Processing of Outliers and Missing Data in Multivariate Manufacturing Data

by

Timothy J. Derksen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1996

Author ....
Department of Electrical Engineering and Computer Science
May 7, 1996

Certified by ....
David H. Staelin
Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
F. R. Morgenthaler
Chairman, Department Committee on Graduate Theses

Eng

Processing of Outliers and Missing Data in Multivariate Manufacturing Data

by

Timothy J. Derksen

Submitted to the Department of Electrical Engineering and Computer Science
on May 7, 1996, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis a strategy for treating outliers and missing data was developed and tested for large, multivariate manufacturing data sets. Three case studies using data from a web process, a batch process, and an assembly process showed the utility of simple visual displays of the data for identifying and characterizing the main features of the data set (including outliers, missing data, drifts, periodicities, excursions, and clustering) during initial exploratory analysis.

The assembly process case study provided data with both partially missing and completely missing observations. Plots of the observation number and variable number for each missing value showed patterns which aided in characterizing the missing data. Filling out the partially missing observations with robust maximum likelihood estimates was an effective precursor to multivariate methods for detecting the main features of the data.

Plots of robust normalized distances from the mean for each observation proved useful for detecting observations which were isolated outliers or members of excursions. Plots showing the principal components and measurement variables on which these observations were extreme gave additional insights which were useful for interpreting the outliers.

Scatter plots of principal components of the data revealed features of the data such as gross outliers, clusters, drifts, and excursions. Comparing plots of standard principal components with plots of robust principal components of the mean and covariance showed that each type of plot highlighted different features in the data.

The most successful treatment for outliers and missing data depended on the main features of the data set learned from the initial exploratory analysis, engineering knowledge of the process, and the type of analysis being done on the data. For example, isolated outliers reflecting poor quality products would be included for process-to-product modelling, but would be removed for inherent variation modelling.

Thesis Supervisor: David H. Staelin
Title: Professor of Electrical Engineering

Acknowledgments

I would like to thank everyone who helped make this thesis possible. I am especially grateful to my advisor Professor David Staelin for his guidance and insights. I would also like to express my gratitude to Elaine Johnson, Peter Sprinz, Marshall Galpern, Mark Strong, Rob Gordon, and Vikas Sharma for their support, suggestions, and help with the three case studies. Thanks also to Professor Roy Welsch, Professor Duane Boning, Mark Rawizza, Dave White, and others who brought up many good issues and ideas at Research Group 4 meetings.

My appreciation also goes to Mike, Bill, Carlos, and Mark for all the help and discussion in the computer lab.

I would also like to thank my family and friends for bearing with me while I concentrated on writing this thesis and for supporting and encouraging me over the years. Finally, I would like to thank the Lord for life and hope.

This research was supported by funds from Research Group 4 of the Leaders for Manufacturing Program at MIT.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context of Research

The research described in this thesis is part of a larger effort of Research Group 4 of the Leaders for Manufacturing (LFM) Program. The LFM Program is a joint effort between leading manufacturing companies and the Massachusetts Institute of Technology. The goal of Research Group 4 is to develop statistical analysis tools for rapidly understanding and improving manufacturing processes and products.

This thesis addresses the issues of outliers and missing data in manufacturing data. Other topics addressed by Research Group 4 include time-series analysis, process-to-product modelling, and real-time mean shift detection.

1.2 Organization of Thesis

This thesis begins with the statement of the problem in Chapter 2. Chapter 3 discusses the strategy for dealing with outliers and missing data in historical data sets. Chapters 4, 5, and 6 give the details of three case studies using manufacturing data sets from three different processes: a web process, an assembly process, and a batch process. Chapter 7 wraps up the thesis with a discussion of the results of the research as well as suggestions for possible future research in this area.

Chapter 2

Problem Description

2.1 Background

Many companies seek to improve the quality of their products and the efficiency of their manufacturing processes by statistically analyzing process and product data. Missing data and outliers are two issues that commonly arise during the data analysis.

2.1.1 Outliers

Definition of Outliers

According to Barnett and Lewis, an **outlier** is “an observation (or subset of observations) which appears to be inconsistent with the remainder of the data” [1, p. 7]. A key point to make here is that outliers are defined relative to the main population—not in terms of possible causes. Seeking a cause for outliers is a separate issue. The following paragraphs define terms related to some of the causes of outliers.

Causes of Outliers

Some outliers may be contaminants. **Contaminants** are those observations which are not “genuine members of the main population” [1, p. 7]. For instance, if the main population was a sample of the heights of women, the height of a man in the sample would be a contaminant whether or not it was distinguishable from the other measurements. Thus, a contaminant need not be an outlier. An example of a contaminant in the manufacturing context is a product made by a broken machine in the midst of products made by a working machine.

Erroneous data is another example of contamination. Erroneous data are those observations with measurement or recording errors. For instance, recording equipment may switch two digits of a measurement. Again, erroneous data may not show up as outliers.

A third possible explanation of outliers is the presence of unusually extreme members of the main population. Most manufacturing processes are very complicated with many factors contributing to the variation of the process. These many factors may occasionally combine to produce an outlier even though the observation is not a contaminant.

Finally, some observations may be declared outliers because the assumed model for the data may be incorrect. For example, a data set assumed to have a normal distribution may have several distinct clusters of observations. A new set of outliers may be defined relative to a new model which assumes each cluster has a normal distribution.

Effects of Outliers

Outliers can cause a statistical analysis to give misleading results. For example, summary statistics such as the sample variance can be greatly inflated by a few extreme values. If these outliers are contaminants, the sample variance will not be an accurate estimate of the variance of the main population.

Outliers can also affect process modelling. An outlying contaminant that is not detected may seriously skew many modelling techniques which seek to minimize the mean-squared error of the model residuals.

In the context of process monitoring, outliers may cause false alarms. If outliers reflecting erroneous data occur frequently, the time required to detect a process excursion may be greater.

Outliers vs. Bad Products

In the manufacturing context, a key subset of the observations are those which correspond to products with unacceptably poor quality (**bad products**). Since a bad product is specified independently of the data set, it is not necessarily an outlier. In fact, the entire main population may consist of bad products while the outliers have high quality.

2.1.2 Missing Data

Definition of Missing Data

The issue of missing data is straightforward—some of the entries in the data matrix are missing. Some observations are **partially missing** while others are **completely missing**. Little and Rubin's *Statistical Analysis with Missing Data* [5] is a good general reference for dealing with partially missing observations.

Causes of Missing Data

Data may be missing for any number of reasons but several common ones include:

- Measurement equipment failed to record a value.
- Data was lost during the storage process.
- The product was pulled from the manufacturing line before measurements were taken.

Effects of Missing Data

The presence of missing data represents a loss of information which can complicate many statistical procedures. Even with modified methods, the results of a statistical analysis may be biased if the mechanism leading to missing data is misunderstood. [5, p. 9]

2.1.3 Manufacturing Data

Types of Manufacturing Data

Manufacturing data are as diverse as the processes from which they come, but the following characteristics are relevant to choosing methods of analysis:

1. Observation Type
    - univariate—one variable per observation
    - multivariate—more than one variable per observation
2. Variable Type(s)
    - measurements of physical parameters such as temperature

- coordinates such as time or location
- categorical variables such as batch or model number
- identification variables such as serial numbers

3. Data Collection Type

    - historical data from regular process operation
    - real-time data from regular process operation
    - data from design of experiments

Most manufacturing data sets have multivariate observations and several types of variables.

Types of Analysis

Statistical analyses of manufacturing data generally fall into one of the following categories:

1. Process-to-Product Modelling

    - specifies relationships between process parameters and product quality
    - based on historical data sets
    - used for determining desirable operating region(s)

2. Inherent Variation Modelling

    - describes the unavoidable variation for a given region of operation
    - based on historical data sets
    - used as the basis for statistical process control

3. Control Settings to Process Modelling

    - specifies relationships between control settings and process parameters
    - based on historical data sets
    - used for understanding and improving process control

4. Statistical Process Control

    - monitors the process for the occurence of unusual events

- based on real-time data
- used to detect process problems and poor quality product

5. Time-Series Analysis
    - describes how the process or product changes over time
    - based on real-time data or historical data sets
    - used to characterize process events and variation

Outliers and missing data may need to be treated differently for these different types of analysis.

2.1.4 Current Approaches

Much research has been done in recent years on analysis methods relevant to manufacturing data including methods for dealing with outliers and missing data. The following paragraphs describe several sources related to outliers and missing data.

Outliers in Statistical Data, 3rd Edition by Barnett and Lewis [1] discusses many statistical methods for the identification and accomodation of outliers including the following:

1. Univariate samples (from several probability distributions)
2. Multivariate samples (from several probability distributions)
3. Linear models
4. Time-series models
5. Directional data

The issue of missing data is not addressed.

Multivariate Analysis by Krzanowski and and Marriott [2] discusses many techniques for multivariate data analysis in general. Several suggestions are given for detecting and treating outliers and missing data.

Little and Rubin's *Statistical Analysis with Missing Data* [5] discusses how to deal with partially missing observations. The issue of potential outliers in additional to partially missing observations is addressed in a paper by Little. [4]

A paper by MacGregor and Kourti [6] gives an overview of methods for implementing statistical process control in multivariate manufacturing processes. Included in the paper are suggestions for treating outliers in process-to-product modelling, historical data analysis, and inherent variation modelling.

2.2 Problem Statement

The problem addressed in this thesis is the development and testing of a strategy for treating outliers and missing data in large, multivariate, manufacturing data. The primary focus will be on the initial exploratory data analysis of historical data sets. The results of applying this strategy will be a better understanding of the nature of the data as well as some protection against the negative effects of outliers and missing data.

Chapter 3

Strategy

3.1 Introduction

According to Krzanowski and Marriott, analysis of data should begin with an initial investigation aimed at identifying the "main features" of the data such as clustering and outliers.[2, p. 43] Other features common in manufacturing data include time-series behavior (drifts, periodicity, excursions), missing data, and skewed distributions. After the features are identified, the main analysis can proceed in an informed manner.

The basic strategy for dealing with outliers and missing data in the initial exploratory analysis of historical manufacturing data sets is shown in Figure 3-1.

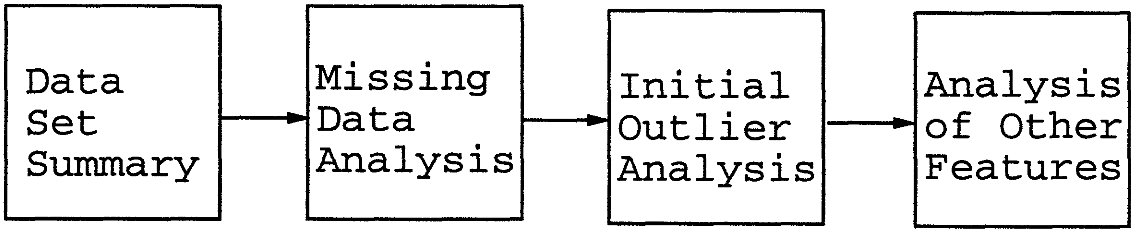

Figure 3-1: Flowchart for the initial exploratory analysis of historical manufacturing data sets.

Each of these steps is discussed in the remainder of this chapter.

3.2 Data Set Summary

A basic understanding of the data set is needed before the initial investigation starts. Useful information to know about the data set includes:

1. Type of process (batch, continuous, etc.).

2. Number of recorded observations.

3. Number and types of variables.

3.3 Missing Data Analysis

3.3.1 Missing Data Detection

Missing data is detected and characterized first since it can complicate the detection of other features. Partially missing observations are usually detected while loading the data, but completely missing observations may be harder to detect. Sometimes completely missing observations can be discovered by scrutinizing identification variables such as serial numbers or plotting the variables versus time.

3.3.2 Missing Data Characterization and Interpretation

Missing data can be characterized in terms of the variable number and observation number for each missing value. Patterns of missing data (such as several consecutive observations missing the same measurements) may emerge. Combining this information with engineering knowledge of the process and measurement system should provide an understanding of the cause of the missing data.

