scaling via logarithmic transformations on column temperatures and end product compositions to improve PLS estimator performance. Kresta (1992) also uses logarithmic transformations, but only on end product compositions. This approach can improve the performance of PCA and PLS by helping "overcome" the limitations of their linear nature.

2.4 Retaining Latent Vectors in PCA and PLS

Earlier, the subject of identifying the dominant latent vectors was briefly mentioned. The goal is to select the number of latent vectors that accurately represent a high percentage of the deterministic information in the data while including minimum stochastic information. Choosing a subset of the latent vectors is where the data compression in the PCA and PLS techniques comes from.

Jackson (1991) discusses a number of techniques used to determine the dominant latent vectors. Most of these tests focus on the latent vector's eigenvalues, which indicate the relative variance explained by each latent vector. Some techniques focus on the

percentage of variance explained by a given set of latent vectors, while others assert that latent vectors which contain stochastic information should all have similar eigenvalues. A graphical technique known as the SCREE test is available which calls for examination of the eigenvalues in order to find a "break" in the values, which is thought to indicate the boundary between deterministic and stochastic information.

Wold (1978) advances the use of cross-validation as a tool for choosing dominant latent vectors. This technique requires partitioning of the data into a control set and a test set. The control set is used to develop the PCA model, and the model is applied to the test set using different numbers of dominant latent vectors, yielding different errors between the predicted and actual test set. Minimization of these errors are the criteria for choosing the appropriate number of dominant vectors. Being data based methods, PCA and PLS can suffer from a pitfall common to such approaches -- the possibility of "overtraining" or "overfitting" such that noise in the data set is modeled as process information. Cross-validation specifically addresses this pitfall. The method has the nice properties of being conceptually simple and having little dependence on assumptions about the data or residuals.

Cross-validation has been adapted for use as a tool in choosing dominant latent vectors in a PLS regression. In PLS, the error of interest is the prediction error -- the difference between the actual and predicted **Y** values.

Wise (1991) uses another approach of choosing the number of dominant vectors. This approach is suited only to data sets which have observations, serially collected in time, which exhibit autocorrelation due to a frequency of observation that is relatively high compared to the process response time. The approach is based on the concept that deterministic process data from a continuous

process will exhibit autocorrelation, while stochastic information will not. As a result, he uses a graphical technique plotting the autocorrelation function of the model residuals for models using different numbers of dominant latent variables. The judgement of an "acceptable" level of autocorrelation in the residuals is subjective, but this method takes advantage of the unique autocorrelated nature of serially collected process data.

2.5 PCA and PLS Metrics

A number of metrics are associated with PCA monitoring. These metrics allow the user to take advantage of the PCA analysis to diagnose process faults and upsets quickly and effectively.

Once the latent vectors $\mathbf{p}_i$ have been selected, new data observations $\mathbf{x}^T_{new}$ can be projected onto these vectors. This yields scores, scalars which represent measures of variation from mean along the associated latent vector:

$$T_i = \mathbf{x}^T_{new}\mathbf{p}_i$$

If the input data is mean-centered, a score of 0 will correspond to an observation which is at its statistical mean with respect to the associated latent vector.

These score values can be plotted as a individual data point versus time, or sets of two scores can be plotted together in an x-y fashion, with multiple observations plotted on the same chart.

When the new observation $\mathbf{x}^T_{new}$ is projected onto the hyperplane spanned by the latent vectors, a PCA estimate of the sample is

formed (Wise 1991.) For a model in which A latent vectors have been retained, this estimate can be expressed as

$$\hat{\mathbf{x}}_{new}^{T} = \mathbf{T}_{new}^{T}\mathbf{P}_{A}^{T} = \mathbf{x}_{new}^{T}\mathbf{P}_{A}\mathbf{P}_{A}^{T}$$

where $\mathbf{T}_{new}$ is the vector of scores on the model $\mathbf{P}_A$ for the data observation $\mathbf{x}^T_{new}$.

The overall variance within the data space spanned by the latent vectors can be measured via the T^2 statistic. T^2 for a given sample is equal to the sum of the squares of the adjusted variance scores (normalized to unit variance) on each of the latent vectors in the model (Jackson 1980)

$$T^2 = \sum_{i=1}^{A}\left(\frac{T_i}{\lambda_i}\right)^2$$

Put another way, T^2 is a measure of how far the PCA estimate of the sample is from the multivariate mean of the data. As such, monitoring T^2 can provide an indication of abnormal process behavior or a shift in the multivariate mean.

Score and T^2 metrics are measures of information which falls within the PCA model or into the space spanned by the latent vectors. Residual metrics are indications of variability of the data outside the latent vector space. The residual vector of a given data observation, $\mathbf{r}_{new}$, can be calculated by subtracting the PCA estimate of the observation from the actual observation:

$$\mathbf{r}_{new}^{T} = \mathbf{x}_{new}^{T} - \hat{\mathbf{x}}_{new}^{T} = \mathbf{x}_{new}^{T}\left(\mathbf{I} - \mathbf{P}_{A}\mathbf{P}_{A}^{T}\right)$$

It is often useful to reduce the residual vector to a single scalar which represents how well the data observation is fit using the PCA

model. This scalar, Q, is defined as the magnitude of the residual vector:

$$Q = \left\| \mathbf{r}_{new} \right\| = \mathbf{r}_{new}^T \mathbf{r}_{new} = \mathbf{x}_{new}^T \left(\mathbf{I} - \mathbf{P}_A \mathbf{P}_A^T \right) \mathbf{x}_{new}$$

For a well-developed PCA model, the residuals should contain primarily uncorrelated information. As a result, it is possible to monitor Q using traditional SPC techniques, developing some "expected" standard deviation of residual variability and using this information to develop control limits. The topic of residual analysis is currently an active area of research.

Kresta (1991) sums up the subject of PCA monitoring metrics by pointing out two types of excursions that can be detected through these methods. Shifts in the scores indicate an abnormality of some kind, but one which is consistent with the basic process model. Shifts in the residuals indicate an abnormality which is not consistent with the basic process model, indicating a possible change in the nature of the process. Jackson and Mudholkar (1978) propose a procedure to use when analyzing score and residual data which illustrate how the two classes of metrics can be used in concert:

- test Q; if it is not significant, PCA model holds and T^2 should be tested
- if T^2 is not significant, procedure is complete and process is in statistical control
- if T^2 is significant, then the individual latent vectors should be tested to determine the nature of the disturbance
- if Q is significant, the residuals should be investigated immediately since the PCA model does not hold. When Q is significant, it is generally an indicator of bad data

In PLS, scores and residuals can be calculated for **X** or **Y** data, allowing monitoring of data based on the latent vectors developed in the PLS model. As a indicator of consistency of new observations with a base data set this method is less effective, since the latent vectors are not based solely on the variance of the **X** or **Y** base data, but also on how it covaries with its companion block. More often, however, the PLS metric of interest is the estimate of **Y** values based on **X** data, developed using the equations listed in section 2.2.

In PLS metrics, the emphasis is different. Although scores and loadings can be calculated using the PLS model, the primary metric of interest is prediction of **Y** quality variables using new observations of **X** process measurements and the PLS model (Kresta 1992). The **X** data is scaled to be consistent with the reference data set, and new **Y** values are calculated using

$$\hat{\mathbf{Y}} = \sum_{i=1}^{A} b_i \mathbf{t}_i \mathbf{q}_i^T$$

Alternatively, the loading vectors can be reduced to a linear regression model in terms of the **X** variables, yielding an expression for the predicted **Y** variables

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\alpha}_{PLS}$$

The regression vector $\hat{\alpha}_{PLS}$ is calculated as

$$\hat{\alpha}_{PLS} = \sum_{i=1}^{A} b_i \left(\prod_{j-1}^{i-1} \left(\mathbf{I} - \mathbf{w}_j \mathbf{p}_j^T \right) \right) \mathbf{w}_i \mathbf{q}_i^T$$

An advantage of the regression form is that the relative contribution of each **X** variable is expressed in terms of a single coefficient. This allows easy interpretation of the relative

importance of the **X** variables in prediction of **Y** values. Once the regression vector is developed, new predictions can be calculated as

$$\hat{\mathbf{y}} = \mathbf{x}^T \hat{\alpha}_{PLS}$$

2.6 Presentation/Use of PCA and PLS Metrics

The real benefit in PCA monitoring is in the use of the metrics as indicators of the process condition by operators, engineers, and managers. In order to make effective use of this information, it must be presented in a concise, easy to understand, and statistically correct fashion.

Kresta (1991) compares the monitoring of scores to SPC X bar charts, where variables are monitored for both positive and negative excursions from the mean via limits developed from observation of past operation. Monitoring of Q data is compared to R charts, where the value is bounded from below by zero in the best case (all process information described in latent vector space) and is compared to some upper limit based on historical data.

Wise and Ricker (1989) and Kresta (1991) point out that the development of multivariable SPC charts is very similar to univariate charts. Variables of interest (T^2, Q, scores) are identified, statistically appropriate charting techniques are selected, and a reference set of data is chosen to define the "normal" operating condition. Choice of a proper base data set is critical to the success of the application.

A unique advantage of PCA monitoring is the use of score plots on an x-y scatter diagram. Normally, T_1 and T_2 are plotted, although different combinations may be examined in situations where more than

two latent vectors are retained. A "control ellipse" can be calculated using a limit based on T^2 which, in effect, produces a 2-dimensional control limit. Jackson (1980) asserts that an x-y diagram displaying T_1 and T_2 contains the maximum amount of process information that can be displayed in two dimensions.

Wise (1991) and Vogel (1993) point out that many traditional SPC techniques are based on the assumption that the samples are uncorrelated in time. This offers a potential pitfall in the application of SPC techniques using PCA metrics with continuous process data, which often exhibits autocorrelation. If the data is autocorrelated, SPC run rules and control limit calculations are not necessarily valid. Residuals are a candidate for traditional SPC approaches if the process data does not exhibit autocorrelation and the number of latent vectors is selected appropriately.

In PLS, the metric of interest is the estimated quality variable. Ideally, it is presented to the operator through the distributed control system (DCS) along with process measurements. Trending and alarming for this variable can be implemented in a fashion similar to other process measurements. If the PLS estimate is determined to be invalid for a period of time (for example, due to the loss of a key measurement or movement of the process into a new mode of operation for which no base data set for PLS modeling is available), clear notification should be given.

2.7 Limitations of PCA/PLS Analysis

PCA and PLS have a number of limitations. Their reliance on linear mathematics is one -- the techniques will not work well with highly nonlinear data (Wise 1991), which is often the case in continuous chemical processes. However, they have been proven to be effective

for such processes if the process behavior is roughly linear over the operating range of interest, or if it can be linearized through some nonlinear transformation during scaling.

Need for data is another limitation. PCA is based on a model created solely from plant data -- extrapolation of the model to process conditions not included during its construction can yield incorrect and misleading results. It is important to start with a data set which includes a wide range of normal operating conditions. This requirement is directly in conflict with the first listed: a set of data representing all operating conditions of interest will likely not exhibit completely linear behavior, and a data set collected from a process run in a "roughly linear" region probably lacks the needed range of conditions for a generally applicable model. Engineering judgement must be used by the model builder to choose the correct combination of operating conditions needed for a useful implementation.

In PLS, both process measurement and quality variables are used as inputs to the regression algorithm. In most cases, these are collected from different sources: process measurement variables from a DCS or process data historian, and quality variables from a management information system (MIS) or laboratory data system. Synchronization of the two sets of data is crucial to the success of the PLS model. Often this is difficult: the exact time of the process measurement is well known, but the exact time when the sample is physically removed from the process to go to a laboratory for analysis is not well known. The plant personnel who remove such samples are responsible for a number of important duties, and it is not unusual for the actual time the sample is taken to vary 15-20 minutes from the "normal" time.

PCA and PLS are not inherently robust with respect to loss of a given input variable. If a field transmitter fails, a

residual-based monitoring technique will most likely provide notification, but the existing model is no longer valid and the monitoring system is rendered useless until the model is reconstructed without the missing variable or the transmitter is repaired. Wise (1991) suggests a construct for handling this scenario by exclusion of a given variable and removal of its contribution from the PCA metrics. In addition, loss of a crucial measurement renders a PLS model invalid, which most likely results in large errors in estimated measurements.

CHAPTER 3

PROCESS DESCRIPTION AND CONTROL SYSTEM DESIGN

3.1 Reactive Distillation and PCA/PLS

Chemical processing, in one simple description, can be thought of as three distinct tasks -- combination of raw materials, chemical reaction producing a desirable product, and separation of the product from undesirable by-products and excess raw materials. In most plants, these three tasks are performed by different pieces of processing equipment. However, there is one piece of equipment in which all three of these can occur -- the reactive distillation column.

The benefits of using reactive distillation are primarily realized in a lower capital cost since fewer pieces of equipment are needed, and the additional piping, pumps, and other items associated with these eliminated pieces of equipment contribute to capital savings. However, the reactive distillation column can present an operating challenge, since it is a "hybrid" operation. The issues which must be considered in analyzing a reactor column are unique. The control strategy applied to the column must consider both the extent of reaction of the raw materials and the extent of separation of the desired product from by-products and raw materials.

These unique needs make a reactive distillation column an attractive candidate for the application of multivariate statistical monitoring tools for process monitoring and inferential measurement. The way in which the process measurements covary when using this piece of equipment is not as intuitively obvious as for example, the

temperatures in a binary distillation column or the product flowrates, temperature, and pressure of an equilibrium CSTR. As a result, the covariance analysis power of PCA might serve to provide useful information about the state of the process measurements and some indication of the process health.

In the same way, the process information needed to control a reactive distillation column might be less traditional than the temperatures, pressures, and flows normally available on a continuous basis through a DCS. Product or by-product composition on a particular stage can be a key measurement needed for improved control of the column. Often, continuous on-line analysis techniques are unavailable or infeasible because of cost. Laboratory analysis, while cost-effective and fairly accurate, does not provide the frequency of process information needed for closed loop engineering process control (EPC) algorithms such as PID. PLS provides a tool by which the continuous measurements available in the DCS can be transformed, based on an initial set of laboratory analyses and corresponding process measurements, into an estimated composition that can be used for continuous closed loop control.

3.2 Description of Target Process

The reactive distillation column model used in this work has the same basic configuration as those addressed by Roat *et al.* (1986) and Simandl (1988). An industrial manifestation of this simulation example is discussed in Agreda and Partin (1984) and Agreda *et al.* (1990.) There are slight differences from Roat and Simandl in the base case steady state conditions due to a change in philosophy regarding the target underflow composition. Detailed model information, including the differences between this column and the two previous treatments, as well as a complete set of physical

properties, vapor-liquid equilibrium relationships, reaction kinetics and volumes, etc. is provided (Appendix A.)

In the target process (Figure 3.1), reactants B and D are introduced into a 25 tray distillation column. Reactant D enters on tray 21, while B is introduced on tray 7. The reaction proceeds via the following stoichiometry:

$$B + D \rightleftharpoons A + C$$

The reaction rate is first order in each component and is expressed in the conventional Arrehenius form. The reaction is liquid phase, and takes place in the accumulated liquid on each tray.

The catalyst stream E is introduced on tray 17. Having a considerably lower relative volatility than the other components, it travels down the column and exits with the bottoms. The desired product, A, leaves the column as distillate and the by-product, C, is removed with the bottoms. Feed and outlet flows from the column are listed (Table 3.1.)

3.3 Control Strategy Design

The reactor column can be thought of as performing three functions -- two separation operations (top and bottom sections of the column), and one reaction operation (middle section of the column). A robust control strategy must address all three objectives effectively. In this work, a conventional EPC approach using PID and ratio controllers will be employed. The production rate for the column will be controlled by setting the distillate flowrate.

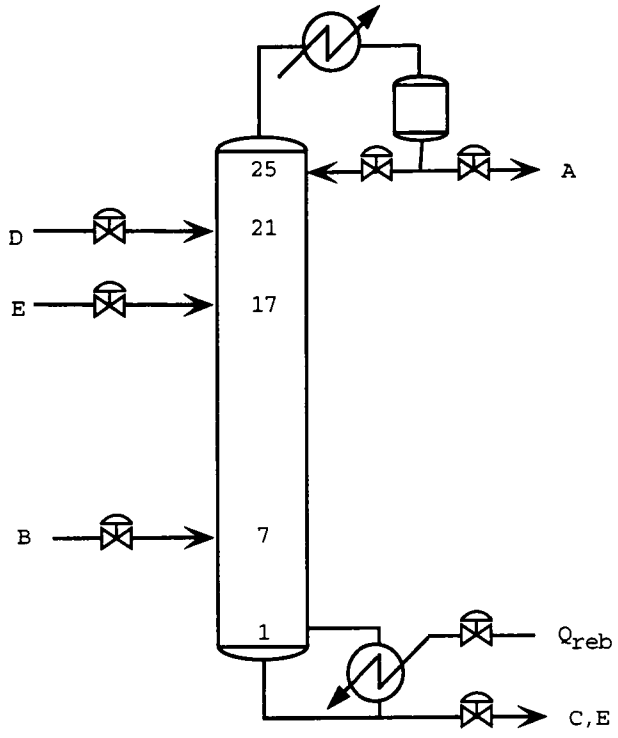


Figure 3.1. Process sketch of reactive distillation column.