3.3.3 Treatment of Missing Data

The treatment of missing data depends on the type of analysis being conducted and the mechanism causing the missing data. Little and Rubin [5] describe maximum likelihood approaches for data sets with partially missing observations. Other possibilities for dealing with data containing partially missing observations include the following:

1. Remove the partially missing observations.

2. Remove the partially missing variables.

3. Remove a combination of observations and variables such that the number of retained measurements is maximized. An example of a "greedy" algorithm used to pick which observations and variables to remove is the following. For a given number n, remove the n variables with the most missing measurements and then remove the remaining partially observed observations. Find the maximum number of retained measurements as n ranges from 0 to P (the total number of variables).

Completely missing observations can either be ignored (for outlier detection and process-to-product modelling) or filled in with estimates based on earlier and later observations (for time-series modelling). In any case, missing data represents a loss of information, and the causes of missing data should be eliminated if possible.

Estimating the covariance matrix and mean in the presence of partially missing observations is one example of the accomodation of missing data. The solution to this problem is to use the EM algorithm described by Little and Rubin. [5, ch. 7]

The EM algorithm is a two step iterative procedure for obtaining the maximum likelihood estimates of the sample covariance matrix and the sample mean. For an assumed normal distribution, the EM algorithm proceeds as follows:

1. Starting state: initially use estimates based on the completely recorded observations.

2. E: estimate the missing values based on the current estimates of the mean and covariance matrix and keep track of the variance of the estimates.

3. M: compute new maximum likelihood estimates based on the filled-in data set and add corrections to the covariance matrix to take into account the use of estimates.

4. Stopping state: stop when the largest percentage change in any of the parameters is less than some number like 1%.

The method for computing the estimates is least squares estimation based on the recorded values:

$$\hat{\mathbf{x}}_{j,miss} = \bar{\mathbf{x}}_{j,miss} + S_{miss,obs}{S_{obs}}^{-1}(\mathbf{x}_{j,obs} - \bar{\mathbf{x}}_{j,obs})$$

where t is the current iteration. The corrections to the new estimate S are based on the variance of the least squares estimator:

$$S_{error} = S_{miss} - S_{miss,obs} S_{obs}{}^{-1} S'_{miss,obs}$$

Many times the variables in manufacturing data are highly correlated and S_{obs} is singular. In that case, the least squares estimator can be based on a modification of the pseudoinverse. S_{obs} can be decomposed as

$$S_{obs} = Q \Lambda Q^T$$

where Q contains orthonormal eigenvectors of S_{obs} and Λ is a diagonal matrix containing the eigenvalues of S_{obs}. The inverse of S_{obs} is then

$$S_{obs}^{-1} = Q \Lambda^{-1} Q^T$$

When an eigenvalue in Λ is less than a threshold value (say 10^{-12}), the corresponding element in Λ^{-1} is set to zero. This basically amounts to projecting the data onto the subspace spanned by a subset of the principal components, inverting the matrix, and then projecting back to the original space.

3.4 Initial Outlier Analysis

3.4.1 Outlier Detection

If there are N observations with P measurements, each observation could be thought of as a point in P-dimensional space. Thus, potential outliers are those points which "stick out" from the main cluster of points. One way to measure the extremeness of a given observation is to order the observations according to a measure of its distance from the mean. [1, p. 306]

Typical manufacturing data sets contain many different types of measurements with different units and scales. Since multivariate measures of distance are combinations of all the measurement variables, each variable should be scaled to eliminate the effect of the choice of units and to give each variable equal weight. [2, p. 78] Usually, the variables are normalized to unit variance, but there are other possibilities for scaling. For example, if the variance of the inherent random 'noise' for a variable is known, the variable could be normalized to unit noise variance.

One measure of distance commonly used for find outliers in manufacturing data is the univariate number of standard deviations from the mean:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where x_{ij} is the value of variable j for observation i, $\bar{x}_j$ is the sample mean of variable j, and s_j is the sample standard deviation of variable j. All observations with $z_{ij} > K$ for a given constant K lie outside a P-dimensional cube centered at the sample mean. This approach will work fine for identifying gross outliers (just make K large) but will probably be misleading and less effective when the variables are correlated as in most manufacturing data sets.

In terms of a probability model, the measurements from manufacturing data sets are usually assumed to arise from a multivariate normal distribution. A natural measure of extremeness in this case is in terms of P-dimensional ellipsoids of constant probability. An appropriate statistic is

$$d_i^2 = \mathbf{z}_i' R^{-1} \mathbf{z}_i = \sum_{i=1}^{P} \frac{p_{ij}^2}{c_i}$$

where $\mathbf{z}_i$ is the vector of z_{ij} for observation i, R is an estimate of the correlation matrix, p_{ij} is the value of principal component i for observation j, and c_i is the variance of principal component i. All observations with $d_i^2 > K$ for a constant K lie outside an ellipsoid centered at the origin with axes along the principal components determined by R. The statistic d_i^2 can be called the "normalized" distance since it is equivalent to computing the Euclidean distance after the principal components have been normalized to unit variance.

One difficulty with outlier identification is that the very statistics used to detect outliers can be distorted by the outliers. To overcome this problem, outlier-robust estimates $\bar{x}^*$, s^*, and R^* can be used in place of the usual estimates of the mean, standard deviation, and correlation matrix for scaling the data and computing z_{ij}^* and d_i^{*2}.

Barnett and Lewis discuss many possibilities for robust estimates of the mean, variance, and covariance matrix. [1, p. 273-283] In this thesis, robust estimates of the mean and standard deviation were obtained by symmetric 5% trimming for the initial scaling of the variables. This means that the values for each variable were sorted, and the estimates were based on the measurements remaining after the largest 5% and smallest 5% of the values were removed.

Estimates of the correlation matrix can be distorted by the presence of observations which break the correlation pattern of the data in addition to the presence of very extreme values. In this thesis, robust estimates were obtained with a 2 step trimming procedure.

The first step in arriving at the robust estimate R^* is to get an initial estimate R_1^* based on the observations remaining after the gross outliers are removed. Observations with $z_{ij} \geq 10$ are considered gross outliers.

The next step is to detect the observations with $d_i^{*2} \geq K$ (with d_i^{*2} computed using R_1^*) as the next tier of outliers. The values of d_i^{*2} should be plotted for each observation, and K should be chosen to identify the observations with distinctly larger d_i^{*2} than the majority. Ideally the value of K should be chosen based on the probability distribution of d_i^{*2}, but that option is impractical for high-dimensional manufacturing data sets with unknown numbers of outliers and missing measurements. For this thesis, the value of K was chosen based on a visual display of d_i^{*2} for each observation. The final estimate R^* is based on the observations which remain after both the gross outliers and the next tier of outliers are removed.

Another scenario which must be considered is the presence of missing data in addition to the possibility of outliers. Little suggests an approach combining the EM algorithm to accomodate the missing data with a maximum likelihood type estimator (M-estimator) to accomodate potential outliers. [4] However, Barnett and Lewis mention difficulties with M-estimators for high-dimensional data. [1, p. 275-281] For the case studies in this thesis, robust estimates were computed as follows:

1. $\bar{x}^*$ and s^*—estimates based on 5% symmetric trimming of the recorded measurements for each variable.

2. R^*—four step process

   (a) Remove observations with $z_{ij}^* \geq 10$.

   (b) Estimate R_1^* and $\bar{x}^*$ with the EM algorithm discussed in the previous section on missing data analysis.

(c) Remove observations with $d_i^{*2} \geq K$ where K is determined as in the previous paragraph.

(d) Estimate R^* and $\bar{x}^*$ with the EM algorithm.

3.4.2 Outlier Characterization

Observations with extreme values of d_i^{*2} can be characterized in terms of the original variables on which they are extreme if any. A measurement is considered extreme if $z_{ij} \geq 2$. Patterns of outliers may emerge from this information. For example, several outliers may be extreme on the same subset of variables—implying that they may have the same cause. Another example is a group of outliers which have the same value for a categorical variable. For example, the categorical variable could be carrier number, and all products held by a broken carrier could be outliers.

Outliers can also be characterized in terms of the principal components on which they are extreme. An outlier may be extreme on one of the principal components while not being extreme on any of the original variables.

3.4.3 Interpretation of Outliers

Based on knowledge of the specific process under consideration and the outlier characterization, the cause of each extreme value is investigated. Observations which are outlying on many variables probably reflect something happening in the process. Outliers on only one variable may be measurement or recording errors. According to MacGregor and Kourti [6, p. 406], the outliers which are extreme only on principal components with low variances may very well be random "noise".

3.4.4 Treatment of Outliers

The treatment of outliers depends both on the type of analysis and the probable causes of the outliers. The following are several types of analysis and the corresponding treatment of outliers:

1. Process-to-product model

    - remove clearly erroneous observations
    - consider modelling distinct groups of outliers by themselves

2. Inherent variation model

    - remove clearly erroneous observations
    - remove any outliers with an assignable cause
    - remove any outliers thought to correspond to bad product

3. Time-series analysis

    - replace clearly erroneous observations with estimates
    - remove effects of outliers with assignable causes not of interest

3.5 Analyzing Other Features

3.5.1 Detecting other features

Visual inspections of multivariate data general fall into three categories:

1. Plots of original variables.
2. Plots of projections of the data.
3. Plots of computed statistics.

These visual inspections are meant to detect possible features in the data which may then be scrutinized in greater depth.

First, the variables are plotted by observation number or time to get a feel for the nature of the data. Also, the variables can be superimposed on the same graph to reveal features that occur in several variables simultaneously. (The variables may need to be scaled and centered so they can be easily compared.) Several features to look for include outliers, excursions, drifts, periodic time behavior, clustering, and skewed distributions. Two- or three-dimensional scatter plots are also investigated for meaningful combinations of variables such as spatial coordinates.

The principal components are orthogonal projections of the data based on the covariance or correlation matrix. The first principal component is the projection of the data which has the largest variance. The second principal component is the projection with the largest variance subject to the constraint that it must be uncorrelated with the first principal component, and so on.

The following procedure can be used to compute the principal component values for each observation:

1. Decompose the covariance (or correlation) matrix of the measurement variables:

$$S = Q\Lambda Q^T$$

   where Q is an orthonormal matrix with the eigenvectors of S for columns, and Λ is a diagonal matrix containing the eigenvalues of S.

2. Create Q_* by reordering the columns of Q so that the first column of Q_* is the eigenvector corresponding to the largest eigenvalue of S, the second column of Q_* is the eigenvector corresponding to the second largest eigenvalue of S, and so on.

3. Compute the principal component values $\mathbf{p}_i$ from $\mathbf{x}_i$ (the measurements for observation i):

$$\mathbf{p}_i = Q_*^T \mathbf{x}_i$$

The text by Krzanowski and Marriott contains some discussion of principal components in addition to an alternate procedure for computing them. [2, Ch. 4]

Plots of the first few principal components either with respect to time or each other usually show the source(s) of greatest variation in the data. Features typically captured by the first few principal components include excursions, clustering, and drifts.

The definition of the principal components is determined by the correlation matrix. Thus, a robust version of the correlation matrix may be used to get a robust set of principal components. A method of getting such a robust correlation matrix has already been described in the context of outlier detection. The interpretation of the robust principal components is that they capture the variation in the main population while the standard principal components capture the variation in the complete data set. In general, both types of principal components should be computed and any differences should be investigated.

Finding interesting projections of the data is the problem addressed by the field of projection pursuit. [2, p. 92-100] Most of these methods try to find the optimum of some criterion (such as nonnormality) over all possible projections of a given dimension.

Another method which results in projections of the data is partial least squares (PLS). PLS provides projections of the data onto a set of orthonormal vectors much like principal components, but the PLS projections are chosen to best predict the variation in another data set rather than to best explain the variation in the original data set. [6, p. 407]

3.5.2 Feature Characterization

If the data set has clusters, an investigation is undertaken to see if any combination of the categorical variables explains the clusters. For instance, each cluster in data from a batch process may correspond to a different batch number. If the clusters are not explained by the categorical variables, the original variables which show the clustering should be investigated.