Table 3.1. Base case material balance for reactive distillation column.

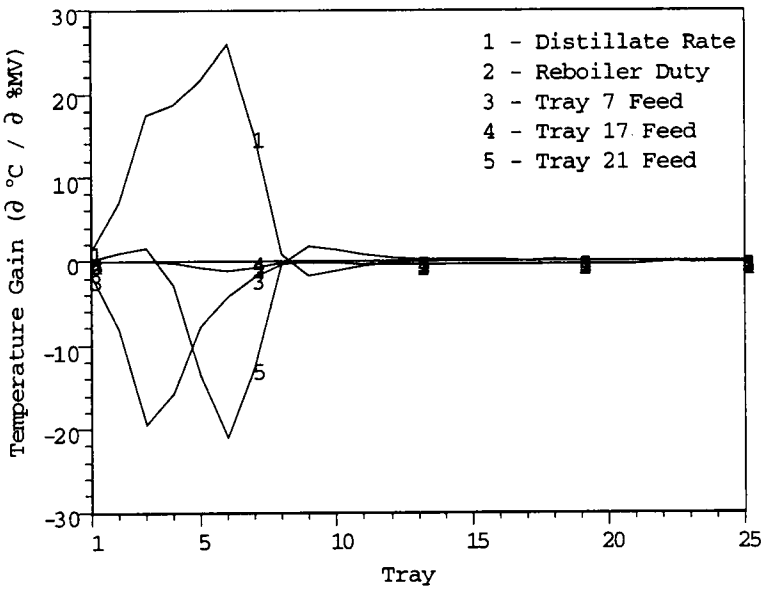
	Tray 7	Tray 17	Tray 21	Distillate	Bottoms
Flowrate (mol/hr)	49.86	0.50	48.90	50.00	49.26
Temperature (C)	50.00	25.00	25.00	50.00	50.00
Composition (mol %):					
A	0.00	0.00	0.01	97.71	0.00
B	99.99	0.00	0.00	1.71	0.31
C	0.01	0.00	0.00	0.55	98.63
D	0.00	0.00	99.99	0.03	0.05
E	0.00	100.00	0.00	0.00	1.01

The development of a control strategy for the column begins with examination of steady state gain information. This approach allows the user to assess which potential process measurements provide the most information about changes in column conditions, and thus serve as effective controlled variables.

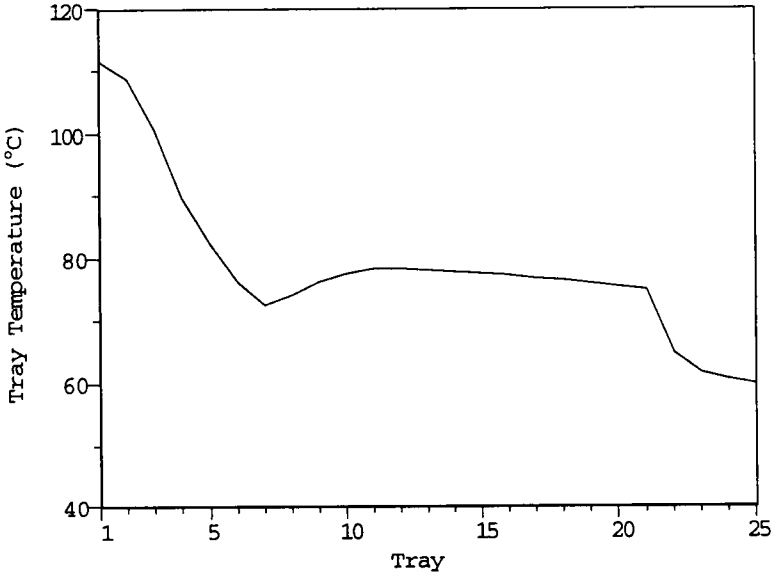
Examination of the steady state temperature gains (Figure 3.2a) indicates that the temperature information lies mostly in the bottom section of the column, around tray 5. This is consistent with the break in the temperature profile (Figure 3.2b) from trays 3-6.

Singular value decomposition (SVD) analysis is helpful in indicating the presence or absence of additional degrees of freedom for problems such as this (Downs 1982). An SVD analysis of the column temperatures is performed (Figure 3.3). The purpose of the SVD analysis is to determine if there are multiple degrees of freedom in the temperature information. Such degrees of freedom are signified by independent temperature gains in different areas of the column. Although the second temperature vector ($\mathbf{u}_2$) shows some weak independent information around trays 12-21, there is strong information corresponding to the first temperature vector around trays 4-6. The analysis indicates that while a second degree of temperature information exists in the column, it is not independent, and a second temperature controller based on this temperature information would likely interact with a temperature controller using the dominant temperature information.

In many columns, temperature alone can serve as a good indicator of changes in column conditions. However, in some columns, composition information is not adequately represented by temperature measurements. The composition gain plot for component D (Figure 3.4a) indicates a large amount of composition information around tray 8. The liquid composition profile plots (Figure 3.4b) show a



(a)



(b)

Figure 3.2. Column temperature information: (a) steady state temperature gains, (b) base case temperature profile.

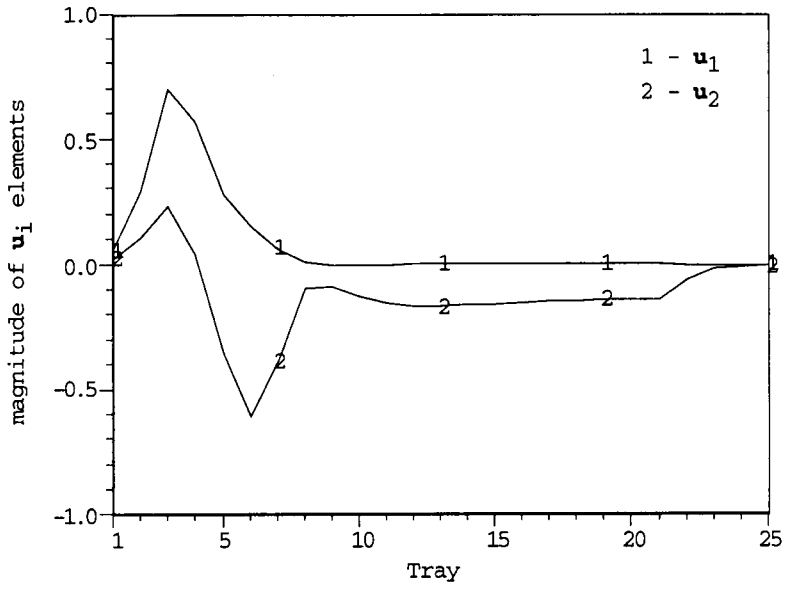
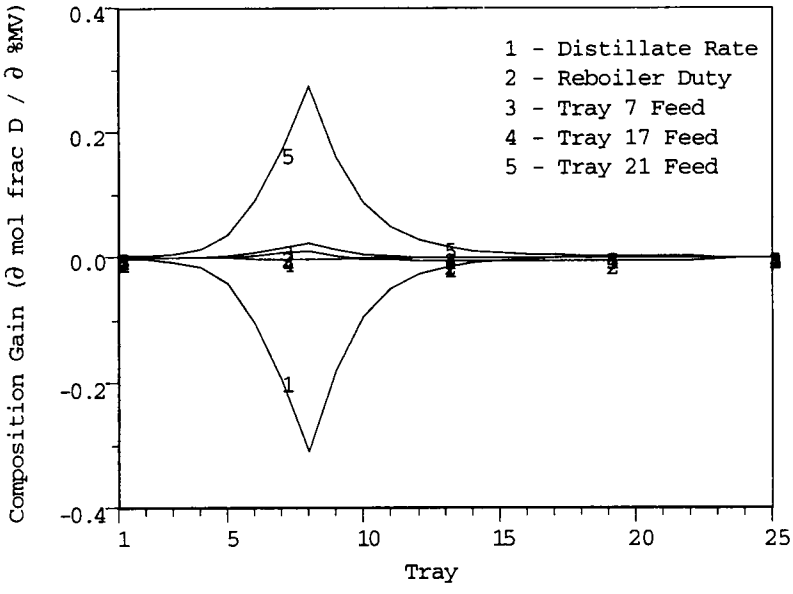
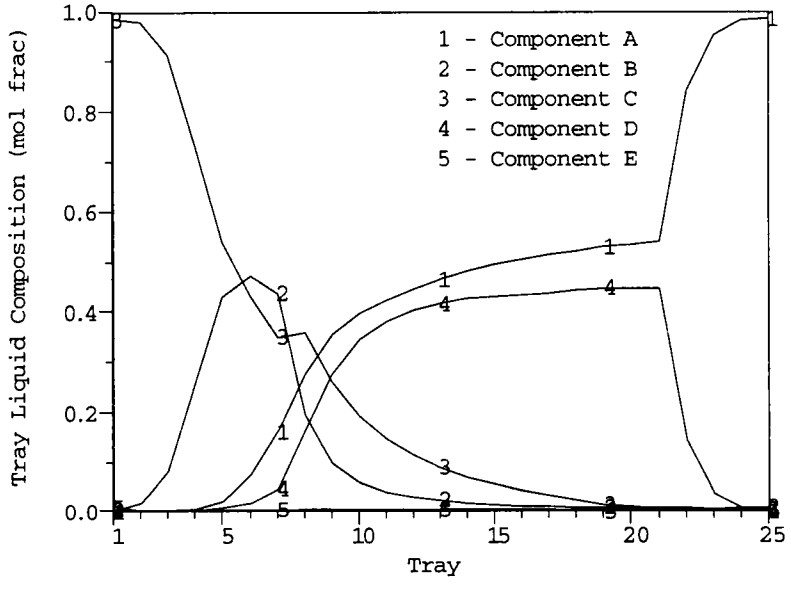


Figure 3.3. SVD analysis of column temperature information.



(a)



(b)

Figure 3.4. Column composition information: (a) steady state composition gains, (b) base case liquid composition profiles for all components.

distinct transition in compositions around tray 8 associated with the introduction of reactant B on tray 7.

The manipulated variables selected were feed to stage 7 and feed to stage 21. These two manipulated variables were chosen primarily because they represent the raw material flows to the column, and because it is critical to maintain the correct raw material feeds for a given distillate production rate. Given that the column production rate is set using the distillate, independent specification of either tray 7 or tray 21 feed would eventually lead to undesirable accumulation or depletion of a reactant, moving the column into a suboptimal or infeasible region of operation. Catalyst flow to tray 17 and reboiler duty were set constant, and the reflux and bottoms flows were used for local inventory control in the reflux drum and column base, respectively. The temperature of interest (tray 4) and the composition of interest (tray 8, component D) are subjected to a relative gain array (RGA) analysis to determine controller pairing (Table 3.2). The preferred controller pairings according to RGA analysis were tray 7 feed controlling tray 4 temperature and tray 21 feed controlling tray 8 D composition.

Dynamic studies are next used to further evaluate feasibility of the strategy and identify any controller interaction problems. There were no interaction problems between the two PID controllers. A feedforward loop between distillate flow and reboiler duty was added to reduce or add base heat to the column as the production rate is lowered or raised. The final strategy is depicted below (Figure 3.5.)

Table 3.2. RGA analysis for reactive distillation column.

	Tray 7 Feed	Tray 21 Feed
Tray 4 Temperature	1.017	-0.017
Tray 8 D Composition	-0.017	1.017

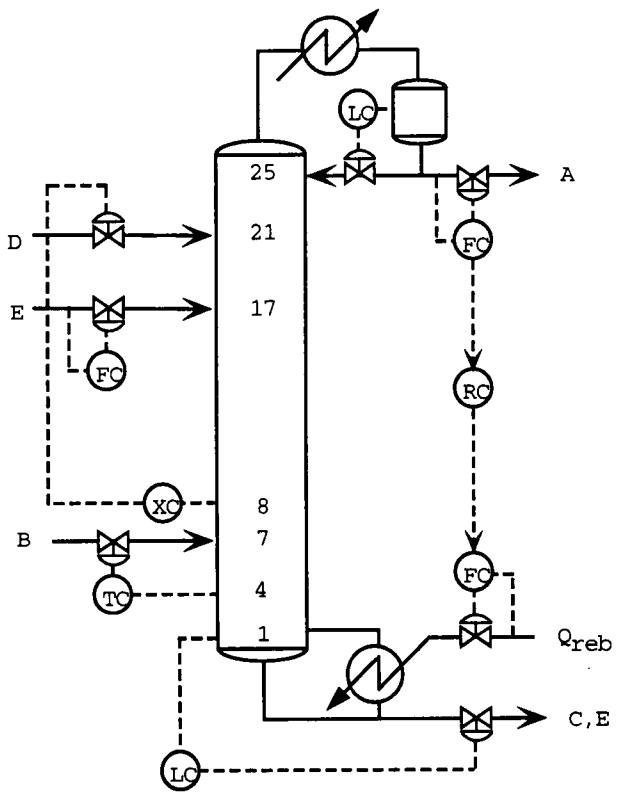


Figure 3.5. Process sketch of reactive distillation column, including control strategy.

3.4 Goals in using PCA and PLS on Target Process

In this strategy development, temperature and composition measurements have been discussed interchangeably. Unfortunately, the cost of providing continuous composition measurements is much higher than the cost of providing continuous temperature measurements. Therefore, many strategies such as this are bypassed in favor of a less effective strategy which makes use of more inexpensive and conventional measurements.

The primary goal of applying PCA and PLS technology to this problem is to create a PLS-based continuous estimate of the composition of component D on tray 8 for use in closed loop control. This estimate will be generated by subjecting the column simulation to perturbations and using the process data set generated through these tests to regress a PLS model.

Composition estimation using PLS is a potential lower cost alternative to continuous analysis. However, the use of a PLS-based estimate for the composition control loop requires careful consideration, because it will most likely be less accurate than a conventional measurement or analytical technique due to measurement noise and error and plant/model mismatch. Characterization of the estimate error is important and negative effects of the error can be mitigated by selection of a composition for control which, when held within the expected error band, produces good control results. In the case of this column, success in control can be assessed by examining the amount of excess raw material B which leaves with the distillate stream (the product stream). Selection of a composition measurement in the middle of the column with a large steady state gain helps insure that control within a reasonable error band will result in consistent product properties.

As shown below (Figure 3.6), selection of the component D composition on tray 8 yields favorable conditions for estimator-based control: at the base case condition, error in the estimate of +/- 5% results in product impurity variation of only +/- 0.12%.

In addition to the PLS-based composition estimator, a PCA model of the column will be regressed in order to assess the nature of the measurement data coming from the process. This information can be used in concert with the PLS model to express a "confidence" in the composition estimate.

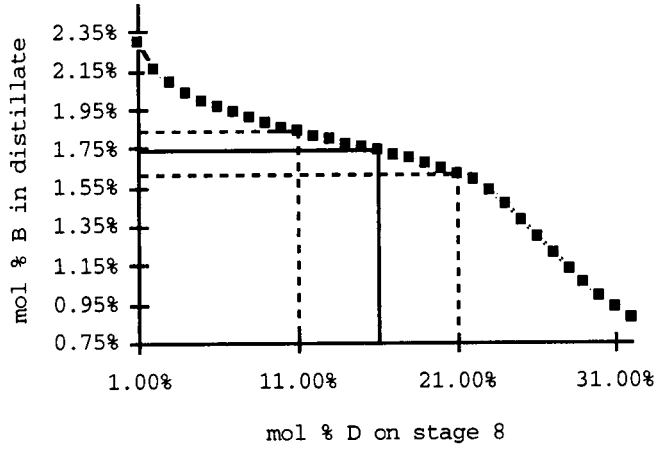


Figure 3.6. Variation of distillate B composition with changes in tray 8 D composition.

CHAPTER 4

COMPOSITION ESTIMATION AND CONTROL USING PARTIAL LEAST SQUARES (PLS)

4.1 Process Measurements and Estimated Composition

The goal of this section of the work is to provide a reliable estimate of the component D composition on tray 8 for closed loop control. Since the work is based on a simulation, a number of process measurements are available for use -- however, only measurements typically available on a continuous basis using conventional instrumentation (temperatures, flows, and pressures) were considered in the development of the PLS model. For the estimator, a data set comprised of all 25 tray temperatures, tray 7 feed, tray 21 feed, distillate flow, reflux flow, bottoms flow, and reboiler duty was chosen. Column pressure and catalyst feed rate were assumed to be constant.

4.2 Baseline Data Set

One of the challenges in developing a PLS model is obtaining a representative data set from which to build the model. The set is comprised of an **X** matrix containing a number of separate observations of a set of input measurements, and a **Y** matrix containing corresponding observations of the desired prediction measurements. The advantage of PLS comes from its ability to interpolate based on the data set with which the model is built;

therefore, it is important to have a data set with observations across the operating range of interest.

Two basic approaches to gathering data are possible. One approach is to look at historical plant data and select sections of operation over which the plant is running "normally". This approach is similar to the one used in identifying data sets on which to base control limits for univariate SPC charts, and lends itself well to situations in which a large number of frequent process measurements and laboratory analysis data are available. The second approach is to conduct a series of process experiments, manipulating column inputs over a range to insure coverage of all combinations of operating parameters. This approach is more appropriate for a pilot plant operation or a simulation than for a full scale plant producing goods for sale to customers, due to the fact that varying operating conditions in a pilot plant or simulation usually carries minimal or no penalties due to off-specification product or upset processes. For this work, the second approach is used.

Mejdell (1991) suggests an experimental design for development of the process data set, using a set of 32 observations to generate an estimator for a simulated distillation column, and 17 observations to generate an estimator for a pilot plant distillation column. Instead of using a true experimental design, Kresta (1992) sets a range for each input variable and varies each independently and in random combination (though the details of this approach are not recorded), using a set of 76 observations to generate a PLS estimator for a benzene-toluene-xylene distillation column; a set of 80 observations are used for construction of a PLS model for a methanol-acetone-water column. Kresta points out that for best results, the closed loop control strategy that will eventually be implemented using the estimator should be used in the development of the data set because feedback control will alter the covariance of the process measurements. The distillation simulation used in this

work allows implementation of feedback control using actual compositions, and such a loop is configured for development of the base data set. Kresta also points out that in the case of gathering data in the presence of a closed loop control strategy, the loop controller setpoints should be perturbed as process inputs instead of the valves associated with the controllers. This is especially important in the case of the composition controller, since it is desirable for the composition estimate to hold over a wide range of conditions.

Four process inputs were selected for perturbation to exercise the simulation over a representative operating range. These inputs were distillate flowrate, reboiler duty, tray 4 temperature setpoint, and tray 8 composition setpoint (Table 4.1.) These were varied individually (observations 1-21), and in combination using two 2^k factorial designs: one with a medium level of perturbation (observations 22-37) and one with a high level of perturbation (observations 38-53.) In the individual variations, care was taken to vary the tray 8 composition setpoint over a wide range of operating conditions, since this is the prediction variable of interest. After each perturbation, the simulation was converged to steady state and a snapshot of the process measurements and tray 8 composition was taken. This resulted in a data set containing 53 observations with 31 process measurements (**X** block) and 1 actual composition (**Y** block) per observation. A full record of the base data set is provided (Appendix B.)

4.3 Development of PLS Model

The PLS model was developed using the MATLAB matrix mathematics software package. The PLS Toolbox by Wise (1994), a set of MATLAB routines developed for PLS and other types of regression, was used.

Table 4.1. Perturbations to model used to generate data set.

Obs	Tray 4 Temp (°C)	Dist Flow (mol/h)	Reboiler Duty (BTU/hr)	Tray 8 % D (mol %)	Obs	Tray 4 Temp (°C)	Dist Flow (mol/h)	Reboiler Duty (BTU/hr)	Tray 8 % D (mol %)
1	90	50	1.80E+06	16.0%	28	85	45	2.00E+06	20.0%
2	90	40	1.80E+06	16.0%	29	95	45	2.00E+06	20.0%
3	90	60	1.80E+06	16.0%	30	85	55	1.60E+06	12.0%
4	90	55	1.80E+06	16.0%	31	95	55	1.60E+06	12.0%
5	90	45	1.80E+06	16.0%	32	85	55	1.60E+06	20.0%
6	90	50	2.05E+06	16.0%	33	95	55	1.60E+06	20.0%
7	90	50	2.20E+06	16.0%	34	85	55	2.00E+06	12.0%
8	90	50	1.60E+06	16.0%	35	95	55	2.00E+06	12.0%
9	90	50	1.40E+06	16.0%	36	85	55	2.00E+06	20.0%
10	90	50	1.80E+06	40.0%	37	95	55	2.00E+06	20.0%
11	90	50	1.80E+06	30.0%	38	80	40	1.40E+06	4.0%
12	90	50	1.80E+06	25.0%	39	100	40	1.40E+06	4.0%
13	90	50	1.80E+06	20.0%	40	80	40	1.40E+06	28.0%
14	90	50	1.80E+06	10.0%	41	100	40	1.40E+06	28.0%
15	90	50	1.80E+06	4.0%	42	80	40	2.20E+06	4.0%
16	90	50	1.80E+06	1.0%	43	100	40	2.20E+06	4.0%
17	90	50	1.80E+06	0.1%	44	80	40	2.20E+06	28.0%
18	85	50	1.80E+06	16.0%	45	100	40	2.20E+06	28.0%
19	80	50	1.80E+06	16.0%	46	80	60	1.40E+06	4.0%
20	95	50	1.80E+06	16.0%	47	100	60	1.40E+06	4.0%
21	100	50	1.80E+06	16.0%	48	80	60	1.40E+06	28.0%
22	85	45	1.60E+06	12.0%	49	100	60	1.40E+06	28.0%
23	95	45	1.60E+06	12.0%	50	80	60	2.20E+06	4.0%
24	85	45	1.60E+06	20.0%	51	100	60	2.20E+06	4.0%
25	95	45	1.60E+06	20.0%	52	80	60	2.20E+06	28.0%
26	85	45	2.00E+06	12.0%	53	100	60	2.20E+06	28.0%
27	95	45	2.00E+06	12.0%					

In this regression, the data were mean centered and scaled to unit variance to avoid artificially weighting variables with larger magnitudes. Mejdell and Kresta both recommend additional scaling of the **Y** values with a logarithm to linearize the end purity composition effects on distillation operations. The ability of the model to predict compositions from the base data set was used to assess different scaling approaches. The logarithm function was tested on the composition, and no improvement in prediction ability was observed. It is assumed that since the composition of interest is not an end composition, as with Mejdell and Kresta's work, the changes in composition with changing conditions have a different characteristic behavior which is not enhanced by a logarithmic transformation. An alternative function, tanh, was tested for improved prediction. Again, no prediction improvement was observed, so the **Y** values were mean centered and unit variance scaled as well.

The PLS regression results in a number of latent variables, each with an associated percent of variation explained for the **X** block and **Y** block (Table 4.2.) The next step is to select the number of latent variables to retain. Cross-validation (Wold 1978), a method in which a section of the data is held back and used as a test set, was used to select the correct number of latent vectors. The goal of cross-validation is to avoid "overfitting" the data, insuring that latent vectors are added only as long as they explain general process information and not information specific to a particular block of data.

Output from the cross-validation algorithm is a plot comparing PRESS (the sum of squares of the prediction error) versus the number of latent vectors included (see Figure 4.1.) In this case, the PRESS statistic goes through a minimum at eight latent vectors. This number of latent vectors was used for formation of the PLS regression vector.

Table 4.2. PLS regression results, showing individual percent of variance explained for each latent vector and total percent variance explained for eight latent vector PLS model.

Latent Vector	**X** Block % Variance Explained	**Y** Block % Variance Explained
1	20.5	75.0
2	45.8	14.5
3	13.0	4.1
4	5.8	2.4
5	5.1	1.2
6	5.4	0.2
7	1.1	0.3
8	1.8	0.1
Total % Explained	98.4	97.8

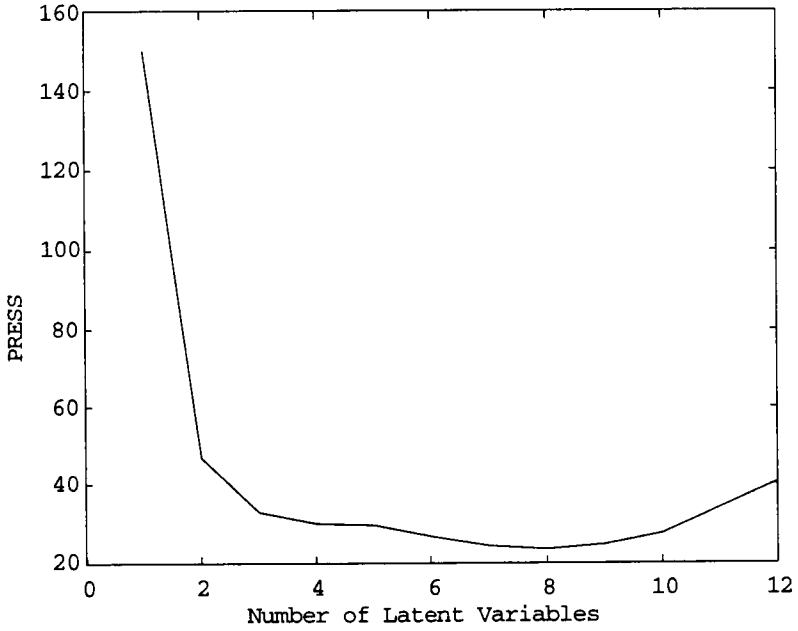


Figure 4.1. Cross-validation PRESS plot showing minimum PRESS using eight latent variables.

Once the number of latent vectors is chosen, a regression vector can be formed which leads directly from a set of **X** block inputs to a **Y** block estimation (Figure 4.2.) This regression vector is applied to the scaled **X** data, and produces a **Y** estimate which must then be rescaled to physically significant units. The estimated value is calculated as

$$\hat{\mathbf{y}} = \mathbf{x}^T \hat{\alpha}_{PLS}$$

The regression vector is a good measure of the relative importance of each input variable in the composition estimate. Variables 1-11, corresponding to temperatures on trays 1-11, have a significant contribution to the estimate. Conceptually this makes sense since the column control analysis (Chapter 3) indicated that the significant temperature and composition gain information is located in this region. The variables associated with column feeds, reflux flow, reboiler duty, and distillate and bottoms rates (variables 26-31) also contribute significantly to the estimator.

It is revealing to contrast the regression vector generated via PLS with a regression vector generated with the same data using an alternative regression technique, multiple linear regression (MLR) (Figure 4.3.) MLR is a technique which uses all the information in the **X** block, both deterministic and stochastic, to form the **Y** estimate. Wise (1991) points out that MLR yields the same results as PLS with all latent vectors included; thus, a comparison of the two regression vectors is a clear indicator of the effect of limiting the PLS model to a set of significant latent vectors.

The MLR regression does a slightly better job of fitting the base data set than the PLS regression. The reason for this is overfitting of the data which leads to better results for the base

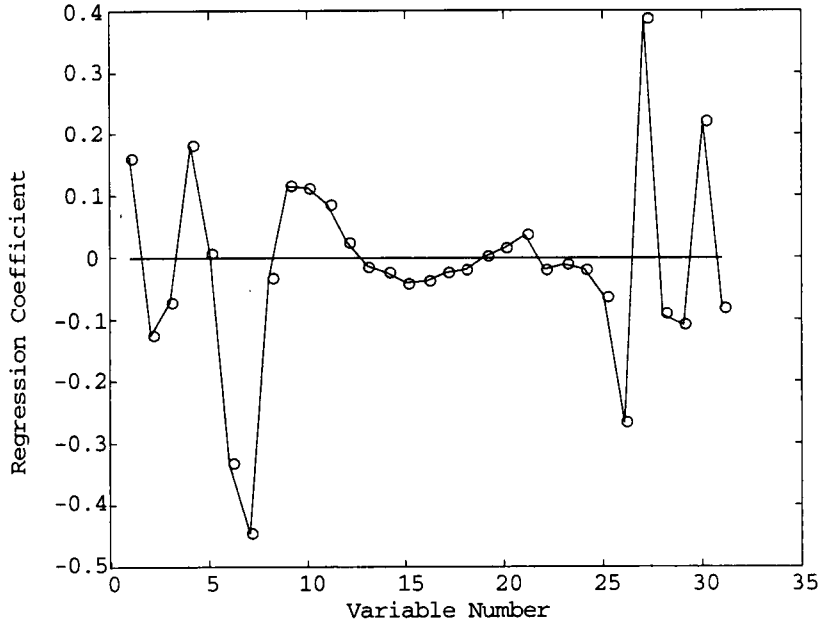


Figure 4.2. Regression vector for PLS estimator.

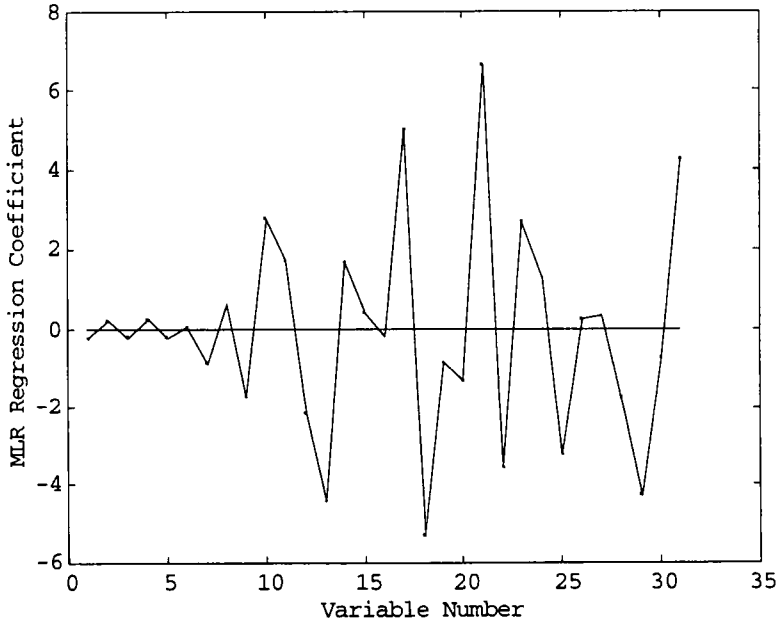


Figure 4.3. Regression vector for MLR regression.

data set but almost always poorer results for on-line estimation. In the MLR regression vector, weight on the bottom temperature information which is key to the PLS predictor is insignificant compared to weight on temperatures in the middle and top of the column. Also, the regression coefficients are an order of magnitude larger for the MLR regression. The regression coefficients are erratic and vary wildly from tray to tray in the temperature region, revealing no general trends in the value of temperature information through the column. These differences are indicators of the value of PLS over MLR for estimation of on-line properties using process data.

Once the PLS regression model is developed, an initial assessment of estimator accuracy can be made by applying the model to the base data set and recording the residuals (Figure 4.4.) This indicates that the estimator does a fairly good job of composition estimation, staying within +/- 3% of the actual value in all but one occurrence. The standard deviation of the error for this data set is 1.4%.

An alternative way to view the prediction information and assess estimator performance is to compare the estimated value to the actual value with the actual values sorted by increasing magnitude (Figure 4.5.) This allows the model builder to assess not only the overall performance, but also how the estimator accuracy varies as the actual composition varies. As might be expected, the estimator seems to be most accurate around the nominal value of 16%, and has the greatest prediction error near the extremes of the composition measurement span.

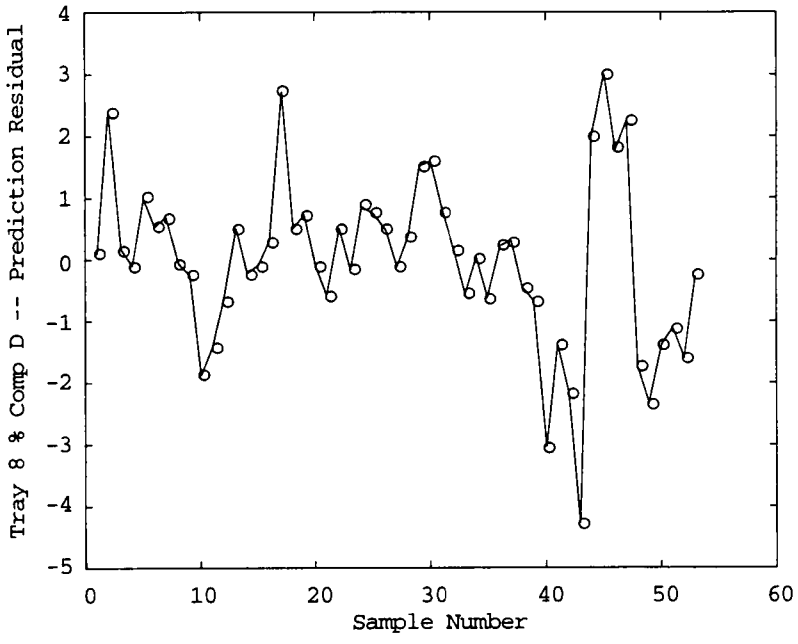


Figure 4.4. Residuals resulting from application of PLS regression to base data set.