Time-series features in manufacturing data may include drifting, periodicity, mean shifts, and excursions. These features are characterized according to the principal components and original variables on which they appear, as well as start and stop times, frequency content, amplitude, shape, and regularity.

If certain variables appear to have a skewed probability distribution, the first step is to understand the type of measurement the variable represents. If the variable is a measurement of a physical quantity such as temperature or pressure, the direction of skewing should be useful in understanding the cause. On the other hand, measurements of length or distance and discrete counts may naturally tend to have skewed distributions. The variables which naturally skewed distributions may be transformed before further processing. [2, p. 59-64]

3.5.3 Treatment of Features

As with the treatment of outliers and missing data, the treatment of the other features in the data depends on the type of analysis being conducted and the interpretation of the features.

For instance, both process-to-product modelling and control settings to process modelling seek to model all excursions, clusters, and outliers which reflect genuine process operating regions. Thus, these features should be retained while outliers due to measurement errors should be removed. On the other hand, inherent variation modelling focuses on one operating region, and all observations from other operating regions should be removed. [6, p. 412]

Time-series analysis usually requires an iterative approach. Gross outliers may need to be replaced with estimates before a drift is characterized. The drift may need to be subtracted from the data before a periodic component can be characterized, and so on.

3.6 Other Approaches

Some other methods besides those discussed in the preceding sections were initially considered but were not pursued in depth. The two primary methods in this category were ordered interpoint distances and Andrews' curves [2].

The ordered interpoint distances methods represents each observation with a curve based on the ordered distances between the given observation and all the other observations in the data set. Thus, the value of the curve at point n for a given observation is the Euclidean distance between that observation and its nth nearest neighbor. The motivation for using ordered interpoint distances plots was to detect outliers and clustering in the data. The reason this method was not pursued was that the robust normalized distance from the mean was more effective for detecting outliers, and scatter plots of principal components were more effective at detecting clusters.

Andrews' curves also represent each observation with a curve. The value of each point of the curve can be thought of as a 1-dimensional projection of the data. [2] Andrews' curves were not pursued because there was no reason to suppose that the projections given by the Andrews' curves would reveal the main features in the data better than the 1-dimensional projections given by the principal components.

3.7 Summary

The initial data analysis of a historical data set provides information about the main features of the data. This information serves as a springboard for further investigation of the data. Features of interest include outliers, missing data, time-series behavior, clustering, and variables with skewed probability distributions.

Chapter 4

Web Process Case Study

4.1 Data Set Summary

The data for this case study came from a continuous web process for producing long sheets of product packaged as rolls. The first data set contains data from a scanner which recorded information about spot defects in the product. The second data set contains in-line settings and measurements. Both data sets were recorded over the same time period.

The data sets are summarized in Table 4.1. For the scanner data, each observation corresponds to one defect detected by the scanner, and the two categorical variables are the roll number and defect type. The identification variable for the in-line measurements specifies the observation number, and some of the in-line measurement variables are switch settings or statistics computed from the measurements.

Table 4.1: Web Data Summary

Data Set Name	Observations	Variables				
		ID	Categorical	Time	Location	Measurement
Scanner	14961	0	2	1	2	4
In-line	2880	1	0	1	0	359

4.2 Scanner Data

4.2.1 Missing Data Analysis

The scanner data did not contain any partially missing observations. Completely missing observations were practically impossible to detect since the observations were not recorded regularly but rather whenever a defect arrived. Defects were recorded only while a roll was being produced so there are some relatively long time intervals with no recorded defects.

4.2.2 Initial Outlier Analysis

Outlier Detection

Figure 4-1 shows the robust normalized distance d_i^{*2} (discussed in Chapter 3) for the measurement variables of each observation. A large excursion in the middle of the data is clearly visible. Another smaller excursion occurs after the large one.

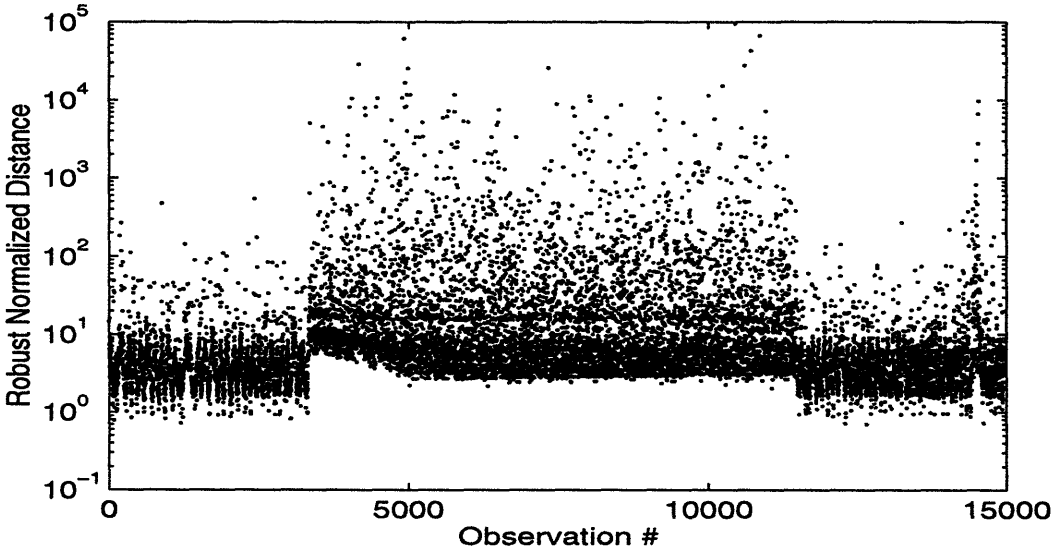

Figure 4-1: Robust normalized distances for the scanner data.

Outlier Characterization and Interpretation

Investigation of the observations with the largest values of d_i^{*2} revealed that the majority of them were of the same defect type. It turned out that part of the definition of the defect types related to the size of the defect. Since the measurement variables predominantly measured defect size, it was not surprising that the observations with the largest normalized distance from the mean were mostly of the largest defect type. However, it was interesting that the largest defects seemed to appear in only two sections of the data instead of sprinkled throughout the data set.

Treatment of Outliers

Since the nature of the excursions in Figure 4-1 was known to be related to size, all the observations were kept and investigated for further features.

4.2.3 Analyzing Other Features

Detecting Other Features

Since there were only nine variables, each variable was individually plotted. One feature which surfaced from this initial investigation was that the majority of the defects (over 8000) were from roll 29. Another feature was that defects tended to occur in the same locations across the web, forming streaks down the web (see Figure 4-2). Note also the streaks across the web near times 30 and 58.

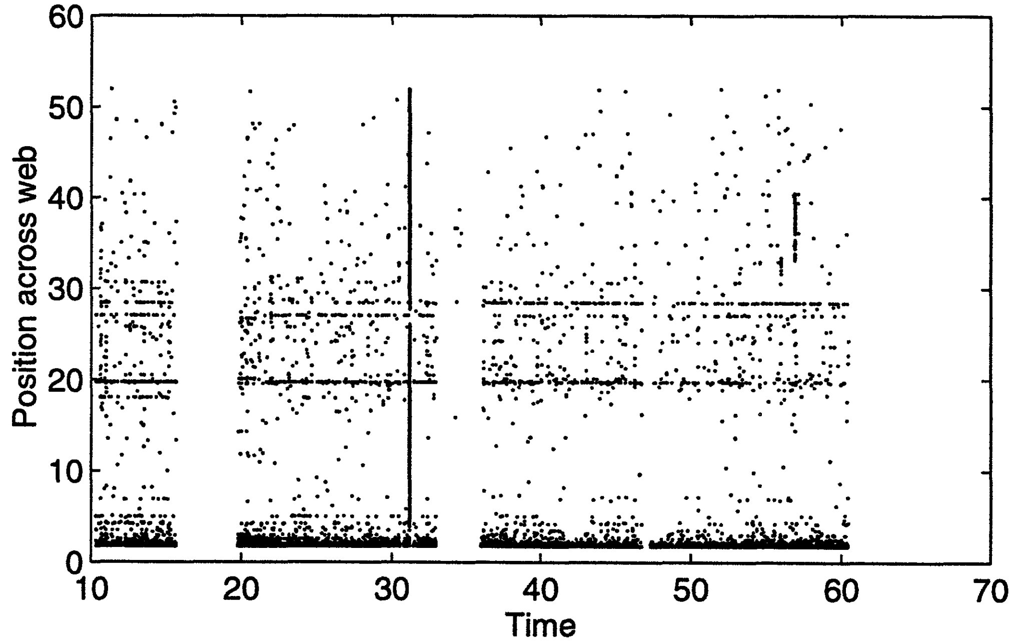

Figure 4-2: Defect location across the web versus time.

Plotting principal components failed to reveal additional features. The first principal component revealed basically the same information as the three highly correlated measurements of size, while the second principal component showed the position across the web.

Feature Characterization

First, the relatively large number of defects in roll 29 were investigated. The initial problem was to determine where and when the defects occurred. To that end, the defects were plotted by position across the web versus time as shown in Figure 4-3. This plot revealed several evenly spaced streaks across the web at the end of roll 29 with several defects sprinkled between the streaks.

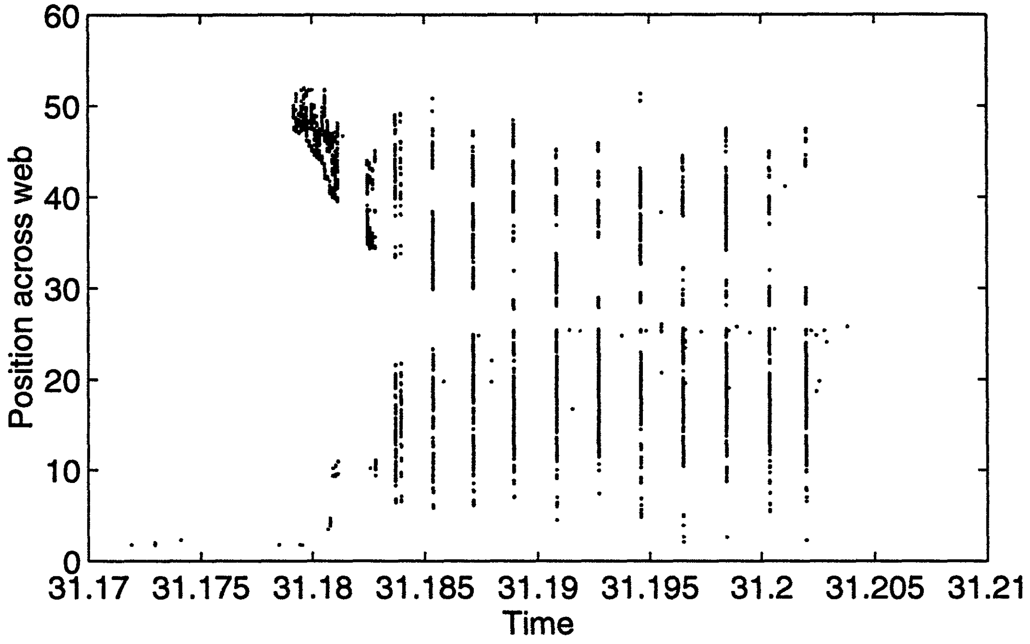

Figure 4-3: Defects at the end of roll 29.

The next problem was to understand the nature of the defects in the streaks across the web in roll 29. Which types of defects occured in these streaks? Table 4.2 shows the distribution of the defects by defect type. Further investigation showed that the defects sprinkled between the streaks were of types 44 and 45. Also, defects of types 42, 43, 46, and 94 were found in each streak and were not confined to specific positions across the web.

Table 4.2: Streaks in Roll 29

Defect Type	35	42	43	44	45	46	47
Number of Defects	1	2354	902	23	22	3479	1362

The next feature to characterize was the streaking along the web corresponding to the defects which occurred at the same position across the web. Plotting the positions of the defects across the web versus time showed that these streaks continued throughout the data collection time period. A count of the number of defects in each streak revealed that there were ten streaks with more than 100 defects (one streak had 1449 defects) and another six with more than 50 defects.