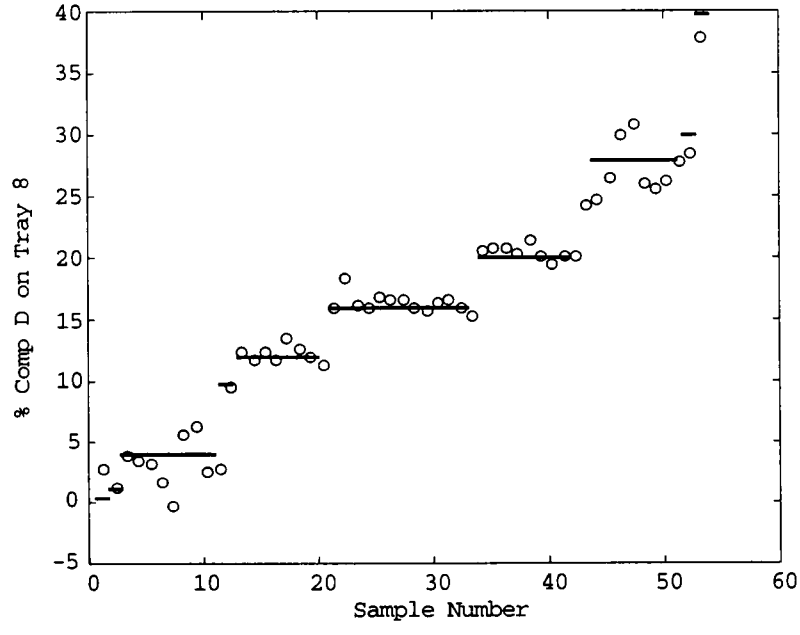


Figure 4.5. Actual (-) and estimated (o) compositions, sorted in ascending order.

4.4 Use of Composition Estimate in Closed Loop Control

Once a PLS regression of the base data set has been performed and the results of the regression are verified, application of the estimator in closed loop control is tested.

Two complications are introduced when the estimator is applied closed loop. The first is the dynamic nature of control. The PLS model discussed here was developed using only steady state process data. As the process is subjected to disturbances, it will go through dynamic responses not accounted for in the PLS model. This will undoubtedly result in estimation error during the period in which the process is not at or near steady state.

The second is the uncertainty of the nature of process disturbances. The PLS model was constructed using known disturbances likely to occur during process operation, such as production rate changes and controller setpoint changes. In addition to these, other unmeasured or infrequent disturbances will affect the way the process operates and thus the way the process measurements vary and covary. The PLS model is only capable of interpreting process information found in the base data set. As these unmeasured or infrequent disturbances enter the process, the PLS model has no information on how to incorporate these changes into the composition estimate. Again, this will almost certainly lead to estimator error.

In the following tests, a series of process changes are made on the reactor column. An EPC control system is implemented on the rigorous distillation simulation as described in Figure 3.5. The tray 8 composition controller is implemented using the PLS estimate. Process responses are presented for both the tray 8 composition and the product quality variable, presented here as the % product impurity in the distillate (nominally 1.72% B.) Four examples

(Cases 1-4) are presented with disturbances included in the base data set, and six (Cases 5-10) with unmeasured or infrequent disturbances not included in the data set.

In Case 1, the column production rate is reduced from 50 mol/hr to 40 mol/hr (Figure 4.6.) [As with all following graphs, the composition estimate for tray 8 is represented as a dashed line, and the actual compositions for tray 8 and tray 25 as solid lines.] For this disturbance, the estimator does a good job of dynamically tracking the actual composition and returning the column to the target tray 8 composition. As a result, the product impurity composition is controlled and returns to near its original value.

For Case 2, the tray 4 temperature controller setpoint is raised from 90°C to 100°C (Figure 4.7.) The dynamic responses of the actual and estimated compositions have similar shapes, but the estimator makes a large dip at t=20 min. This may be related to the heavy reliance on temperature information low in the column for the estimate -- for this particular disturbance, the temperatures in the bottom of the column are drastically affected. Even though the estimated composition ends up about 1.5% lower than the actual value, the product impurity composition returns to within 0.02% of its target value after the process disturbance and resulting "bump."

For Case 3, the tray 8 estimated composition controller setpoint is raised from 16% D to 21% D (Figure 4.8.) The dynamic responses of the actual and estimated compositions are very similar in shape and respond slowly compared to Cases 1 and 2. This response speed is due to the conservative tuning of the composition controller. The estimated composition ends up with around 1.1% error. The response of the top product impurity is also slow and gentle, with an offset of around 0.15% from target.

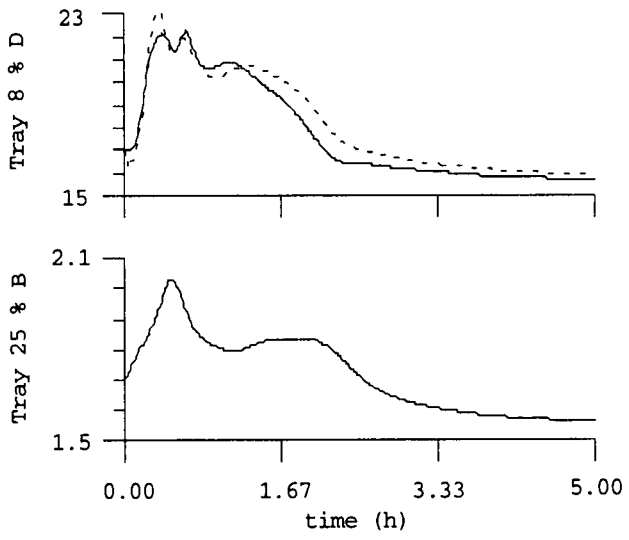


Figure 4.6. Process responses for Case 1 (production rate reduced from 50 mol/hr to 40 mol/hr)

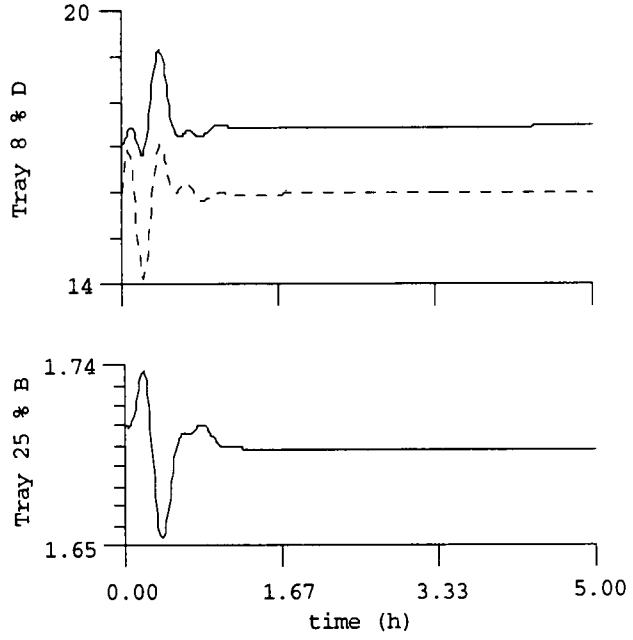


Figure 4.7. Process responses for Case 2 (tray 4 temperature setpoint changed from 90 C to 100 C)

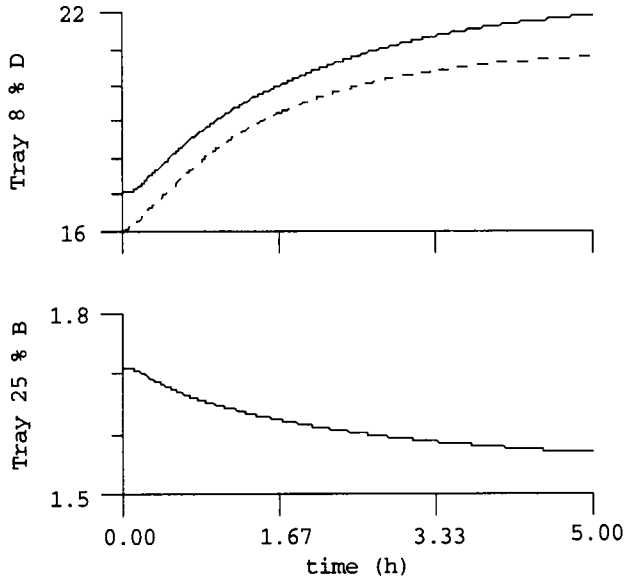


Figure 4.8. Process responses for Case 3 (tray 8 estimated composition controller setpoint changed from 16% D to 21% D)

In Case 4, the column reboiler duty is increased from 1.80E+06 BTU/hr to 2.05E+06 BTU/hr (Figure 4.9.) For this test, the feedforward controller on the reboiler duty is disabled, and the reboiler duty is manipulated directly. Reboiler duty obviously has a drastic effect on the column, as the tray 8 composition drops from 16% to 8% in less than 15 minutes. Dynamically, the estimator accurately tracks the actual composition and ultimately yields an error of approximately 1.3%. Even so, the top impurity composition moves from 1.72% to around 2.4%. This indicates that reboiler duty is a strong handle for influencing top product composition in this control strategy. This case makes the point that controlling the tray 4 temperature and tray 8 composition to target does not guarantee top product purity for all disturbances using this control strategy.

It is important to note that this result is not an artifact of disabling the feedforward controller and violating Kresta's guidelines on consistency between the control strategy used with the

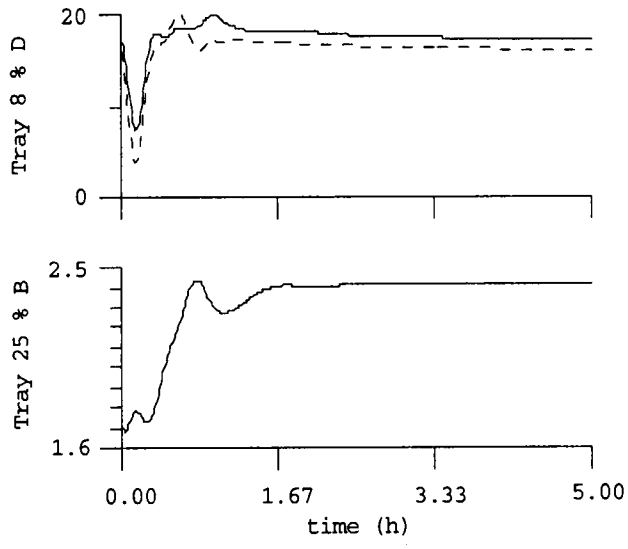


Figure 4.9. Process responses for Case 4 (reboiler duty increased from 1.80E+6 BTU/hr to 2.05E+6 BTU/hr)

base data set and with the on-line estimator. The base data set was constructed using the reboiler duty as a manipulated variable. Addition of controllers other than feedback controllers (such as feedforward, lead/lag, or ratio controllers) does not affect the covariance of the process measurements and as a result, the resulting PLS regression model. As proof, note that the actual and estimated composition for this case track well -- the best indicator of whether the PLS model is valid or not.

For Case 5, the first of the unmeasured or infrequent disturbances, the composition of the tray 7 feed is changed from 0.01% C to 15% C (Figure 4.10.) The dynamic responses of the actual and estimated compositions are similar in shape, but the estimated composition yields approximately 3.5% error. This is due mainly to the increase in tray 7 feed to supply the needed B to maintain production rate -- the feed composition change resulted in less introduction of B for the same feed flowrate. The top product impurity remains near the same level -- an indicator of either the robustness of the control

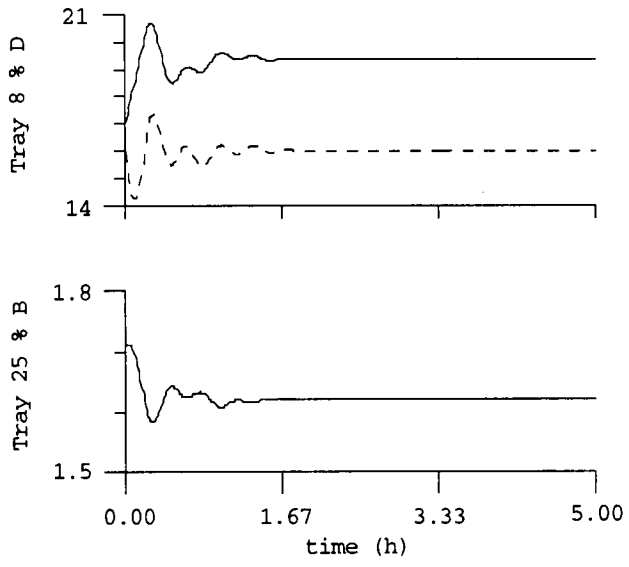


Figure 4.10. Process responses for Case 5 (tray 7 feed composition changed from 0.01% C to 15% C)

strategy or the insensitivity of product impurity to tray 7 feed composition.

For Case 6, the composition of the tray 21 feed is changed from 0% C to 15% C (Figure 4.11.) Again, the dynamic responses of the actual and estimated composition are similar, and the estimated composition yields about 2.5% error. The top product impurity increases drastically, going from 1.72% to around 3.5%. This indicates that top purity is much more sensitive to tray 21 feed composition changes than tray 7 changes, since the estimated composition for Case 6 is actually better than Case 5, but the top impurity composition is much higher for Case 6 than Case 5.

For Case 7, the top column pressure is increased from 850 torr to 1000 torr (Figure 4.12.) The dynamic responses of the actual and estimated composition are again similar in shape, with a higher final prediction error of about 5.5%. The top purity increases around 0.05%. Increasing the column pressure will raise all column temperatures and thus induce error into the estimate information

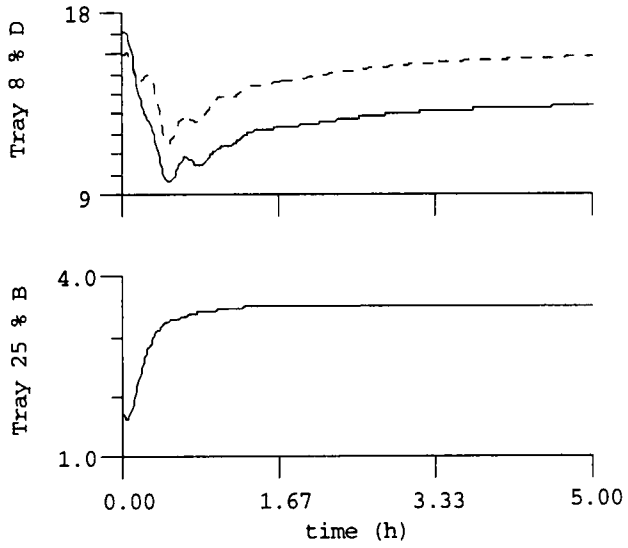


Figure 4.11. Process responses for Case 6 (tray 21 feed composition changed from 0% C to 15% C)

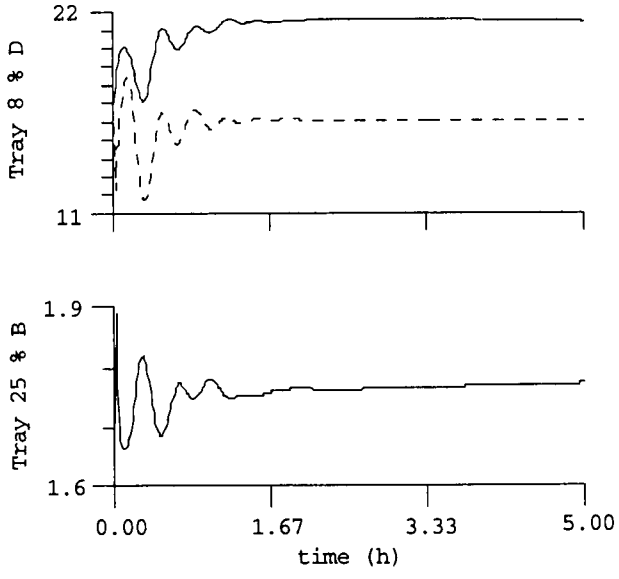


Figure 4.12. Process responses for Case 7 (column top pressure changed from 850 torr to 1000 torr)

coming from the tray temperatures. However, the higher pressure also affects the vapor-liquid equilibria inside the column, and these two effects likely counteract each other to produce only a slight change in top impurity composition.

For Case 8, the tray 17 catalyst feed flow is reduced from 0.5 mol/hr to 0.05 mol/hr (Figure 4.13.) Since the catalyst composition is included in the kinetic model, this has the effect of drastically slowing the reaction. This results in gross error in the estimator -- more than 9% low. The product impurity moves to around 7% as well. When the catalyst flow is reduced, the reactive distillation moves from a mode of separating products to separating unreacted raw materials. Obviously, this will completely invalidate the PLS model, since the base data set contained only data from the reaction region of operation. This invalidation is evidenced by the failure of the estimator in this case.

Cases 9 and 10 are demonstrations of the effect that measurement failure has on the estimator. In Case 9, a temperature which has a high weight in the PLS regression vector (tray 7) is "frozen" or fails to its current value. A production rate increase from 50 mol/hr to 60 mol/hr is then made (Figure 4.14.) The estimator quickly fails, indicating 16% D when in reality all D has reacted or drained off the tray. The top product impurity quickly rises to over 5% B. This indicates that the loss of a key measurement can quickly incapacitate the PLS estimator.

For Case 10, a temperature with a lower weight in the PLS regression vector (tray 14) is "frozen", and the same production rate increase is made (Figure 4.15.) Here, the estimator performs well, maintaining less than 3% estimation error and holding top product impurity composition within 0.15% of target. As expected, loss of

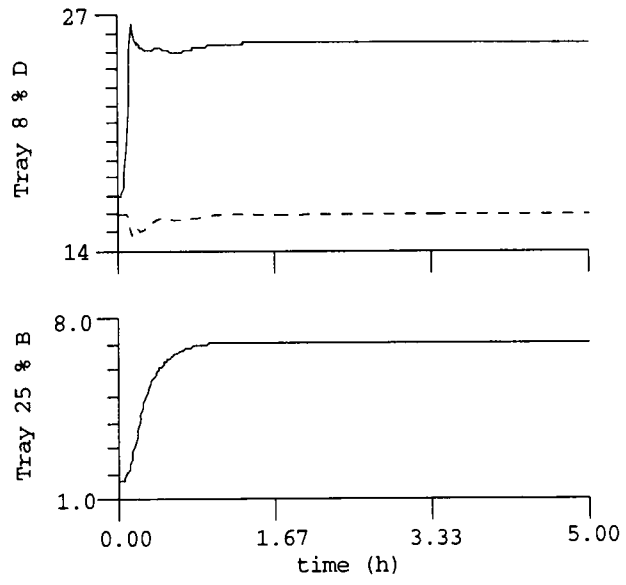


Figure 4.13. Process responses for Case 8 (tray 17 catalyst feed reduced from 0.50 mol/hr to 0.05 mol/hr)

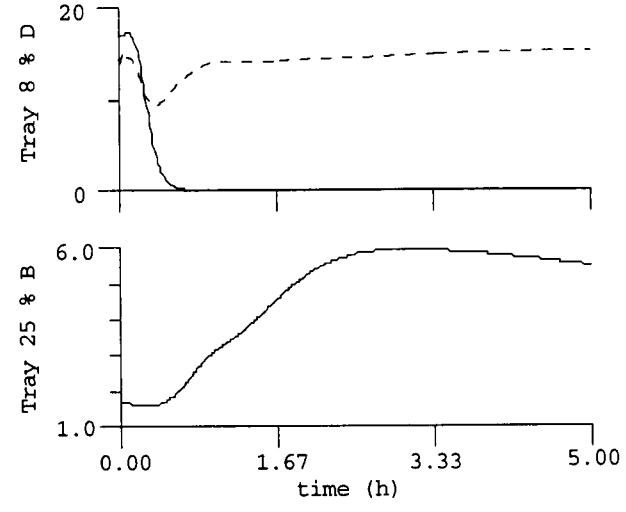


Figure 4.14. Process responses for Case 9 (production rate increase from 50 mol/hr to 60 mol/hr, and temperature measurement on tray 7 fails to last good value)

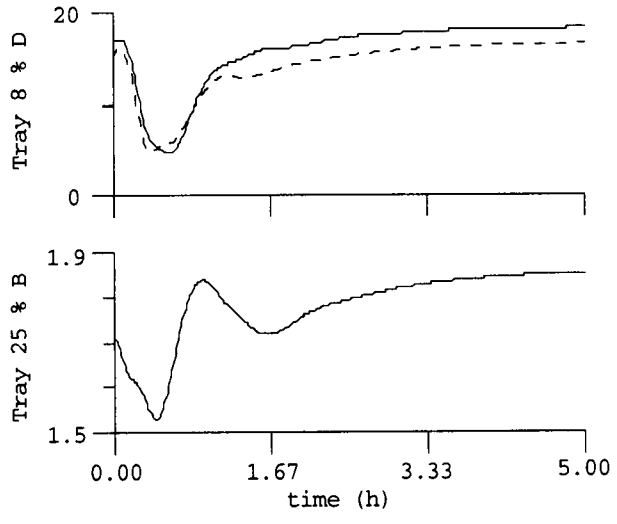


Figure 4.15. Process responses for Case 10 (production rate increase from 50 mol/hr to 60 mol/hr, and temperature measurement on tray 14 fails to last good value)

the less critical temperature measurement had much less of an effect on the PLS estimator than the loss of the heavily weighted temperature.

4.5 Discussion of Results

Overall, the PLS estimator does a good job of providing a reliable estimate of tray 8 D composition for closed loop control for disturbances included in the base case data set (Table 4.3.) In cases 1-4, the steady state value of the estimator is within 2% of the actual composition. The estimate-based controller also does a good job of controlling distillate impurity for Cases 1-3, resulting in top impurities within 0.15% of the target value. Case 4 demonstrated that accurate prediction does not always guarantee optimum impurity control, as the reboiler duty influences top

Table 4.3. PLS estimator error and top impurity concentration using both estimator-based control and composition measurement control for disturbance cases.

Case #	PLS estimator error @ t=5 hr	Tray 25 % B using estimator @ t=5 hr	Tray 25 % B using composition @ t=5 hr
1	0.42%	1.56%	1.55%
2	-1.47	1.70	1.73
3	-1.14	1.57	1.61
4	-1.28	2.43	2.48
5	-3.38	1.62	1.70
6	2.45	3.50	3.48
7	-5.46	1.77	1.92
8	-9.50	7.11	8.97
9	15.48	5.49	1.92
10	-2.57	1.86	1.92

impurity composition even with the estimated composition controlled to target. This is a result of the control strategy design, and serves as additional justification for the feedforward controller setting the reboiler duty discussed in Section 3.3.

Cases 5-10 were developed using disturbances not included in the base data set, and the estimator was less successful. Almost all of these cases have significant estimator error, ranging as high as 16% for Case 9. Some cases maintain acceptable top impurity control even with estimator error (Cases 5, 7, & 10). Others (Cases 6 & 8) exhibit estimator error, but result in top impurity values comparable to those achieved using an actual composition measurement for control. Case 9 is a clear case in which estimator error results in high product impurity and control using composition measurement produces good results.

The steady state nature of the data with which the PLS model was built contributes to some of the estimator error, especially during the initial transient response. Case 2 is a good example of a

process dynamic resulting in high estimator error for a short period of time. Generally the estimator does well in process dynamics, but these results show that it is least dependable during upsets.

As discussed earlier, the PLS estimator is incapable of accurate extrapolation outside its base data set. Process dynamics and unmeasured or infrequent disturbances are two causes of inconsistency with the base data set, and there does not yet exist a technology for minute-to-minute compensation of the PLS estimate for these causes. If compensation is not possible, it would at least be valuable to detect when the potential for estimator error is great; that is, when the current input data are inconsistent with the base data set on which the PLS model is based. PCA residual analysis is a tool capable of this type of detection, and this subject is addressed in the next chapter.

CHAPTER 5

PROCESS MONITORING AND ESTIMATOR VALIDATION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

5.1 Process Measurements and Process Monitoring

The goal of this section of the work is to provide a convenient, accurate means for monitoring column performance using a multivariate approach. Process monitoring using PCA is examined through the use of the Q, or residual, statistic. Excursions in process data that are inconsistent with variability and covariability in the base data set result in an abnormally high value of Q. In addition, the use of Q as a measure of validity for the PLS estimator examined in chapter 4 will be addressed.

As with the Chapter 4 PLS discussion, this chapter's work is based on a simulation and uses the same base data set comprised of all 25 tray temperatures, tray 7 feed, tray 21 feed, distillate flow, reflux flow, bottoms flow, and reboiler duty. Column pressure and catalyst feed rate are assumed to be constant.

5.2 Baseline Data Set

PCA, like PLS, is a data-based approach and requires a good representative data set for construction of a reliable model. In this technique, only an **X** matrix of separate observations of a set of input measurements is required. PCA provides the user with measures indicating the consistency of measurement observations with

"normal" process behavior, so it is key to have a base case data set with observations across all valid operating conditions.

The issues regarding gathering data are similar to those discussed in section 4.2. One difference is the availability of input measurement data. In PLS, both input measurement and quality measurement data must be provided on a consistent observation frequency. In some cases, the quality measurements of interest might be routinely sampled every 6 or 12 hours, and special arrangements must be made to schedule more frequent sampling in order to get the data needed for the PLS model. The need for more frequent quality measurement acquisition tends to drive the user toward the process experiment approach for data acquisition. Input measurement data, however, are often available at very high frequency (on the order of seconds or minutes) and are archived for months or years. Therefore, with PCA modeling the model builder has a wealth of historical plant data from which to select sections in which the process is running "normally".

In this work, the PLS model was constructed with a base data set determined through process experiments. Since one process monitoring goal in this case is to provide some indication of the validity of the PLS estimator, it is logical to use the same **X** block of data as was used for the PLS regression in PCA modeling. Therefore, the PCA model will provide a measure of consistency of new data observations with the input data used for formation of the PLS estimator. Inconsistency with the input data (evidenced by a high Q statistic via PCA analysis) indicates a departure in process behavior from the relationships on which the PLS estimator was developed, which leads to error in the composition estimate itself. The **X** block from the PLS base data set (generated by the process experiment runs detailed in Table 4.1) is used for construction of the PCA model. This resulted in a data set containing 53 observations, with 31 process measurements in each observation.

5.3 Development of PCA Model

The PCA model was developed using the MATLAB matrix mathematics software package and Wise's PLS Toolbox. Scaling issues in PCA are similar to those encountered in PLS. Mean centering and scaling to unit variance were again used to avoid artificially weighting variables with larger magnitudes.

The PCA regression results in a number of latent variables, each with an associated percent of variation explained for the **X** block (Table 5.1.) The next step is to select the number of latent variables to retain. Since eight latent vectors were retained in construction of the PLS estimator, it seems reasonable to initially consider eight latent PCA vectors, using the argument that both techniques are based on the concept of an inherent "dimensionality" of the data which is less than the actual number of measurements. Inclusion of eight latent vectors results in containment of over 99% of the base data set variance in the space spanned by these vectors.

Table 5.1. PCA regression results, showing individual percent of variance explained for each latent vector and total percent variance explained for eight latent vector PCA model.

Latent Vector	**X** Block % Variance Explained
1	58.2
2	12.7
3	9.3
4	8.9
5	5.4
6	2.3
7	1.5
8	0.9
Total % Explained	99.2

It is interesting to compare the percent variance captured in the **X** block by each latent vector in the PCA regression to the same measure for the PLS regression (Table 5.2.) Several differences support the fact that PLS and PCA use different "objective functions" in performing their regressions; PLS is based on both **X** variance and covariance between **X** and **Y**, while PCA is based strictly on **X** variance. The first latent vector in the PCA regression explains more variance than any other latent vector, and each subsequent vector accounts for a decreasing amount of variance. In contrast, the maximum amount of **X** variance explained in the PLS regression is in the second latent vector, and the amount of variance, while generally decreasing as latent vectors are added, does not always decrease for an additional latent vector. This is due to the influence of covariance between **X** and **Y** on the PLS regression. Also, the PCA latent vectors explain 99.2% of the variance, while the PLS latent vectors explain 98.4%. These results support Jackson's theorems regarding PCA (section 2.1.)

Table 5.2. Comparison of PLS and PCA percent of **X** block variance explained for each latent vector.

Latent Vector	PCA **X** Block % Variance Explained	PLS **X** Block % Variance Explained
1	58.2	20.5
2	12.7	45.8
3	9.3	13.0
4	8.9	5.8
5	5.4	5.1
6	2.3	5.4
7	1.5	1.1
8	0.9	1.8
Total % Explained	99.2	98.4

Once the projections of the **X** data onto the latent vectors have been removed from the data set, what is remaining is referred to as the residual. In PCA, the Q statistic is used as a measure of the amount of residual information remaining after information projected onto the latent vectors is removed. Q can be thought of as the perpendicular distance of the observation above or below an *A*-dimensional hyperplane spanned by the latent vectors.

After a PCA model is constructed, the Q statistic for the base data set can be calculated and plotted (Figure 5.1.) Using the Q values for these data, a 95% confidence limit can be developed and used as an indicator of "normal" or "abnormal" deviation from the PCA latent vectors, in much the same way that a range limit on an R or R_{bar} chart indicates process deviation based on past process performance (Wise 1991).

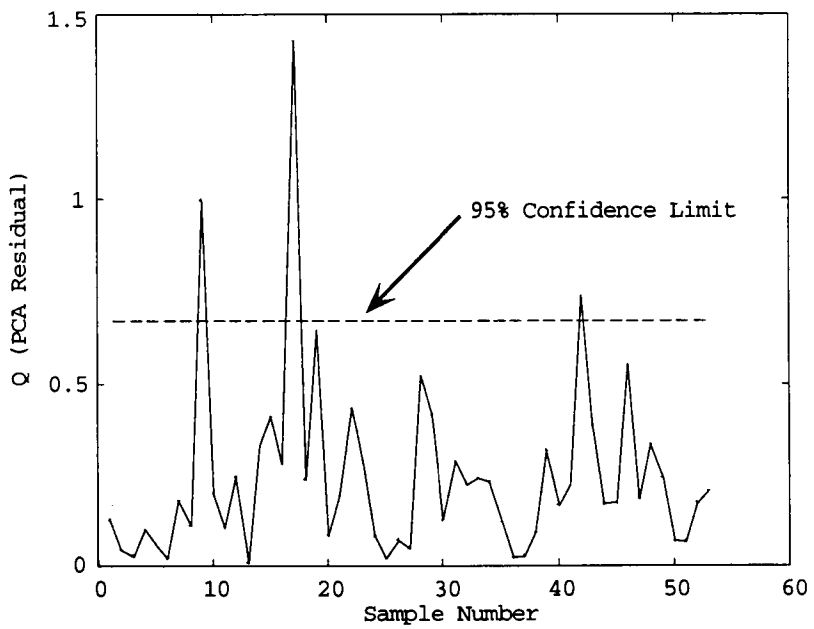


Figure 5.1. Residual information for base data set with 95% confidence limit.

5.4 Use of PCA Q Statistic for Process Monitoring

Once a PCA model has been developed using the base data set, the model should be applied to new data observations in order to test and evaluate its performance.

An ideal process monitor would signal the user whenever the process departed even slightly from its fundamental process variable relationships, determined by reaction kinetics, equilibrium conditions, and other physical and chemical relationships, while not producing "false alarms" based on measurement noise or failures.

The previously developed PLS estimator relies on process conditions similar to the base data set for accurate estimation of composition. If the process conditions change significantly, the very relationships on which the estimator is based may have changed, and the estimate is subject to severe error. Therefore, an ideal process monitor would provide value to the PLS estimator by indicating consistency of the current measurement data to the base data set, and therefore indicating in some form the validity of a composition estimate based on the current data.

While not an ideal process monitor, PCA can provide a good indication of deviation of the current data from the base data set performance through the Q statistic. In this section, the Q statistic is used to indicate inconsistency of the current measurement data from the base data set and evaluated as a tool in determining the validity of the PLS-based composition estimate.

This implementation of PCA Q statistics shares one complication with the PLS estimator implemented earlier (section 4.4) -- the dynamic nature of control and continuous processes. The PCA model discussed here was developed using only steady state process data. As the

process is subjected to disturbances, it will go through dynamic responses not accounted for in the PCA model. This will likely result in Q statistics indicating inconsistency with the base data set during dynamic process upsets.

In the following tests, the same series of process changes are made on the reactor column as were made in Section 4.4 for testing of the PLS composition estimate. An EPC control system is implemented on the rigorous distillation simulation as described in Figure 3.5. The tray 8 composition controller is implemented using the PLS estimate. In these tests however, the Q statistic for each run is also calculated and presented. The values for Q are normalized around the 95% confidence limit for Q as determined using the base data set, with a Q value of 1 corresponding to the 95% confidence limit point. Process responses are presented for both the tray 8 composition and Q. Again, four examples (Cases 1-4) are presented with disturbances included in the original data set, and six (Cases 5-10) with unmeasured or infrequent disturbances not included in the data set. The Q statistic should provide an indication that disturbances in Cases 1-4 are consistent with the base data set, and in Cases 5-10 should indicate process performance inconsistent with the base case data set.

In Case 1, the column production rate is reduced from 50 mol/hr to 40 mol/hr (Figure 5.2.) [As with all following figures, the composition estimate for tray 8 is represented as a dashed line, and the actual compositions for tray 8 and the Q statistic as solid lines. The dotted line on the Q statistic graph indicates the 95% confidence limit for Q.] Q initially rises, but quickly returns to a level indicating normal process behavior.