The next step was to investigate the streaks individually. There did not appear to be any pattern in the arrivals of the defects in a given streak, and a χ^2 goodness-of-fit test [3, p. 138-144] was performed to see if the number of defect arrivals in disjoint time intervals followed a Poisson distribution. An acceptance of the Poisson distribution hypothesis would imply that the defects arrived randomly in the sense that the number of defects arriving in a given time interval could not be predicted. The results were inconclusive as shown in Table 4.3.

Table 4.3: Test of Poisson distribution: streak along the web

Location of Streak	Time Interval	Number of Intervals	Acceptance/Rejection of Poisson distribution at a 1% level of significance
1.7750	T	2516	rejected
1.7750	$5T$	501	accepted
1.7750	$10T$	214	accepted

The types of defects in the streaks were also considered. It was found that the streaks along the web consisted primarily of defect types 44 and 45 with one or two defects of other types included occasionally. Were the defect types 44 and 45 related in some way? In each of the streaks with the five largest number of defects, there were over twice as many defects of type 44 than those of type 45 so there was no one-to-one correspondence. Additionally, plots of defect type versus time revealed no repeated patterns suggesting that one type of defect followed the other.

Interpretation and Treatment of Features

Over 80% of the defects are accounted for by the streaks across the web at the end of roll 29 and the ten largest streaks along the web. Possible explanations for these streaks include protrusions on rollers or drips from machinery. These features are treated as the primary focus of the data analysis since the goal of looking at this data was to understand the causes for the defects.

4.3 In-line Data

4.3.1 Missing Data Analysis

The in-line data had no partially missing observations. Checking the sequence of sample numbers revealed that there were no completely missing observations either.

4.3.2 Initial Outlier Analysis

Outlier Detection

Before computing the distance from the mean, the 29 measurement variables which were a constant value throughout the time period were removed. These constant variables added no information about the variation in the data, but they did make the correlation matrix singular.

Figure 4-4 shows the normalized distance from the mean, d_i^2, for each observation. Two excursions are the most prominent features in this plot. Also, there are several isolated points which "stick out" as well as an outlier "cloud" which sticks above the main population.

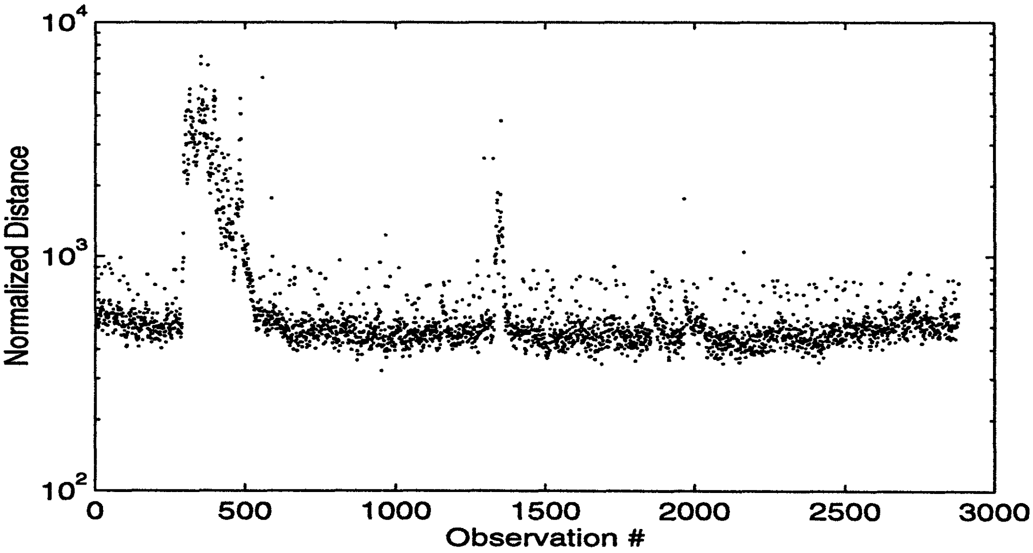

Figure 4-4: Multivariate distance from mean.

Figure 4-5 shows the robust normalized distance from the mean, d_i^{*2}, for each observation. This plot shows the two excursions as well as three very extreme outliers. It turned out that these three observations contained the only three measurements which were different on an otherwise constant variable.

The majority of the values for the robust normalized distance from the mean are about 200 while the majority of the values for the standard normalized distance from the mean are over 400. This occurred because the large excursion moved the estimate of the sample mean away from the main population. This example shows the benefits of robust analysis—the removal of the effects of a few extreme samples.

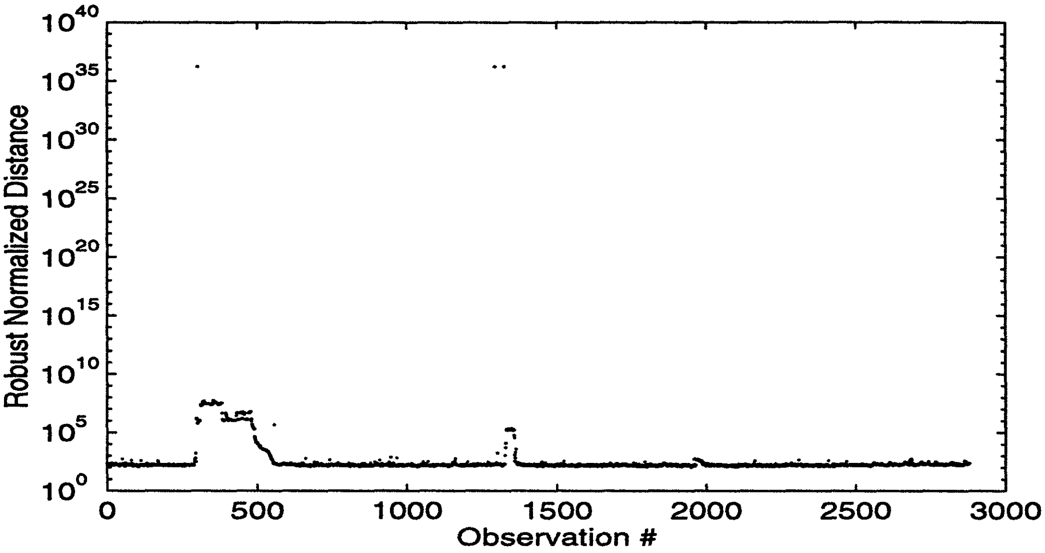

Figure 4-5: Robust multivariate distance from mean.

Outlier Characterization

Figure 4-6 and Figure 4-7 compare d_i^2 with d_i^{*2}. The circled points in both plots are the same observations and so are the asterisks in both plots.

The observations flagged by d_i^{*2} occur near the edges of excursions while the observations flagged by d_i^2 seem to appear randomly throughout the time interval. Thus, the two statistics give different results. The outliers at the edges of excursions have a straightforward explanation, but the outliers found using d_i^2 need further investigation.

One way to characterize outlying observations is in terms of the principal components on which they have extreme values. Figure 4-8 shows the principal component values greater than two standard deviations from the mean for the observations with $d_i^2 \geq 650$.

Principal component 321 seems to have an unusually large number of outlying values. Also, the outlying values on principal component 321 seem to correspond to the outlier

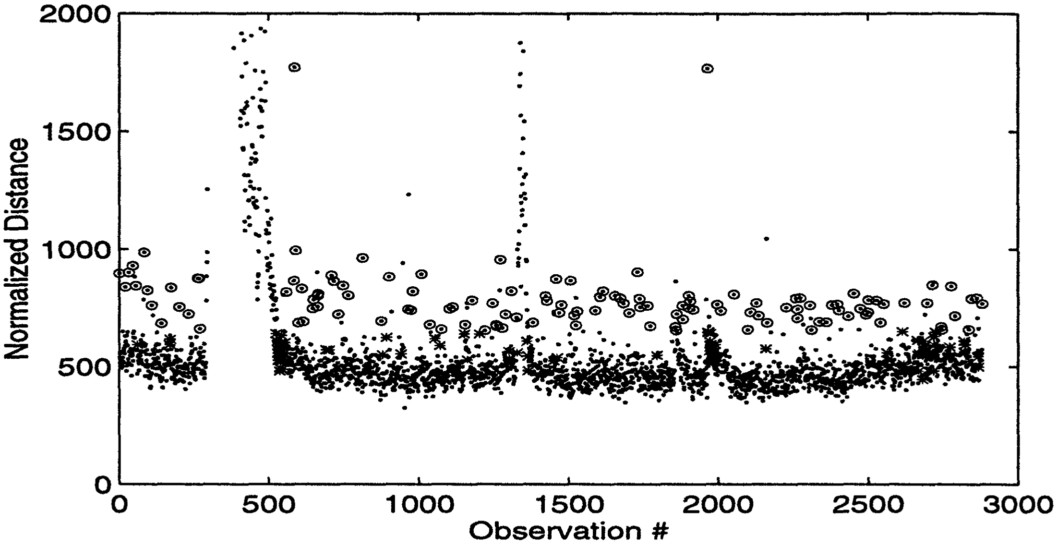

Figure 4-6: Normalized distance from the mean. The circled observations have $d_i^2 \geq 650$ but $d_i^{*2} < 250$. Asterisks have $d_i^{*2} \geq 250$ but $d_i^2 < 650$.

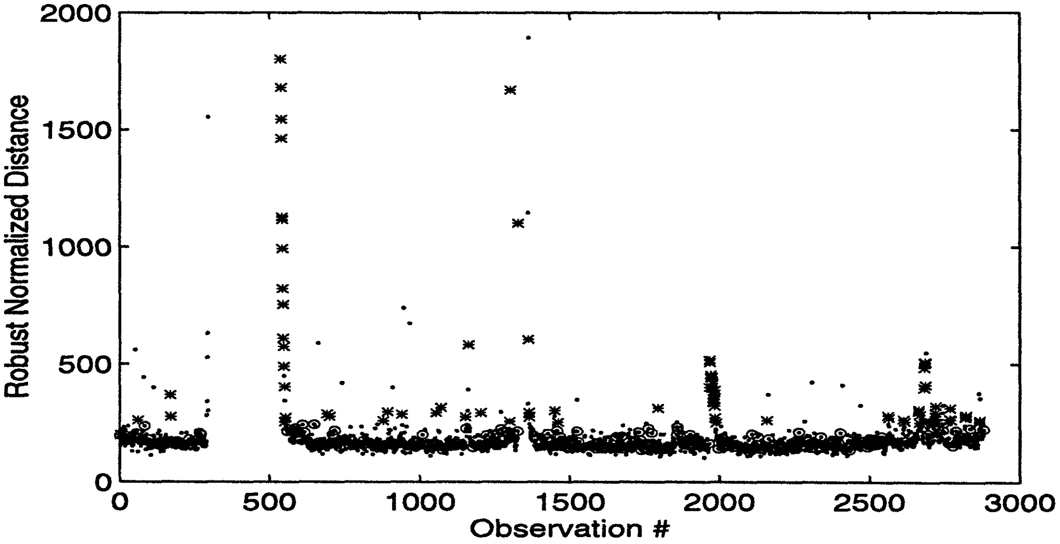

Figure 4-7: Robust normalized distance from the mean. The circled observations have $d_i^2 \geq 650$ but $d_i^{*2} < 250$. Asterisks have $d_i^{*2} \geq 250$ but $d_i^2 < 650$.

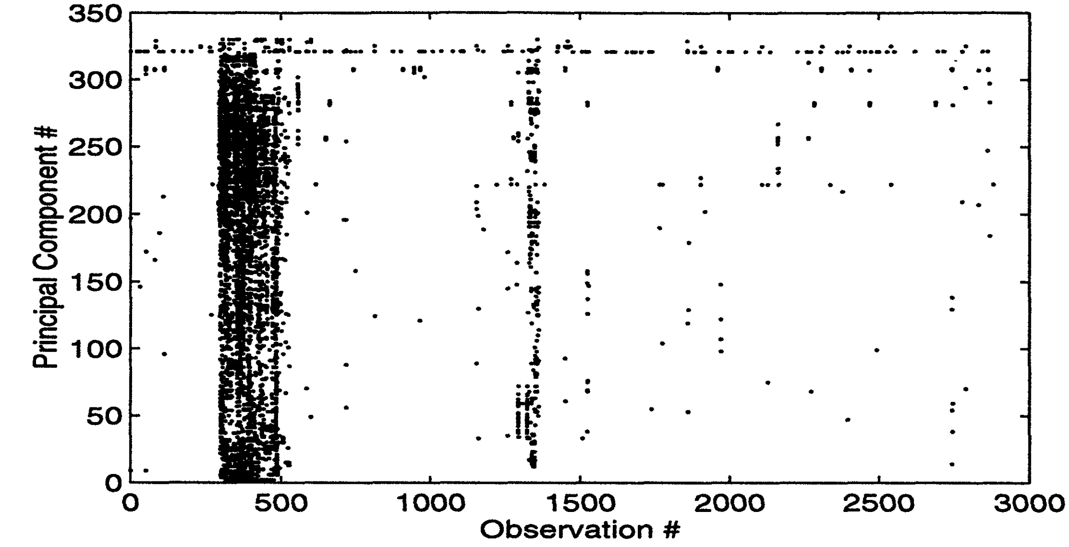

Figure 4-8: Principal component values more than 2 standard deviations from the mean for the observations with $d_i^2 \geq 650$.

cloud under investigation. Figure 4-9 shows principal component 321. The circled points specify the same observations as the circled points in Figure 4-6 and Figure 4-7. This one principal component seems to explain the majority of the discrepancy between d_i^2 and d_i^{*2}.

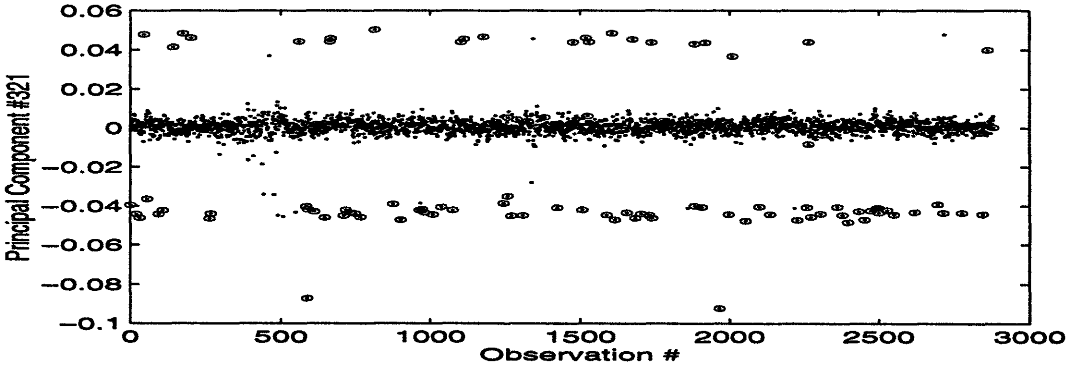

Figure 4-9: Principal component 321. The circled observations have $d_i^2 \geq 650$ but $d_i^{*2} < 250$.

Each principal component is a linear combination of the measurement variables, so one way to interpret a principal component is in terms of the variables with the largest weighting. The loadings of each variable for principal component 321 are shown in Figure 4-10.

The dominant variables are variables 12 and 14—which are highly correlated. Thus, d_i^2 picks out observations whose measurements on variables 12 and 14 are slightly different while d_i^{*2} does not.

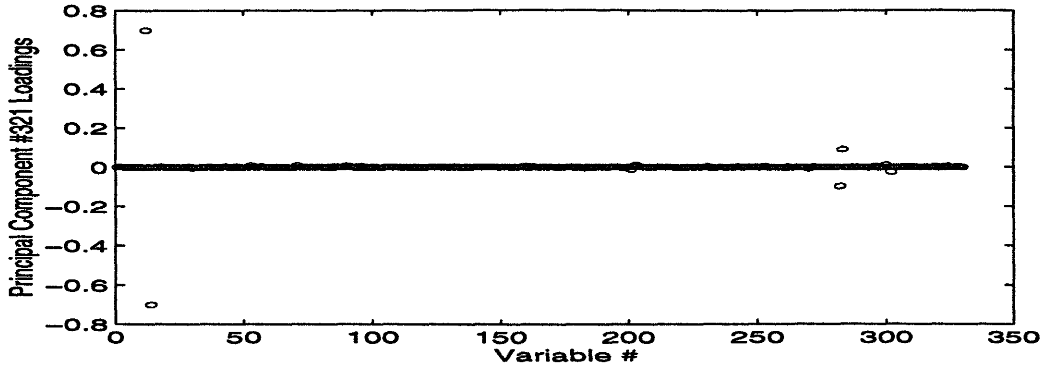

Figure 4-10: Principal component loadings.

Figure 4-11 shows the measurements more than two standard deviations from the mean for observations with $d_i^2 \geq 650$. The large excursion prior to observation 500 occurs on the majority of the variables while the smaller excursion near observation 1300 occurs on relatively few variables. The number of variables on which a feature occurs serves as an upper bound to the dimensionality of that feature.

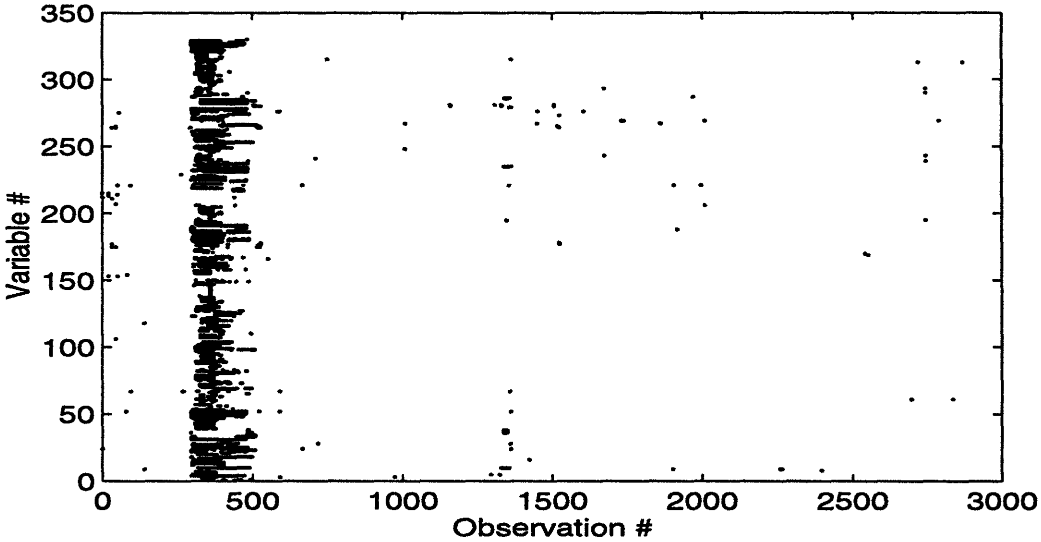

Figure 4-11: Measurements more than 2 standard deviations from the mean for the observations with $d_i^2 \geq 650$.

Figure 4-12 shows the measurements greater than two robust standard deviations from the mean for the observations with $d_i^{*2} \geq 250$. This plot shows many more outlying measurements than Figure 4-11. The reason for this is that the larger excursion masked other features by inflating the standard deviation estimates. Thus, the robust statistics seem to give a clearer picture of the nature of the outlying measurements in this case.

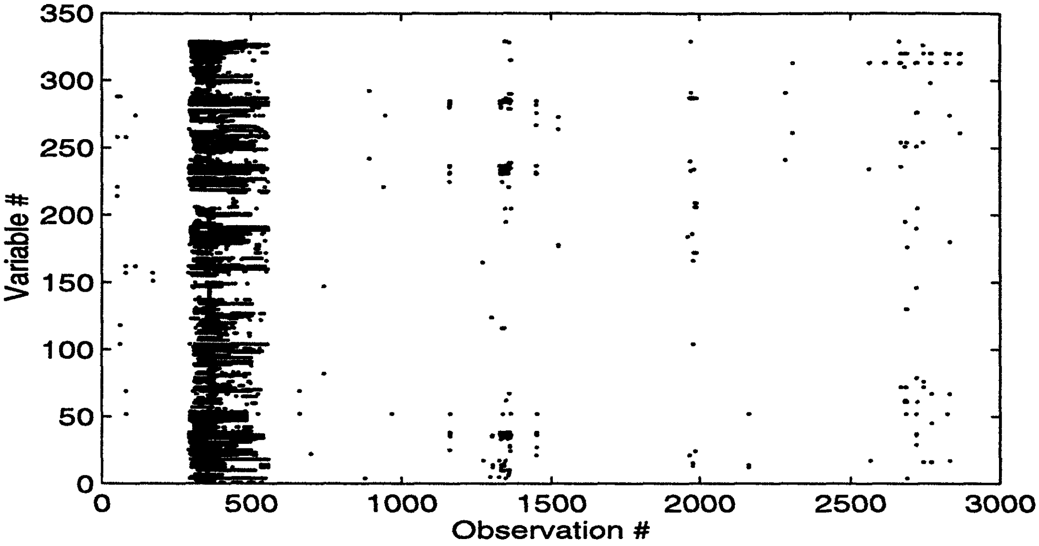

Figure 4-12: Measurements more than 2 standard deviations from the mean for the observations with $d_i^{*2} \geq 250$.

Interpretation of Outliers

Knowledge of the process revealed that the excursions had an assignable cause. It turned out that the excursions were more of a symptom of process maintenance than a prediction of process problems. Engineering knowledge about variables 12 and 14 should determine whether the outlier cloud in Figure 4-4 is meaningful or whether it represents random "noise".

Treatment of Outliers

Since the postulated cause of the excursions is unrelated to normal process operation, probably the best thing to do is to remove the excursions before proceeding with any further analysis. The outliers due to differences between variables 12 and 14 should probably be treated as valid observations if it was determined that they represent random noise.

4.3.3 Analyzing Other Features

Detecting Other Features

The principal components showed definite time-series behavior. For example, Figure 4-13 shows the first principal component values for each observation. A drift is clearly visible in addition to a large excursion.

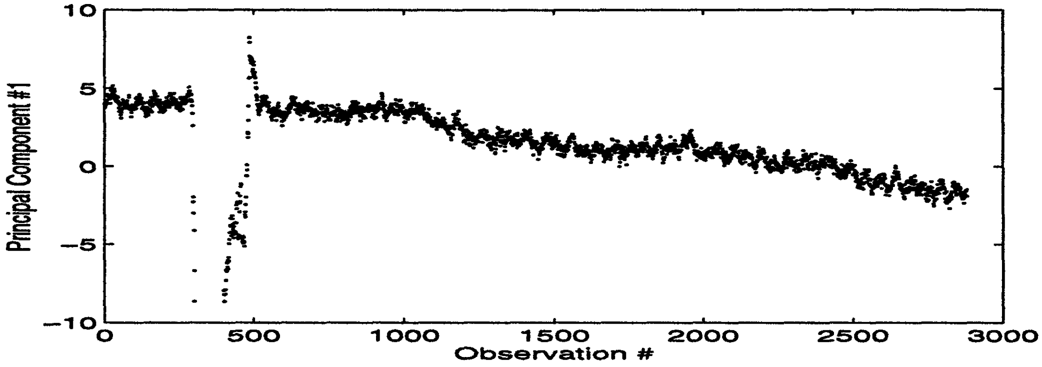

Figure 4-13: First principal component.

Figure 4-14 shows the robust first principal component values. The drift is much more pronounced in the robust first principal component than in the standard first principal component. The reason for this is that the large excursion dominates the first several regular principal components while the robust principal components are based effectively on the remainder of the data set after the large excursion is removed. Again the benefits of robust analysis are evident.