For Case 2, the tray 4 temperature controller setpoint is raised from 90°C to 100°C (Figure 5.3.) Q generally remains below the 95% confidence limit, but has a large spike corresponding to the dip in

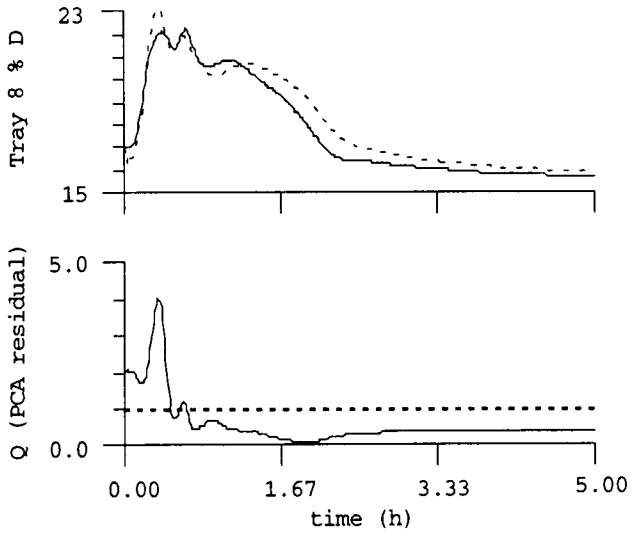


Figure 5.2. Process and residual responses for Case 1 (production rate reduced from 50 mol/hr to 40 mol/hr)

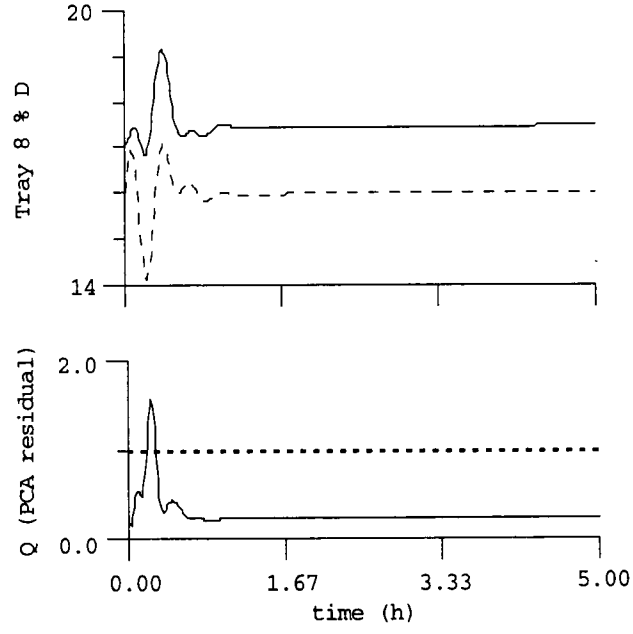


Figure 5.3. Process and residual responses for Case 2 (tray 4 temperature setpoint changed from 90°C to 100°C)

the composition estimate at t=20 min. This is a good example of process data inconsistent with the base data set leading to poor composition estimation being detected via Q.

For Case 3, the tray 8 estimated composition controller setpoint is raised from 16% D to 21% D (Figure 5.4.) Q remains well below the confidence limit, indicating normal process behavior throughout the disturbance.

In Case 4, the column reboiler duty is increased from 1.80E+06 BTU/hr to 2.05E+06 BTU/hr (Figure 5.5.) Q goes up dramatically during the initial upset, again corresponding to dynamic process changes. As the estimator lines out after about 1.5 hours, Q drops below the confidence limit, indicating normal operation.

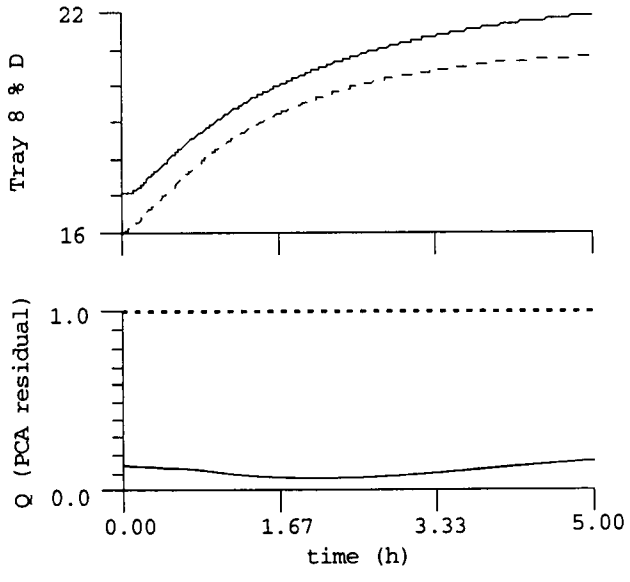


Figure 5.4. Process and residual responses for Case 3 (tray 8 estimated composition controller setpoint changed from 16% D to 21% D)

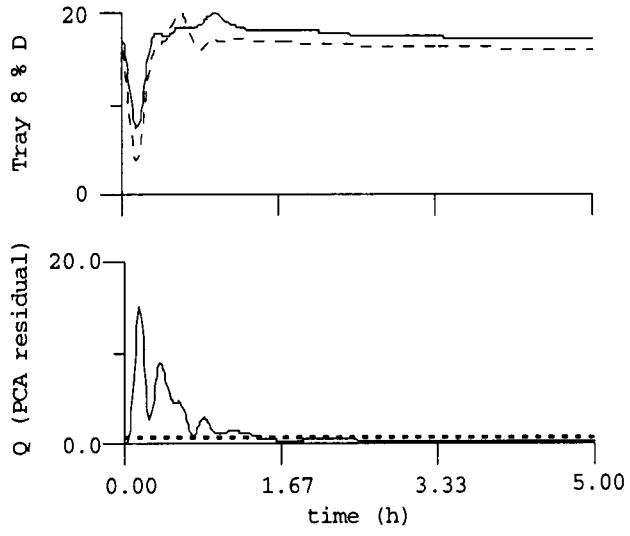


Figure 5.5. Process and residual responses for Case 4 (reboiler duty increased from 1.80E+6 BTU/hr to 2.05E+6 BTU/hr)

For Case 5, the first of the unmeasured or infrequent disturbances, the composition of the tray 7 feed is changed from 0.01% C to 15% C (Figure 5.6.) Q experiences an initial peak associated with introduction of the disturbance, and returns to a "steady state" value as before. In this case however, Q ends up well above the confidence limit for normal operation. This indicates inconsistency between the new process data and the base case data -- a shift in the process operation which may invalidate the PLS composition estimate.

In Case 6, the composition of the tray 21 feed is changed from 0% C to 15% C (Figure 5.7.) Q again holds at a value above the confidence limit, indicating a shift in process conditions.

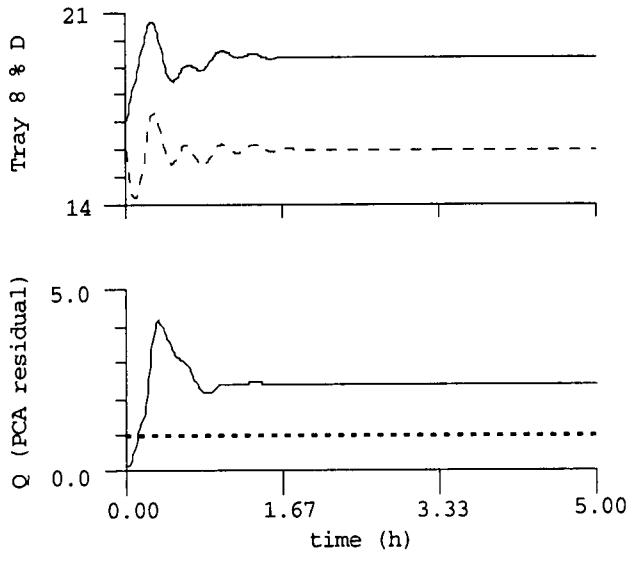


Figure 5.6. Process and residual responses for Case 5 (tray 7 feed composition changed from 0.01% C to 15% C)

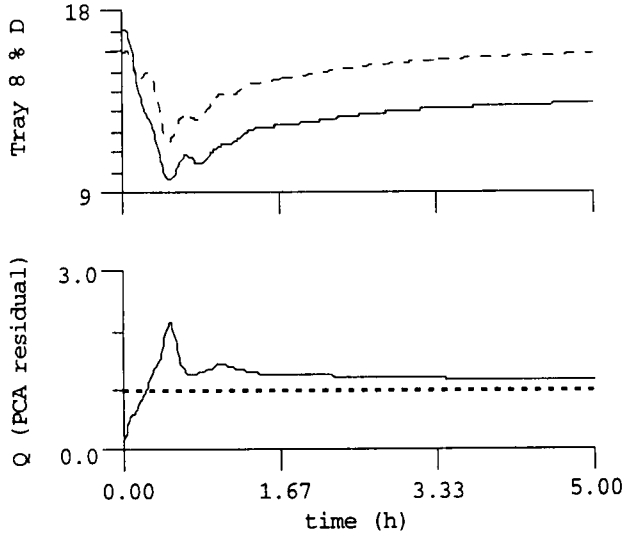


Figure 5.7. Process and residual responses for Case 6 (tray 21 feed composition changed from 0% C to 15% C)

For Case 7, the top column pressure is increased from 850 torr to 1000 torr (Figure 5.8.) Q indicates a departure from the base data set, which in this case is likely due to the elevated temperature measurements resulting from the higher top pressure.

In Case 8, the tray 17 catalyst feed flow is reduced from 0.5 mol/hr to 0.05 mol/hr (Figure 5.9.) As mentioned in Section 4.4, reduction of catalyst flow dramatically changes column operation, moving from a mode of separating products to separating unreacted raw materials. This invalidates the PLS model, which uses a base data set containing data from the reaction region of operation.

In this case, the PCA monitor fails. Q indicates no abnormal process behavior. The relationships between the process measurements used by PCA for calculation of Q do not drastically change, and as a result no problem is detected. This is a shortcoming of the data-based nature of PCA -- no fundamental or first principles relationships are used, only observations of past and current process behavior. If a crucial input such as the catalyst flow is unavailable or excluded from the PCA model, the process monitor has no way to deduce or infer the influence of that input.

Inclusion of the catalyst flow as a measurement would result in a better indication of the process abnormality. Including additional variables comes at a price, especially if a process experiment approach is used for collecting a base data set, as each additional input adds complexity and possibly another manipulated variable to be exercised during the data collection process.

Cases 9 and 10 are demonstrations of the effect measurement failure has on the estimator. In Case 9, a temperature which has a high weight in the PLS regression vector (tray 7) is "frozen" or fails to

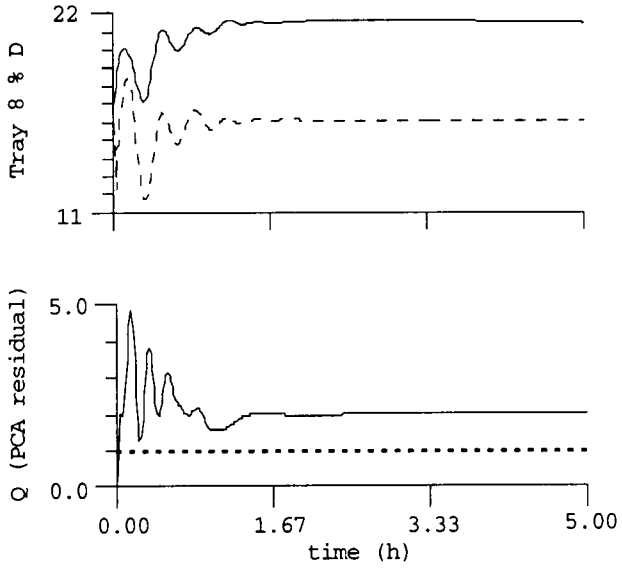


Figure 5.8. Process and residual responses for Case 7 (column top pressure changed from 850 torr to 1000 torr)

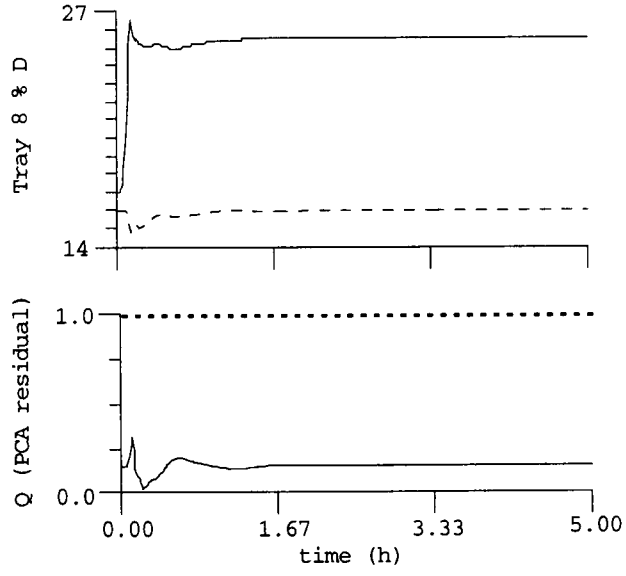


Figure 5.9. Process and residual responses for Case 8 (tray 17 catalyst feed reduced from 0.50 mol/hr to 0.05 mol/hr)

its current value. A production rate increase from 50 mol/hr to 60 mol/hr is then made (Figure 5.10.) Q rises above the confidence limit almost immediately, and continues to climb as the composition decreases.

For Case 10, a temperature with a lower weight in the PLS regression vector (tray 14) is "frozen", and the same production rate increase is made (Figure 5.11.) As loss of the less critical temperature measurement has less effect on the PLS estimate than the previous case, it also has less effect on Q. Q rises above the confidence limit for a short time, but returns to an indication of normal operation.

5.5 Discussion of Results

PCA residual analysis serves as a useful indicator of consistency of process data with the base data set (Table 5.3). When disturbances are introduced that are part of the base data set (Cases 1-4), the Q statistic has a steady state value below the 95% confidence limit. For disturbances not included in the base data set, the Q statistic remains above the 95% confidence limit at steady state for 4 of 6 cases.

Cases 1, 2, and 4 display similar characteristics: a short-lived increase in Q, followed by a return to a value below the 95% confidence limit. This response indicates that the process data associated with the initial disturbance and the associated transient are inconsistent with the base data set -- not surprising since the base data set is based on steady state runs only. Case 3 remains well below the 95% confidence limit for the entire run, due mostly to the gentle nature of the disturbance introduced.

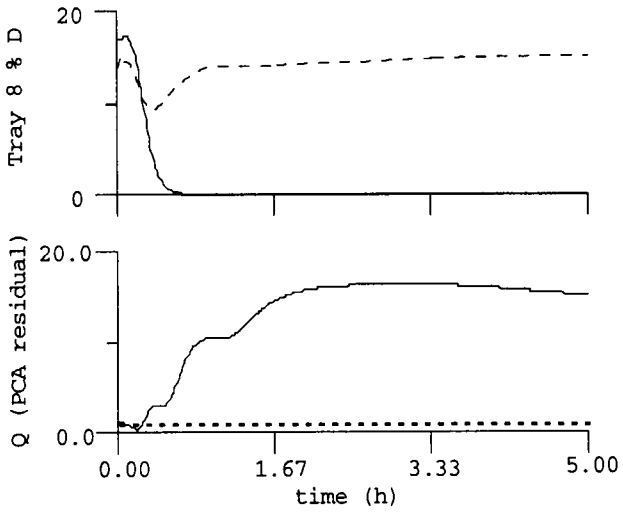


Figure 5.10. Process and residual responses for Case 9 (production rate increase from 50 mol/hr to 60 mol/hr, and temperature measurement on tray 7 fails to last good value)

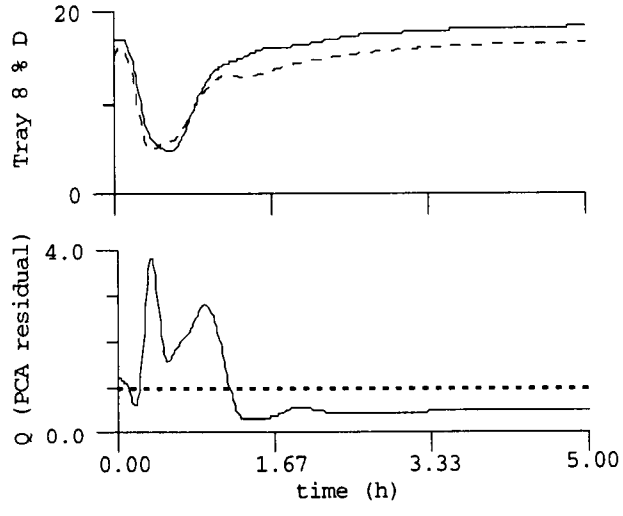


Figure 5.11. Process and residual responses for Case 10 (production rate increase from 50 mol/hr to 60 mol/hr, and temperature measurement on tray 14 fails to last good value)

Table 5.3. PLS estimator error and normalized PCA residual (Q) values for disturbance cases.

Case #	PLS estimator error @ t=5 hr	PCA residual (Q) value @ t=5 hr
1	0.42%	0.44
2	-1.47	0.25
3	-1.14	0.21
4	-1.28	0.50
5	-3.38	2.43
6	2.45	1.19
7	-5.46	2.01
8	-9.50	0.19
9	15.48	15.25
10	-2.57	0.51

On Cases 5-10, the Q statistic identifies 4 of the 6 cases as inconsistent with the base data set. One of the two not identified (Case 10) was associated with failure of a measurement with a low amount of variance; therefore, its failure to a normal value had little contribution to the residual, resulting in no detection via Q.

The second case not identified, Case 8, is not as easily explained. With a 90% reduction in catalyst flow, the column is drastically affected, but these changes do not show up in Q. This indicates that the specific set of column operation changes associated with this disturbance have little or no effect on the input data set, comprised primarily of process temperatures and flows. There are processes, typically distillation columns, for which "multiple steady states" have been identified -- cases in which similar sets of physical measurements do not always result in the same set of products and by-products. In previous work, both Roat and Simandl identified two steady-states associated with similar process measurements for this column. As a data based technique, PCA is

limited to the input information it is provided. For a change in process operation which is not accompanied by a difference in PCA input measurements, the technique will fail to detect the change.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE STUDY

6.1 Conclusions

The simulation runs indicate that PLS and PCA are useful for constructing an estimator of mid-column composition for closed loop control and developing a monitoring system to detect excursions in column operation not observed in the base data set. For cases 1-4, based on disturbances included in the base data set, the PLS estimator accurately estimates the tray 8 D composition. Additionally, the control of the estimated composition leads to low distillate impurity levels in all cases except Case 4, where reboiler duty is increased, causing more impurity to be taken overhead. This exception however, is due to control strategy design rather that estimator error since the composition estimate for Case 4 is accurate to within 2%. Cases 5-10 were developed using disturbances not included in the base data set and have considerably larger amounts of PLS estimator error and corresponding higher levels of distillate impurity. In particular, Cases 8 and 9 exhibit gross error in estimator prediction.

A process monitoring system using a PCA residual, or Q analysis, is successful in identifying departure from normal process behavior as defined by the base data set. Cases 1-4 exhibit Q values below the 95% confidence limit after brief increases associated with dynamics caused by the disturbance. In Cases 5-10, the Q statistic correctly identifies the abnormal process excursion 4 out of 6 times. One failure (Case 10) is due to the low level of model error introduced by the particular disturbance, and the other failure (Case 8) is due

to the lack of process measurement input changes to accompany the shift in process operating condition. This may be attributable to the phenomena of multiple steady states observed in simulations of this column by Roat and Simandl.

6.2 Recommendations for Future Study

PLS and PCA technology are just beginning to gain acceptance in industrial circles as useful techniques for estimation and process monitoring. There are a number of areas of future study which if explored, could address the goal of further industrial implementation of these technologies to solve real world chemical processing problems.

The issue of how to assemble a base data set is an area of potential work. Currently, the process is fairly arbitrary and is based primarily on the guideline of ensuring all operating conditions of interest are represented. A treatment of this issue, exploring the differences between models based on experimental data versus models based on historical plant data, might provide a more systematic and efficient method of gathering needed base data sets. Exploration of how to extract useful information from existing process measurement and quality measurement databases would also be useful in addressing concerns regarding data alignment and inclusion of data taken during process transitions versus data taken when the process is at "steady state".

This work demonstrates the usefulness of PCA residual analysis in monitoring process data and asserts that an abnormally high PCA residual indicates the potential for error in the PLS composition estimate. A useful additional step would be to develop an "uncertainty" measure using the PCA and PLS statistics, calculating

the potential error in the estimate based on the residual information from the PCA model. Such an approach could yield an estimated value with associated error bars, graphically indicating to the user the validity of the current estimate. Further examination of closed loop composition control using PLS composition estimates could explore the potential of adapting the gain on the composition controller based on the PCA Q statistic or the estimator "uncertainty" discussed above.

The steady state nature of the base data set in this work led to an estimator which had some problems associated with dynamic behavior in the process. A treatment of estimators and monitors built using steady state data compared to those built using dynamic data would be helpful in clarifying the benefits of a dynamic-based PLS and PCA model and in specifying a guideline for the construct of such a base data set.

Unfortunately, industrial processes do not offer the high number of process measurements available through a process simulation. In the case of the reactive distillation column, one might expect to find 4 or 5 trays, at most, instrumented for temperature measurement in a 25 tray column. As a result, it is key to properly identify measurements of value during plant design. Analysis using steady-state temperature gains derived from rigorous process simulations is routinely used at Eastman for identification of crucial temperature measurements for temperature control early in a distillation column's design. A similar approach could be explored for using PLS analysis to identify key measurements needed to predict a composition of interest in a column or other piece of equipment.

Over time, it is common for processes to be modified and improved to achieve higher production rates, less by-product production, and better quality. These changes can invalidate an existing PLS/PCA

model, since such improvements will most likely change the nature of the process measurement data. Development of methods by which PLS and PCA models can be updated in a systematic fashion is important to the long-term industrial use of these technologies. An approach which allows for the "forgetting" of old process behavior in favor of new process behavior would enhance the usefulness of PLS/PCA applications over the life of a plant.

LIST OF REFERENCES

Agreda, V.H. and L.R. Partin, U.S. Patent 4,435,595, Mar. 1984.

Agreda, V.H., L.R. Partin and W.H. Heise, *High-Purity Methyl Acetate via Reactive Distillation*. Chemical Engineering Progress, 1990. (2): p.40-46.

Downs, J.J., *The Control of Azeotropic Distillation Columns*. 1982, University of Tennessee:

Harris, T.J. and W.H. Ross, *Statistical Process Control Procedures for Correlated Observations*. The Canadian Journal of Chemical Engineering, 1991. **69**: p.48-57.

Hoskuldsson, A., *PLS Regression Methods*. Journal of Chemometrics, 1988. **2**: p.211-228.

Hunter, J.S., T.J. Harris, and J.F. MacGregor, *SPC Interfaces -- Engineering and Statistical Process Control*. presented at Eastman Chemical Company, Kingsport, TN, June 28-30, 1993.

Jackson, J.E., *A User's Guide to Principal Components*. 1st ed. 1991, New York: John Wiley & Sons.

Jackson, J.E. and G.S. Mudholkar., *Control Procedures for Residuals Associated With Principal Component Analysis*. Technometrics, 1979. **21**(3): p.341-349.

Jackson, J.E., *Principal Components and Factor Analysis: Part I -- Principal Components*. Journal of Quality Technology, 1980. **12**(4): p.201-213.

Jacobs, D.C., *Watch Out for Nonnormal Distributions*. Chemical Engineering Progress, 1990. (11): p.19-27.

Kozub, D.J. and C.E. Garcia, *Monitoring and Diagnosis of Automated Controllers in the Chemical Process Industries*. presented at AIChE Meeting, St. Louis, MO, November 9, 1993.

Kresta, J.V., *The Application of Partial Least Squares to Problems in Chemical Engineering*. 1992, McMaster University:

Kresta, J.V., *Development of Inferential Process Models using PLS*. Computers in Chemical Engineering, 1994. **18**(3): p.597-611.

Kresta, J.V., J.F. MacGregor and T.E. Marlin, *Multivariate Statistical Monitoring of Process Operating Performance*. The Canadian Journal of Chemical Engineering, 1991. **69**: p.35-47.

MacGregor, J.F., T.E. Marlin, J.E. Kresta, and B. Skagerberg, *Multivariate Statistical Methods in Process Analysis and Control*. in *CPC-IV*, 1991. South Padre Island, TX:

MacGregor, J.F., *On-Line Statistical Process Control*. Chemical Engineering Progress, 1988. (10): p.21-31.

Mejdell, T., *Estimators for Product Compositions in Distillation Columns*. 1990, University of Trondheim, The Norwegian Institute of Technology:

Nomikos, P. and J.F. MacGregor, *Monitoring of Batch Processes using Multi-way Principal Component Analysis*. Unpublished, 1993.

Roat, S.D., J.J. Downs, E.F. Vogel, and J.E. Doss, *The Integration of Rigorous Dynamic Modeling and Control System Synthesis for Distillation Columns: An Industrial Approach*. in *CPC-III*, 1986. Asilomar, CA:

Roffel, J.J., J.F. MacGregor, and T.W. Hoffman, *The Design and Implementation of a Multivariable Internal Model Controller for a Continuous Polybutadiene Polymerization Train*. in *DYCORD+'89*, 1989. Maastricht, The Netherlands:

Simandl, J., *Simulation of Distillation Towers with Chemical Reactions*. 1988, University of Calgary:

Strang, G., *Linear Algebra and its Applications*. 3rd ed. 1988, Fort Worth: Harcourt Brace Jovanovich.

Vogel, E.F., *Minimizing Product Quality Variability with SPC and EPC*. presented at Eastman Chemical Company, Kingsport, TN, May 1993.

Williams, V.J., *Management of Variation in Plantwide Process Control System Design and Analysis*. 1992, University of Tennessee:

Wise, B.M., *Adapting Multivariate Analysis for Monitoring and Modeling of Dynamic Systems*. 1991, University of Washington:

Wise, B.M., *PLS Toolbox 1.4*. 1994.

Wise, B.M. and N.L. Ricker, *Feedback Strategies in Multiple Sensor Systems*. AIChE Symposium Series, 1989. **267**(85): p.19-23.

Wise, B.M. and B.R. Kowalski, *Process Chemometrics: The Case for Chemometrics in Chemical Engineering*. AIChE Computing and Systems Technology Division Communications, 1993. **16**(2), p.5-11.

Wold, S., *Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Analysis*. Technometrics, 1978. **20**(4): p.397-405.

APPENDICES

APPENDIX A

DATA FOR REACTIVE DISTILLATION COLUMN SIMULATION

As discussed in the text (Chapter 3), the column studied in this work has the same basic configuration as those addressed by Roat *et al.* and Simandl. Column specifications, reaction data, physical properties, and vapor-liquid equilibrium parameters are provided below. Where there are differences between this work and Roat or Simandl, they are noted.

I. Column Description

25 trays plus a total condenser, numbered 1 (reboiler) thru 25

Stage efficiency	1.0
Reflux ratio	1.68
Column top pressure	850 torr
Column bottom pressure	1073 torr

Note: Roat used 1084 torr

Tray reaction volume	1 ft^3
Reboiler Duty	1.8 x 10^6 BTU/hr

Note: Roat used 1.96 MBTU/hr, Simandl used 1.90 MBTU/hr

Column Feeds

	Tray 7	Tray 17	Tray 21
Flowrate (mol/hr)	49.86	0.50	48.90
Temperature (C)	50.00	25.00	25.00
Composition (mol %):			
A	0.00	0.00	0.01
B	99.99	0.00	0.00
C	0.01	0.00	0.00
D	0.00	0.00	99.99
E	0.00	100.00	0.00

Note: In Roat, feed rates were 50.42 (7), 0.50 (17), and 49.00 (21), (7) was 100% B, and (21) was 100% D. In Simandl, feed rates were 50.00 (7), 0.50 (17), and 50.00 (21).

Column Product Streams

	Distillate	Bottoms
Flowrate (mol/hr)	50.00	49.26
Temperature (C)	50.00	50.00
Composition (mol %):		
A	97.71	0.00
B	1.71	0.31
C	0.55	98.63
D	0.03	0.05
E	0.00	1.01

Note: In Roat, distillate rate was 50.00, with compositions (A-E) of 97.94, 1.56, 0.48, 0.02, and 0.00. Bottoms rate was 49.92, with compositions of 0.00, 1.34, 97.61, 0.04, and 1.00. In Simandl, distillate rate was 50.00, with compositions of 97.79, 1.73, 0.46, 0.02, and 0.00. Bottoms rate was 50.54, with compositions of 0.04, 0.44, 96.33, 2.20, and 0.99.

II. Reaction Data

$$B + D \rightleftharpoons A + C$$

Reaction takes place in the liquid phase

Catalyst: Component E, delivered as liquid

Rate Constants:

$$K_{for} = 6.970 \times 10^{10} * e^{(-9198/RT)}$$

$$K_{rev} = 1.338 \times 10^{10} * e^{(-9198/RT)}$$

[R = 1.987 cal/(gmol K), T in K]

Reaction Rate: $K_{for}\ C_E\ C_B\ C_D - K_{rev}\ C_E\ C_A\ C_C$

[lbmol/(hr ft^3), C_x in lbmol/ft^3]

Heat of Reaction: 1687 BTU/lbmol A formed

III. Physical Property Data

Molecular Weight (lb/lb mol):

A	74.08
B	32.04
C	18.02
D	60.05
E	98.00

Liq Molar Volume (cc/gmol):

A	80.17
B	39.26
C	18.80
D	55.83
E	98.00

Vapor Pressure [torr, $P_i = \exp(A_{vpi} + B_{vpi}/(T + C_{vpi})$, T in °C]:

	A_{vpi}	B_{vpi}	C_{vpi}
A	16.26831	-2665.54	219.73
B	18.71857	-3720.83	242.66
C	18.67190	-4030.92	235.00
D	16.81508	-3404.00	216.75
E	2.00000	- 0.01	0.00

Liquid Heat Capacity [BTU/lb °F, $h_i = A_{li} + B_{li}*T + C_{li}*T^2$, T in °C]:

	A_{li}	B_{li}	C_{li}
A	0.4540	1.43E-3	0.0
B	0.53	2.83E-3	0.0
C	1.0034	-2.11E-4	2.26E-6
D	0.4710	7.70E-4	0.0
E	1.00	0.0	0.0

Heat of Vaporization [BTU/lb, $\Delta H_i = A_{vi} + B_{vi}*T\ C_{vi}*T^2$, T in °C]:

	A_{vi}	B_{vi}	C_{vi}
A	195.00	-0.317	0.0
B	529.00	-0.867	0.0
C	1056.32	-3.59	1.3E-3
D	201.00	-0.232	0.0
E	200.00	0.0	0.0

IV. Vapor-Liquid Equilibrium Data

Wilson Parameters, given in terms of A_{ij}s, where

$$\ln \gamma_i = 1 - \ln\left(\sum_j x_j A_{ij}\right) - \sum_j \frac{x_j A_{ij}}{\sum_k x_k A_{kj}}$$

		component i				
		A	B	C	D	E
component j	A	1.0000	0.55770	0.25200	0.30650	1.0000
	B	0.58950	1.0000	1.0565	0.40610	1.0000
	C	0.21480	0.42840	1.0000	0.31570	1.0000
	D	1.7405	1.9870	1.4826	1.0000	1.0000
	E	1.0000	1.0000	1.0000	1.0000	1.0000

APPENDIX B

BASE DATA SET FOR PLS AND PCA MODELING

Base Data Set - X Data Obs #	T1	T2	T3	T4	T5	T6	T7	T8	T9
1	109.6	107.4	100.9	90.0	81.6	75.4	71.3	72.6	74.7
2	109.8	108.1	102.6	90.0	77.5	70.5	68.4	69.8	71.1
3	109.5	107.0	100.1	90.0	83.0	78.1	74.0	76.1	80.0
4	109.5	107.2	100.5	90.0	82.4	76.9	72.7	74.3	77.1
5	109.7	107.7	101.5	90.0	80.2	73.3	69.8	71.1	72.7
6	109.7	107.8	101.7	90.0	79.9	73.0	69.8	70.9	72.3
7	109.7	108.0	102.2	90.0	78.6	71.6	69.1	70.2	71.3
8	109.5	107.2	100.4	90.0	82.6	77.2	72.8	74.5	77.8
9	109.5	106.9	99.9	90.0	83.4	78.7	74.4	77.0	82.2
10	111.9	109.7	104.1	90.0	77.4	73.5	73.4	80.2	83.6
11	110.4	109.1	104.7	90.0	75.1	70.8	70.5	75.6	78.4
12	109.9	108.8	104.6	90.0	74.6	69.9	69.5	73.7	76.1
13	109.7	107.9	102.0	90.0	78.7	71.9	69.8	72.7	75.0
14	109.6	107.3	100.5	90.0	83.0	78.6	74.4	73.2	74.4
15	109.6	107.2	100.3	90.0	83.6	80.5	77.8	76.0	74.5
16	109.6	107.2	100.3	90.0	83.7	81.1	79.6	80.7	76.8
17	109.6	107.2	100.3	90.0	83.8	81.3	80.2	83.6	83.9
18	108.8	104.0	94.0	85.0	79.5	74.6	71.1	72.5	74.7
19	99.1	88.3	82.8	80.0	77.2	73.4	70.6	72.2	74.5
20	109.8	108.6	104.7	95.0	84.1	76.3	71.6	72.7	74.8
21	109.9	109.1	107.0	100.0	88.0	77.9	72.0	72.9	74.9
22	108.6	103.3	93.1	85.0	80.8	77.1	73.0	72.7	74.7
23	109.9	108.5	104.6	95.0	85.2	78.7	73.6	73.0	74.8
24	109.2	105.5	96.2	85.0	76.7	71.4	69.5	72.7	75.3
25	109.9	108.7	105.1	95.0	81.7	73.0	69.9	72.8	75.4
26	109.1	104.8	94.8	85.0	79.2	74.4	70.4	70.0	70.8
27	109.9	108.7	105.0	95.0	83.9	76.1	71.0	70.2	70.9
28	109.9	108.4	102.4	85.0	71.9	68.6	68.3	70.7	72.0
29	110.0	109.1	106.5	95.0	76.8	69.7	68.5	70.8	72.1
30	108.0	101.6	91.6	85.0	81.8	79.1	75.6	76.9	80.9
31	109.8	108.4	104.3	95.0	86.2	80.7	76.2	77.3	81.2
32	108.4	102.7	92.7	85.0	80.3	75.7	72.3	76.3	81.0
33	109.8	108.4	104.5	95.0	84.8	77.4	72.8	76.6	81.2
34	108.6	103.5	93.3	85.0	80.7	77.0	73.1	72.8	74.3
35	109.8	108.5	104.6	95.0	85.1	78.6	73.7	73.0	74.4
36	109.3	105.8	96.7	85.0	76.3	71.3	69.7	72.6	74.7
37	109.8	108.7	105.2	95.0	81.4	72.8	70.1	72.7	74.8
38	90.7	83.4	81.0	80.0	79.3	78.3	76.3	75.0	74.3
39	110.0	109.1	106.9	100.0	89.7	82.9	78.7	76.6	75.1
40	109.9	107.6	98.0	80.0	71.3	69.5	69.6	74.9	77.9
41	110.2	109.4	107.8	100.0	82.2	71.9	70.1	75.1	78.0
42	103.4	91.6	83.5	80.0	78.3	76.6	73.7	70.2	68.3
43	110.1	109.2	107.1	100.0	88.4	80.5	75.2	70.8	68.5
44	112.0	108.2	97.6	80.0	72.0	70.3	70.2	73.0	74.2
45	112.3	110.6	108.4	100.0	82.5	72.7	70.7	73.1	74.1
46	88.0	82.4	80.7	80.0	79.5	78.6	77.2	78.9	81.2
47	109.9	109.0	106.8	100.0	90.1	83.7	80.0	81.0	82.8
48	106.7	98.2	87.7	80.0	74.1	70.9	70.3	77.6	83.5
49	109.9	109.1	107.2	100.0	85.6	74.4	71.2	78.0	83.8
50	92.1	84.0	81.1	80.0	79.3	78.3	76.5	75.5	74.0
51	109.9	109.1	106.9	100.0	89.5	82.7	78.6	76.7	74.7
52	109.9	107.5	97.8	80.0	71.6	70.0	70.2	74.7	77.0
53	110.2	109.5	107.8	100.0	82.5	72.5	70.8	74.9	77.1