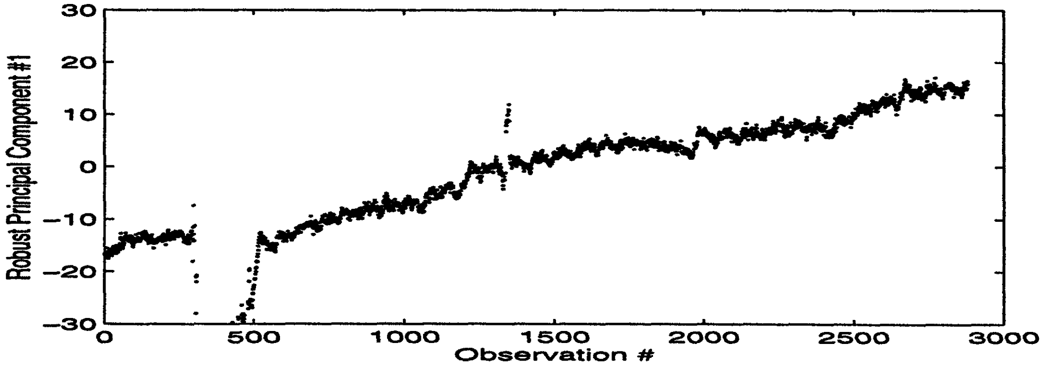

Figure 4-14: Robust first principal component.

Feature Characterization and Interpretation

The time-series behavior of the principal components can be characterized in terms of the variables with the highest weights in determining the particular principal component. These variables can then be plotted for visual inspection by someone familiar with the process. Further steps require a time-series analysis.

Treatment of Features

The treatment of the time-series nature of the data depends on the type of analysis being conducted. Inherent variation modelling would typically require that drifts and periodicities be removed. Process-to-product modelling may call for no special treatment of the time-series features.

Since the goal for this particular data set was to explore the nature of the web defects, the features in the in-line data were compared (in terms of beginning and ending times) with the patterns of the defects in the scanner data. The large excursions were found to coincide with time periods when the scanner was turned off, but no correspondence between the in-line features and the patterns in the defects was found.

4.4 Discussion

The initial analysis of the scanner data showed definite patterns in the web defects—streaks along the web and across the web. Identifying and eliminating the causes of these patterns would yield a great decrease in the number of defects.

Although no relationship was found between the features in the in-line measurements and the patterns in the scanner data, the features in the in-line measurements may be related to other quality measurements. A follow-on analysis should explore this possibility.

Chapter 5

Assembly Process Case Study

5.1 Data Set Summary

The data from this case study came from the split assembly process shown in Figure 5-1. Dimensional measurements are taken on the two components (B and C) and on the final product (A). The data sets include measurements for two product types (P1 and P2), each produced on split assembly processes.

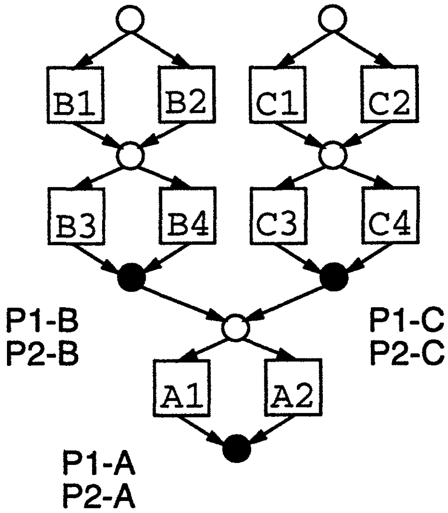

Figure 5-1: Assembly process diagram

A description of the data sets is given in Table 5.1. The categorical variables specify

Table 5.1: Assembly Data Summary

Data Set Name	Recorded Observations	Identification Variables	Time Variables	Categorical Variables	Measurement Variables
P1-C	1175	1	6	4	12
P1-B	1231	1	6	4	13
P1-A	1229	1	6	3	75
P2-C	1241	1	6	4	14
P2-B	1301	1	6	4	15
P2-A	1300	1	6	3	69

Table 5.2: Missing Data Summary—Assembly Data

Data Set Name	Completely Missing Observations	Partially Missing Observations	Completely Recorded Observations	Total Observations
P1-C	57	668	507	1232
P1-B	1	590	641	1232
P1-A	3	658	571	1232
P2-C	62	104	1137	1303
P2-B	0	473	830	1303
P2-A	3	737	563	1303

the path (the combination of machines such as B1-B3 or B1-B4) through the split process as well as model number and carrier number.

5.2 Missing Data Analysis

5.2.1 Missing Data Detection

The initial process of loading the data revealed many partially missing observations. In addition, a comparison of the identification numbers for the three data sets related to a given product (such as P1-A, P1-B, and P1-C) revealed several completely missing observations. In addition, several clearly erroneous observations (from a different time period and randomly inserted in the data set) were removed at the outset.

5.2.2 Missing Data Characterization

Table 5.2 summarizes the missing data in each data set in terms of the number of completely missing, completely recorded, and partially missing observations.

As an example, Figure 5-2 shows the missing measurements for the P1-C data in terms of the specific variables and observations which had missing values. One interesting pattern is that many pairs of variables are missing from exactly the same observations (variables 17 and 18, variables 14 and 15, and variables 12 and 13). Also note that many of the completely missing observations are consecutive.

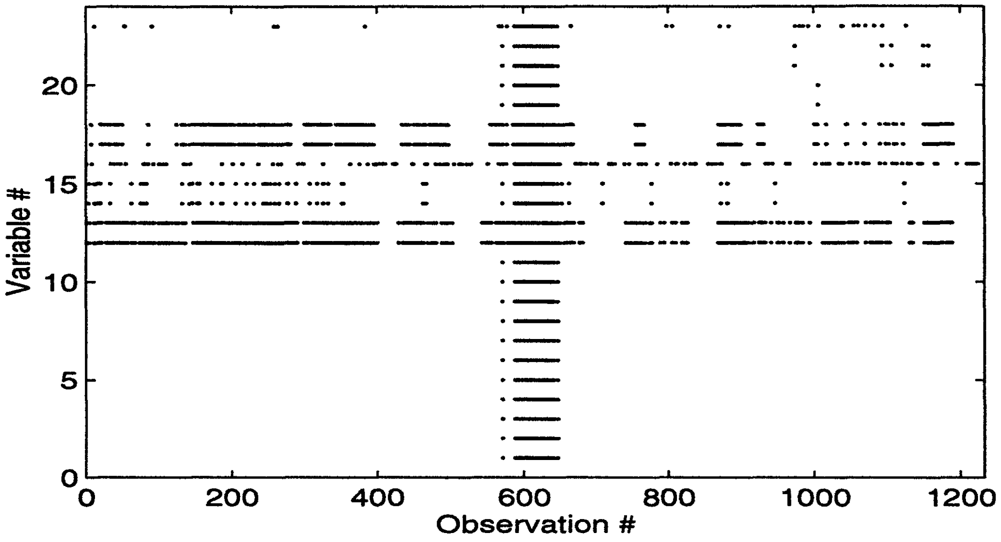

Figure 5-2: Missing data summary for P1-C data.

5.2.3 Missing Data Interpretation

The following are two causes of partially recorded observations:

1. Misalignment of the component or product in the measurement device. (This can explain the pairs of variables missing on the same observations. One variable could be horizontal displacement and the other vertical displace of a particular product feature.)

2. Garbling two observations during the data recording and storage process.

The first cause was revealed by asking someone familiar with the process, and the second cause was observed when loading the data. Both causes would seem to imply that the recorded measurements are a representative sample.

Discovering the cause of the completely missing observations also requires knowledge of the process and measurement system. The most likely guess is that the data was lost in the recording and measurement system since final product measurements (the A data) were recorded for most of the completely missing observations in the C data.

5.2.4 Treatment of Missing Data

The treatment of missing data depends on the which type of analysis is being done. Some suggestions for the following types of analysis include:

1. Process-to-product modelling—fill in estimates for the partially missing observations and ignore the completely missing observations.

2. Inherent variation modelling—same procedure as process-to-product modelling.

3. Time-series analysis—fill in estimates for partially missing observations. The P1-C and P2-C data have many consecutive completely missing observations so the data may need to be split into two time intervals corresponding to before and after the missing observations.

5.3 Initial Outlier Analysis

5.3.1 Outlier Detection

Figures 5-3, 5-4, and 5-5 show values of the robust normalized distance from the mean d_i^{*2} for the P2-A, P2-B, and P2-C data, respectively. The statistic d_i^{*2} for each observation was computed as described in Chapter 3 using $K = 250$ for the P2-A data and $K = 60$ for the P2-B and P2-C data. The circled data points in each figure specify the most extreme observations in the P2-A. In this way, the question of whether good parts (B and C data) can make bad product (A data) or bad parts can make good product is addressed.

5.3.2 Outlier Characterization and Interpretation

In addition to describing the outliers in terms of the variables on which they have extreme measurements, the outliers can be characterized in terms of the data set(s) (A, B, or C) in which they were identified. Thus, observation number 1217 may have a large value of d_i^{*2} in data P2-C but a normal value in P2-A.

The figures showing d_i^{*2} for the P2 data seem to indicate that bad parts can indeed make a good product and good parts can make a bad product. Two other possible explanations are the following:

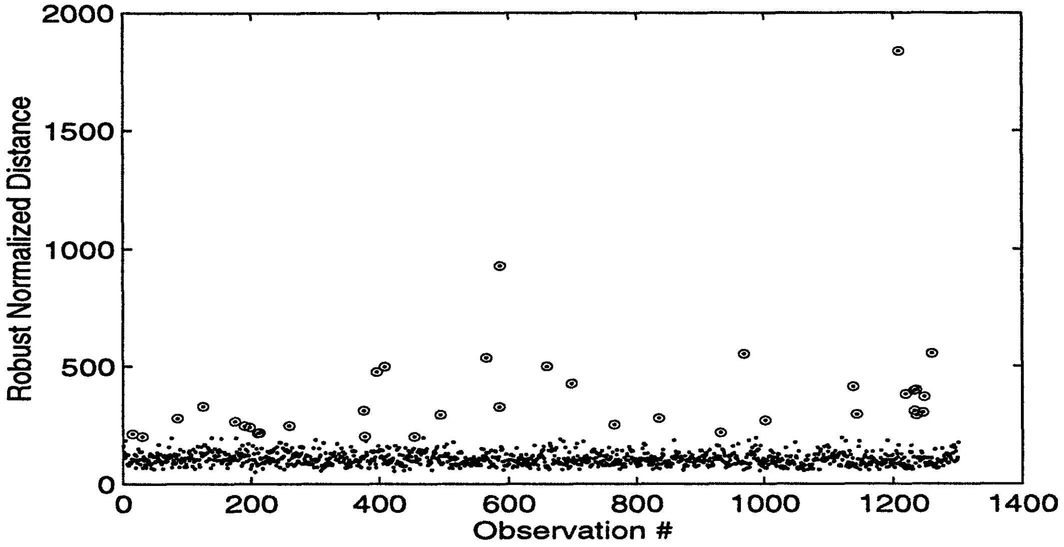

Figure 5-3: Robust multivariate distance from the mean for the P2-A data. Circled points are the observations with $d_i^{*2} \geq 200$ for the P2-A data.

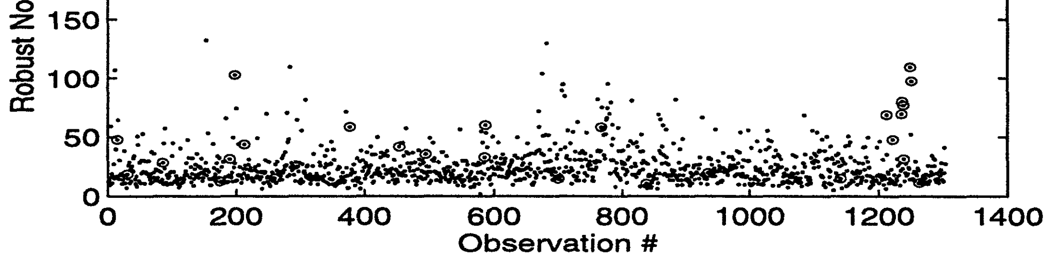

Figure 5-4: Robust multivariate distance from the mean for the P2-B data. Circled points are the observations with $d_i^{*2} \geq 200$ for the P2-A data.

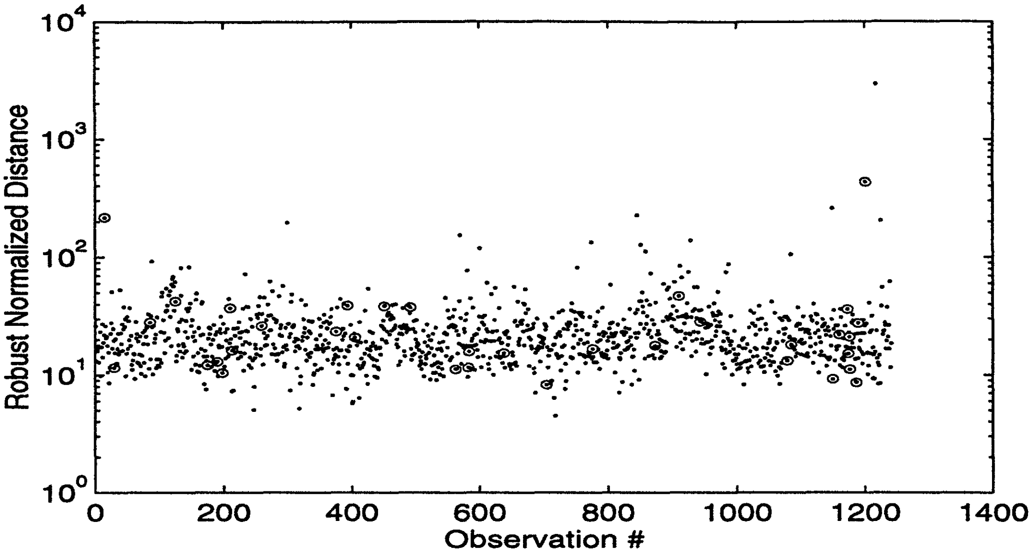

Figure 5-5: Robust multivariate distance from the mean for the P2-C data. Circled points are the observations with $d_i^{*2} \geq 200$ for the P2-A data.

1. The variables measured for P2-A are unrelated to the variables measured for P2-B and P2-C.

2. The outliers reflect errors in measurement or recording.

Knowledge of the particular process and measurement system is needed to completely determine the nature of the outliers.

5.3.3 Outlier Treatment

The treatment of the outliers depends on the subsequent data analysis. Outliers with assignable causes can be addressed separately from the main population for process-to-product modelling, removed or downweighted before inherent variation modelling, or replaced with estimates for time-series analysis. Gross outliers which likely correspond to erroneous measurements should generally be removed before further analysis.

5.4 Analysis of Other Features

5.4.1 Detecting Other Features

Figures 5-6, 5-7, and 5-8 show the first and second robust principal components for the P2-A, P2-B, and P2-C data, respectively. Each scatter plot clearly shows clustering in the data.

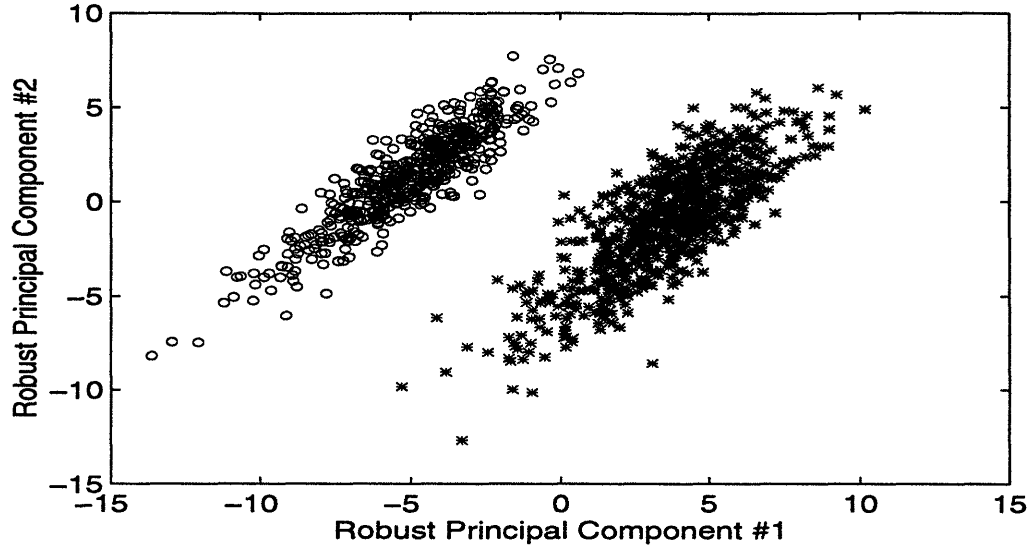

Figure 5-6: First two principal components of the P2-A data. The circles are from path A1 while the asterisks are from A2.

Figure 5-6 shows distinct clustering between the process paths.

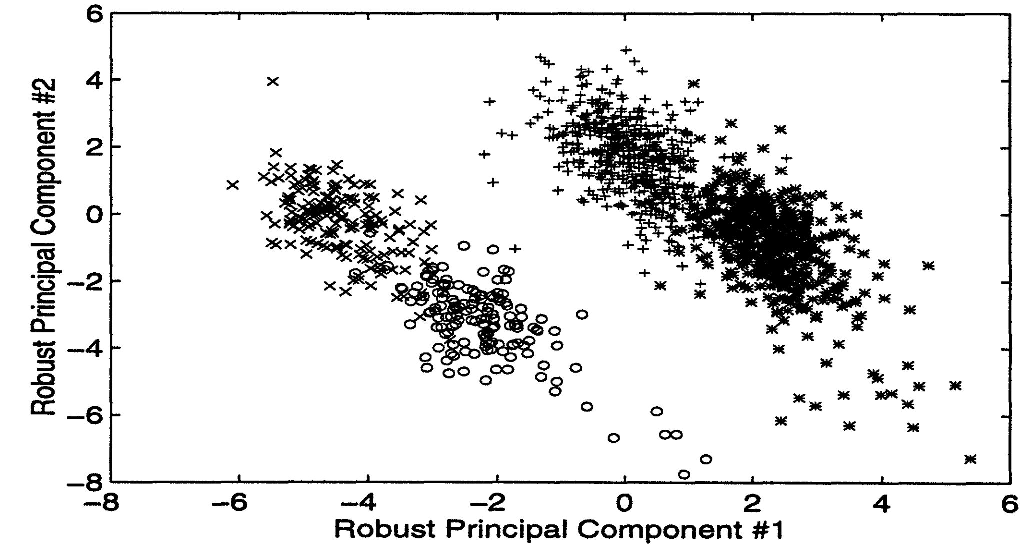

Figure 5-7: First two principal components of the P2-B data: 'o' from path B1-B3, '*' from B1-B4, 'x' from B2-B3, and '+' from B2-B4.

Figure 5-7 also shows distinct clustering between the process paths. However, the clustering is more distinct between process steps 3 and 4 than between process steps 1 and 2. This agrees with intuition because steps 3 and 4 come after steps 1 and 2.

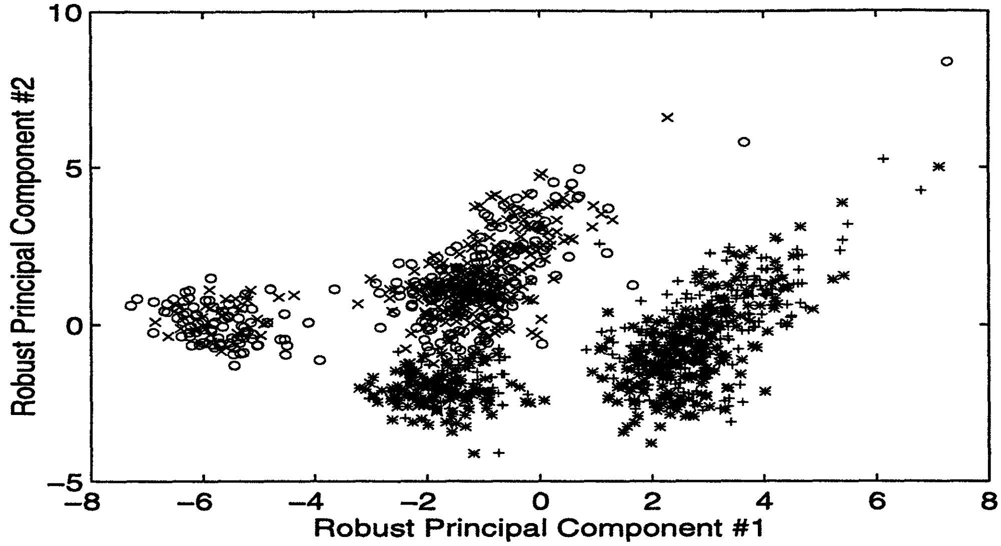

Figure 5-8: First two principal components of the P2-C data: 'o' from path Cl-C3, '*' from Cl-C4, 'x' from C2-C3, and '+' from C2-C4.

Figure 5-8 shows clusters based on process steps C3 and C4 but not on steps Cl and C2. None of the categorical variables could explain this clustering so one conclusion is that a relevant variable has been left out of the data set. In order to see if steps Cl and C2 could be separated, a linear regression was performed against a vector that had ones where the product went through Cl and zeros when it went through C2. Figure 5-9 shows that the resulting linear combination of the variables does indeed show clustering based on Cl and C2.

5.4.2 Feature Characterization and Interpretation

As has already been mentioned, the clusters in all the data sets except P2-C corresponded to the process paths. The clusters in P2-C showed a separation between steps C3 and C4, but also showed clustering which could not be explained with the categorical variables included in the data set. This mysterious clustering came from measurement variables 19 and 20.

5.4.3 Feature Treatment

One logical treatment of the clusters is to address each cluster individually. Thus, a process-to-product model may be different for each process path. Another possible treatment of

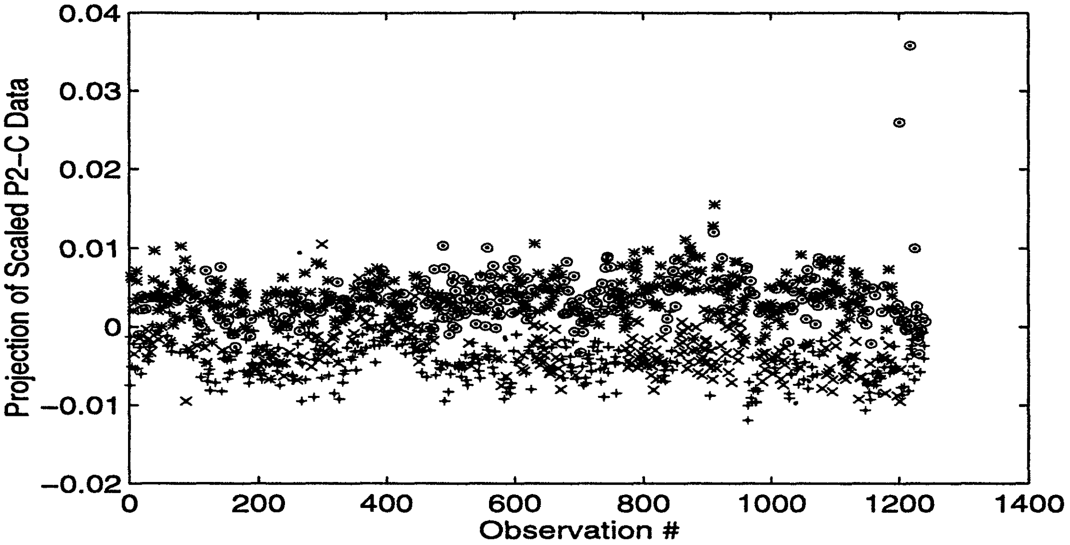

Figure 5-9: One dimensional projection of the P2-C data which shows clustering based on C1 and C2: 'o' from path C1-C3, '*' from C1-C4, 'x' from C2-C3, and '+' from C2-C4.

clusters is to center each cluster at a common point—in effect combining all the clusters into one.