Base Obs	Data #	Set - X Data (continued) T10	T11	T12	T13	T14	T15	T16	T17	T18
	1	76.1	76.7	76.9	76.8	76.6	76.4	76.1	75.9	75.5
	2	71.8	72.2	72.3	72.3	72.2	72.1	71.9	71.7	71.5
	3	82.7	84.1	84.7	84.7	84.3	83.8	83.2	82.6	82.0
	4	79.0	79.9	80.1	80.0	79.7	79.3	79.0	78.6	78.2
	5	73.8	74.2	74.4	74.4	74.3	74.1	73.9	73.7	73.3
	6	73.2	73.6	73.8	73.8	73.7	73.5	73.3	73.1	72.8
	7	72.0	72.3	72.5	72.5	72.4	72.3	72.1	71.9	71.7
	8	79.9	80.9	81.1	81.0	80.7	80.3	79.8	79.4	79.0
	9	86.1	88.4	89.9	90.7	91.0	91.1	90.8	90.3	89.5
	10	85.2	85.7	85.8	85.5	85.2	84.7	84.3	83.8	83.4
	11	79.7	80.1	80.1	80.0	79.7	79.4	79.0	78.7	78.3
	12	77.3	77.8	77.9	77.8	77.5	77.3	77.0	76.7	76.3
	13	76.3	76.8	77.0	76.9	76.7	76.4	76.2	75.9	75.6
	14	75.8	76.5	76.8	76.8	76.6	76.4	76.1	75.8	75.5
	15	75.5	76.3	76.6	76.7	76.5	76.3	76.1	75.8	75.5
	16	75.4	76.0	76.5	76.6	76.5	76.3	76.0	75.8	75.4
	17	79.6	76.6	76.4	76.5	76.4	76.2	76.0	75.7	75.4
	18	76.0	76.7	76.9	76.8	76.6	76.4	76.1	75.9	75.5
	19	75.9	76.5	76.7	76.7	76.5	76.3	76.0	75.8	75.4
	20	76.1	76.7	76.9	76.8	76.7	76.4	76.1	75.9	75.5
	21	76.2	76.8	76.9	76.9	76.7	76.4	76.2	75.9	75.5
	22	76.3	77.0	77.3	77.2	77.0	76.7	76.5	76.2	75.8
	23	76.3	77.1	77.3	77.2	77.0	76.8	76.5	76.2	75.8
	24	76.7	77.3	77.4	77.3	77.1	76.8	76.5	76.2	75.9
	25	76.7	77.3	77.4	77.3	77.1	76.8	76.5	76.3	75.9
	26	71.6	72.0	72.2	72.3	72.2	72.1	72.0	71.8	71.5
	27	71.6	72.1	72.3	72.3	72.2	72.1	72.0	71.8	71.5
	28	72.7	73.0	73.1	73.0	72.9	72.8	72.6	72.4	72.1
	29	72.7	73.0	73.1	73.1	72.9	72.8	72.6	72.4	72.1
	30	84.3	86.2	87.2	87.4	87.2	86.7	86.0	85.2	84.4
	31	84.6	86.5	87.5	87.7	87.5	87.0	86.3	85.4	84.6
	32	84.1	85.8	86.6	86.7	86.5	86.1	85.4	84.7	84.0
	33	84.3	86.0	86.8	86.9	86.7	86.2	85.6	84.9	84.1
	34	75.5	76.2	76.4	76.4	76.3	76.0	75.8	75.5	75.2
	35	75.6	76.3	76.5	76.4	76.3	76.1	75.8	75.6	75.2
	36	75.9	76.4	76.6	76.5	76.3	76.1	75.9	75.6	75.3
	37	76.0	76.5	76.6	76.6	76.4	76.2	75.9	75.6	75.3
	38	75.9	77.1	77.5	77.6	77.4	77.1	76.8	76.5	76.1
	39	76.5	77.4	77.8	77.7	77.5	77.2	76.9	76.6	76.3
	40	79.4	79.9	79.9	79.7	79.4	79.1	78.7	78.4	78.0
	41	79.4	79.9	79.9	79.7	79.4	79.0	78.7	78.4	78.0
	42	68.2	68.6	68.8	68.9	69.0	69.0	68.9	68.8	68.6
	43	68.3	68.6	68.8	69.0	69.0	69.0	68.9	68.8	68.6
	44	74.6	74.8	74.8	74.7	74.5	74.3	74.1	73.8	73.5
	45	74.6	74.7	74.7	74.6	74.4	74.2	74.0	73.7	73.4
	46	85.8	89.2	91.5	93.1	94.2	95.1	95.8	96.4	97.0
	47	87.0	90.1	92.3	93.7	94.8	95.7	96.4	96.9	97.5
	48	87.3	89.8	91.5	92.7	93.7	94.4	95.1	95.6	96.1
	49	87.6	90.0	91.7	92.9	93.9	94.6	95.2	95.8	96.3
	50	74.7	75.4	75.8	75.8	75.8	75.6	75.4	75.2	74.9
	51	75.1	75.7	76.0	76.0	75.9	75.7	75.5	75.3	75.0
	52	78.2	78.6	78.7	78.6	78.3	78.1	77.8	77.5	77.1
	53	78.2	78.6	78.7	78.5	78.3	78.0	77.7	77.4	77.1

Base Data Set - X Data (continued)

Obs	#	T19	T20	T21	T22	T23	T24	T25	F7	F21
	1	75.2	74.8	74.5	64.5	61.5	60.5	59.9	49.8	48.9
	2	71.2	70.9	70.6	62.9	61.0	60.3	59.7	39.8	38.7
	3	81.4	80.8	80.3	67.8	62.6	60.8	60.0	59.6	58.6
	4	77.8	77.4	77.0	65.7	61.9	60.6	60.0	54.8	53.9
	5	73.0	72.7	72.4	63.6	61.2	60.4	59.9	44.8	43.8
	6	72.6	72.2	71.9	63.4	61.1	60.3	59.8	49.8	48.5
	7	71.4	71.1	70.8	63.0	61.0	60.2	59.7	49.8	48.2
	8	78.6	78.1	77.7	66.2	62.0	60.7	60.0	49.8	49.0
	9	88.6	87.5	86.3	73.4	64.5	61.3	60.1	49.0	48.2
	10	83.0	82.6	82.2	68.3	62.7	60.9	60.2	49.9	77.1
	11	77.9	77.6	77.2	65.6	61.9	60.7	60.1	49.8	59.1
	12	76.0	75.6	75.3	64.8	61.6	60.6	60.0	49.8	51.8
	13	75.2	74.9	74.5	64.5	61.5	60.5	60.0	49.8	49.0
	14	75.2	74.8	74.5	64.4	61.5	60.5	59.9	49.8	48.8
	15	75.1	74.8	74.4	64.4	61.4	60.5	59.9	49.8	48.7
	16	75.1	74.7	74.4	64.4	61.4	60.4	59.8	49.8	48.5
	17	75.0	74.7	74.3	64.3	61.3	60.3	59.8	49.7	48.2
	18	75.2	74.8	74.5	64.5	61.5	60.5	59.9	50.1	48.9
	19	75.1	74.7	74.4	64.4	61.5	60.5	59.9	54.5	48.9
	20	75.2	74.8	74.5	64.5	61.5	60.5	59.9	49.8	48.9
	21	75.2	74.9	74.5	64.5	61.5	60.5	59.9	49.7	49.0
	22	75.5	75.1	74.8	64.6	61.5	60.5	60.0	45.1	44.0
	23	75.5	75.1	74.8	64.6	61.5	60.5	60.0	44.8	44.1
	24	75.5	75.2	74.8	64.6	61.6	60.5	60.0	45.0	44.2
	25	75.6	75.2	74.9	64.6	61.6	60.5	60.0	44.8	44.2
	26	71.2	70.9	70.6	62.9	61.0	60.2	59.7	45.0	43.3
	27	71.3	71.0	70.7	62.9	61.0	60.2	59.7	44.8	43.3
	28	71.8	71.5	71.2	63.1	61.1	60.3	59.8	44.8	45.6
	29	71.8	71.5	71.2	63.1	61.1	60.3	59.8	44.8	45.7
	30	83.6	82.8	82.0	69.2	63.0	60.9	60.0	54.9	53.5
	31	83.7	82.9	82.2	69.3	63.0	60.9	60.0	54.3	53.4
	32	83.3	82.6	82.0	69.2	63.0	60.9	60.1	54.8	53.7
	33	83.4	82.7	82.1	69.3	63.0	61.0	60.1	54.4	53.7
	34	74.9	74.5	74.2	64.3	61.4	60.5	59.9	55.1	53.6
	35	74.9	74.6	74.2	64.3	61.4	60.5	59.9	54.7	53.6
	36	74.9	74.6	74.3	64.3	61.5	60.5	59.9	54.9	53.8
	37	75.0	74.6	74.3	64.4	61.5	60.5	59.9	54.7	53.9
	38	75.8	75.4	75.1	64.7	61.6	60.5	59.9	51.8	39.1
	39	75.9	75.5	75.2	64.8	61.6	60.5	59.9	39.8	39.1
	40	77.6	77.2	76.9	65.5	61.9	60.7	60.1	39.9	44.0
	41	77.6	77.2	76.9	65.5	61.8	60.7	60.1	39.8	43.9
	42	68.4	68.1	67.9	61.9	60.5	59.8	59.1	41.5	36.8
	43	68.4	68.1	67.9	61.9	60.5	59.8	59.2	39.7	36.9
	44	73.2	72.9	72.6	63.4	61.2	60.4	59.9	40.1	62.7
	45	73.1	72.8	72.5	63.4	61.2	60.4	59.9	39.9	62.0
	46	97.6	98.3	99.0	98.4	97.1	92.9	80.4	69.6	56.5
	47	98.1	98.7	99.4	98.8	97.6	93.5	81.2	48.2	56.1
	48	96.8	97.5	98.3	97.7	96.2	91.6	78.9	51.4	57.6
	49	96.9	97.6	98.4	97.8	96.4	91.9	79.3	50.2	57.9
	50	74.5	74.2	73.9	64.1	61.3	60.4	59.8	74.3	58.1
	51	74.6	74.3	73.9	64.2	61.3	60.4	59.8	59.6	58.2
	52	76.8	76.4	76.0	65.0	61.7	60.6	60.0	59.8	67.9
	53	76.7	76.4	76.0	65.0	61.7	60.6	60.0	59.8	67.7

Base Data Set - X Data (continued)					Y Data
Obs #	D	R	B	Q	X 8, D
1	50.0	76.3	49.2	1.80E+06	16.0%
2	40.0	88.3	39.0	1.80E+06	16.0%
3	60.0	64.2	58.7	1.80E+06	16.0%
4	55.0	70.3	54.2	1.80E+06	16.0%
5	45.0	82.4	44.2	1.80E+06	16.0%
6	50.0	95.3	48.8	2.05E+06	16.0%
7	50.0	106.7	48.5	2.20E+06	16.0%
8	50.0	61.1	49.3	1.60E+06	16.0%
9	50.0	45.7	47.6	1.40E+06	16.0%
10	50.0	65.7	77.5	1.80E+06	40.0%
11	50.0	72.7	59.4	1.80E+06	30.0%
12	50.0	75.4	52.1	1.80E+06	25.0%
13	50.0	76.3	49.4	1.80E+06	20.0%
14	50.0	76.3	49.1	1.80E+06	10.0%
15	50.0	76.3	49.0	1.80E+06	4.0%
16	50.0	76.3	48.8	1.80E+06	1.0%
17	50.0	76.3	48.4	1.80E+06	0.1%
18	50.0	76.4	49.5	1.80E+06	16.0%
19	50.0	76.9	53.9	1.80E+06	16.0%
20	50.0	76.3	49.2	1.80E+06	16.0%
21	50.0	76.3	49.2	1.80E+06	16.0%
22	45.0	67.2	44.6	1.60E+06	12.0%
23	45.0	67.1	44.3	1.60E+06	12.0%
24	45.0	67.1	44.6	1.60E+06	20.0%
25	45.0	67.1	44.5	1.60E+06	20.0%
26	45.0	97.6	43.7	2.00E+06	12.0%
27	45.0	97.5	43.6	2.00E+06	12.0%
28	45.0	96.8	45.9	2.00E+06	20.0%
29	45.0	96.8	46.0	2.00E+06	20.0%
30	55.0	55.1	53.8	1.60E+06	12.0%
31	55.0	54.9	53.2	1.60E+06	12.0%
32	55.0	55.0	54.0	1.60E+06	20.0%
33	55.0	54.9	53.5	1.60E+06	20.0%
34	55.0	85.6	54.2	2.00E+06	12.0%
35	55.0	85.5	53.9	2.00E+06	12.0%
36	55.0	85.6	54.2	2.00E+06	20.0%
37	55.0	85.5	54.2	2.00E+06	20.0%
38	40.0	58.3	51.4	1.40E+06	4.0%
39	40.0	57.9	39.4	1.40E+06	4.0%
40	40.0	56.3	44.4	1.40E+06	28.0%
41	40.0	56.2	44.3	1.40E+06	28.0%
42	40.0	118.5	38.8	2.20E+06	4.0%
43	40.0	118.2	37.1	2.20E+06	4.0%
44	40.0	110.0	63.3	2.20E+06	28.0%
45	40.0	110.2	62.4	2.20E+06	28.0%
46	60.0	31.0	66.6	1.40E+06	4.0%
47	60.0	31.0	44.8	1.40E+06	4.0%
48	60.0	30.7	49.5	1.40E+06	28.0%
49	60.0	30.4	48.6	1.40E+06	28.0%
50	60.0	95.4	72.9	2.20E+06	4.0%
51	60.0	94.7	58.3	2.20E+06	4.0%
52	60.0	91.5	68.2	2.20E+06	28.0%
53	60.0	91.5	68.0	2.20E+06	28.0%

VITA

Harold David Miller, Jr. was born in Knoxville, Tennessee on April 20, 1966. He attended elementary and middle schools in Knoxville and graduated from Holston High School in June 1984. The following September he entered the University of Tennessee, Knoxville, and in June 1988 he received a Bachelor of Science degree in Electrical Engineering.

In June 1988 he accepted a position with the Engineering & Construction Division of Eastman Chemical Company's Tennessee Eastman Division facility in Kingsport, Tennessee. In 1989 he began work toward a graduate degree in Chemical Engineering via the University of Tennessee Distance Learning program. He received his Professional Engineering license in 1993, and the Master of Science degree in Chemical Engineering in December 1994.

He is married to the former Kellie Jeane Staley.