5.5 Discussion

The preliminary analysis described in the previous sections provides information about the assembly process which can be used as a basis for further analysis. The next few paragraphs discuss the implications of the missing data, outliers, and clustering.

All of the data sets had missing measurements. An investigation of the measurement and recording procedures may lead to an understanding of the causes for the partially and completely missing observations. Improvements in the measurement and recording procedures to reduce the amount of missing data would increase the quality of the data analysis because the missing measurements are a loss of information.

One primary question about the outliers is whether or not they correspond to "bad" products. If they do represent bad products, the outliers can be used to detect problems with the process. On the other hand, if the outliers represent measurement or recording errors then they can be used to evaluate the measurement and recording procedures.

The clusters in the data represent a major source of variation in the product. In all the data sets except P2-C, each cluster reflected on of the possible process paths. Investigating the measurement variables which best show the differences between clusters would be one step toward reducing this variation.

If the variation between clusters is acceptable, the clusters can be used to provide information about the individual process paths. For instance, a problem with a specific process machine would only affect some of the process paths. Monitoring each process path separately could lead to the identification of the broken machine from post-process measurements.

Chapter 6

Batch Process Case Study

6.1 Data Set Description

This case study deals with a data set containing end-of-line measurements on individual products which are processed in batches. Each observation contains 60 measurement variables, a categorical variable (specifying the batch number), and an identification variable (specifying the product within a given batch).

The measurement variables are divided into 2 groups. Two of the measurement variables are critical to product quality so these variables can be called "output" variables. The relationship between the output variables and the other 58 measurement variables ("input" variables) is of interest to the company.

6.2 Missing Data Analysis

While examining the identification variable, it was noted that some observations were completely missing. Table 6.1 summarizes the missing data. It turned out that the cause for the missing observations was the removal of very poor quality products before they reached the end-of-line measurement stage. Since the primary interest in the data was the relationship between input and output measurement variables, the analysis used the recorded observations and ignored that fact that some observations were missing.

Table 6.1: Missing Data Summary—Batch Process

Completely Missing Observations	Partially Missing Observations	Completely Recorded Observations	Total Observations
85	0	2465	2550

6.3 Initial Outlier Analysis

6.3.1 Outlier Detection

Figure 6-1 shows the normalized distance from the mean d_i^2 for the input measurements. (Chapter 3 contains a discussion on the normalized distance from the mean.) This plot shows several gross outliers as well as an outlying batch.

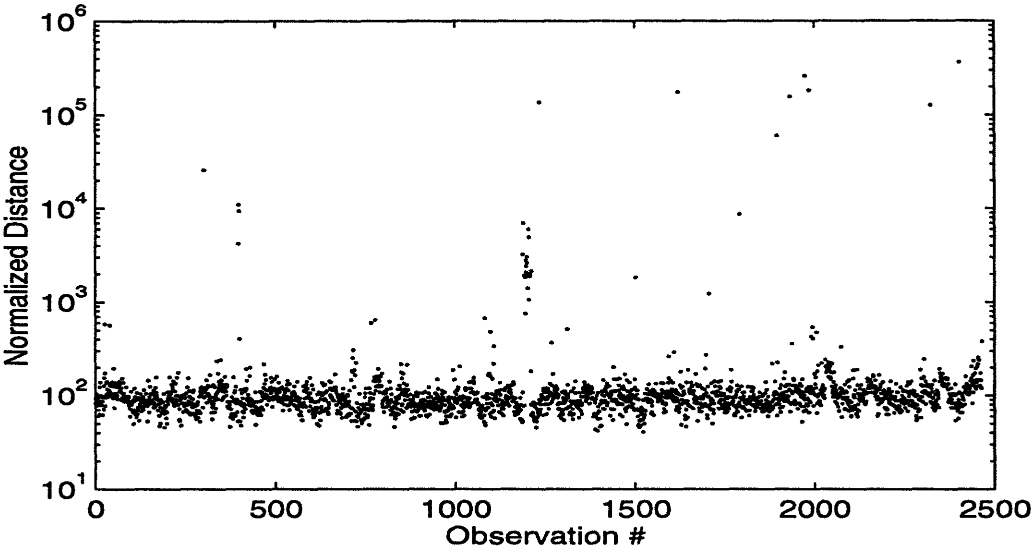

Figure 6-1: Normalized distance from the mean for the input variables.

The robust normalized distance from the mean d_i^{*2} for the input variables of each observation is shown in Figure 6-2. shows that some of the outliers were much more extreme than Figure 6-1 seems to suggest. The reason for this is that the gross outliers greatly inflated the estimates of the standard deviation initially used to scale the data.

Another interesting thing to note is that observations 1620, 1933, and 1986 have the same value for d_i^{*2}. Thus, these points may form a cluster. Alternatively, these observations could be completed unrelated since Figure 6-2 only shows distance and not direction.

Figure 6-3 shows the values of d_i^2 for the 2 output measurements of each observation. This plot also shows several gross outliers, and once again observations 1620, 1933, and

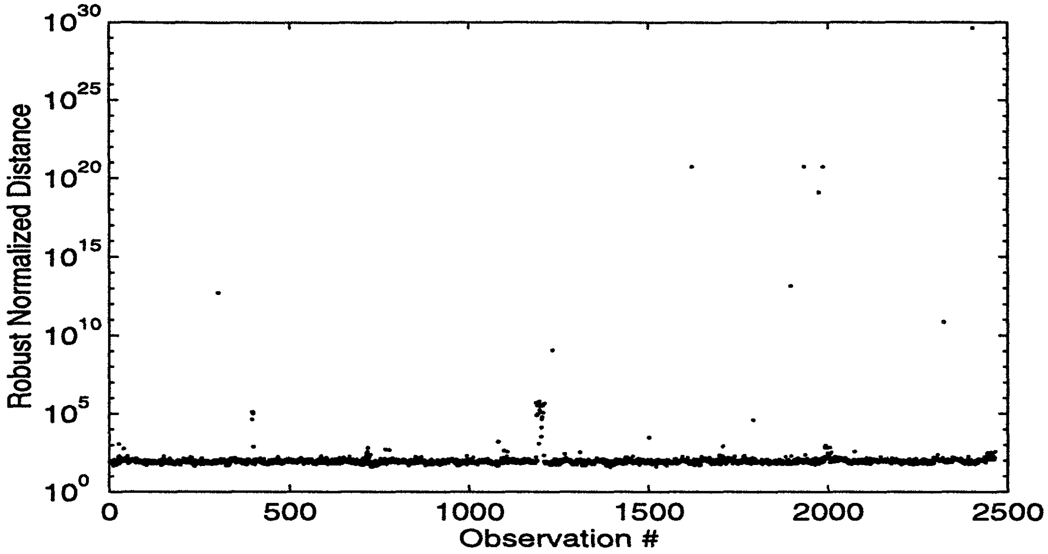

Figure 6-2: Robust normalized distance from the mean for the input variables.

1986 have the same values.

6.3.2 Outlier Characterization

Figure 6-4 shows the measurements greater than 2 robust standard deviations from the mean for each of the observations with a multivariate normalized distance, $d_i^{*2} \geq 300$. Each outlier or outlying batch is characterized by the variables on which it "sticks out".

One interesting thing to note in Figure 6-4 is that most of the gross outliers are extreme on many input variables while the outlying batch is extreme on relatively few input variables. Also, observations 1620, 1933, and 1986 were extreme on many of the same measurement variables.

Figure 6-5 shows the input principal component values greater than 2 standard deviations from the mean for the observations with $d_i^{*2} \geq 300$. Figure 6-6 shows a similar plot based on the robust principal components.

Comparing Figure 6-5 and Figure 6-4 reveals that the outlying batch is extreme on many more principal components than measurement variables while the observation 2403 is extreme on fewer principal components than measurement variables. Furthermore, some observations had extreme principal component values while having no extreme input measurements.

A comparison of Figure 6-5 and Figure 6-6 shows that the gross outliers stick out on

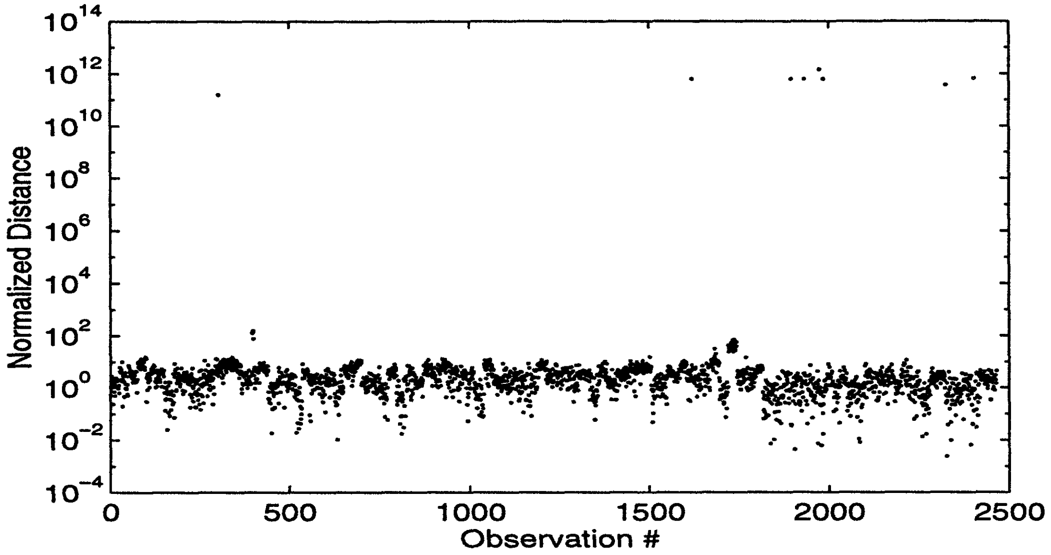

Figure 6-3: Normalized distance from the mean for the output variables.

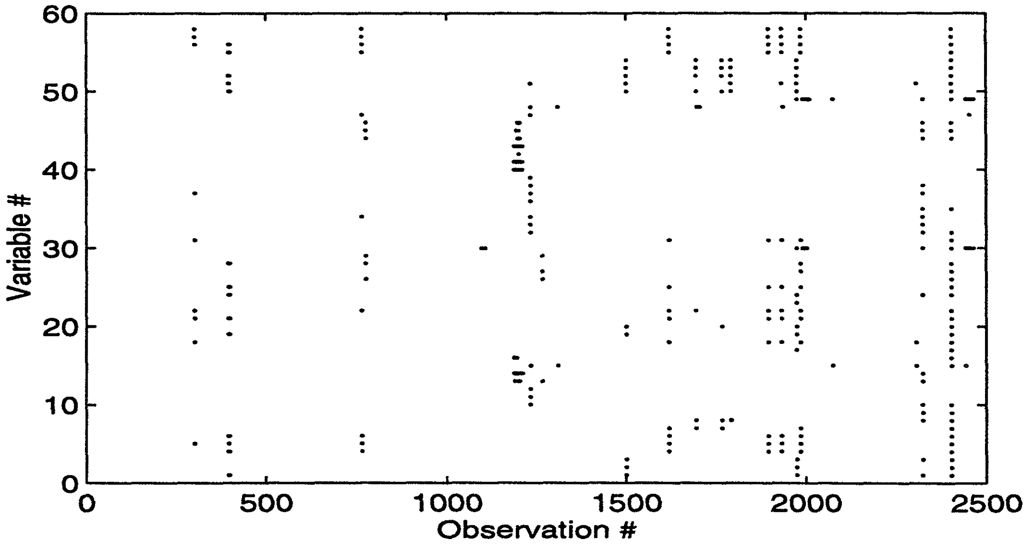

Figure 6-4: Input measurements greater than two robust standard deviations from the mean for observations with ${d_i^*}^2 \geq 300$.

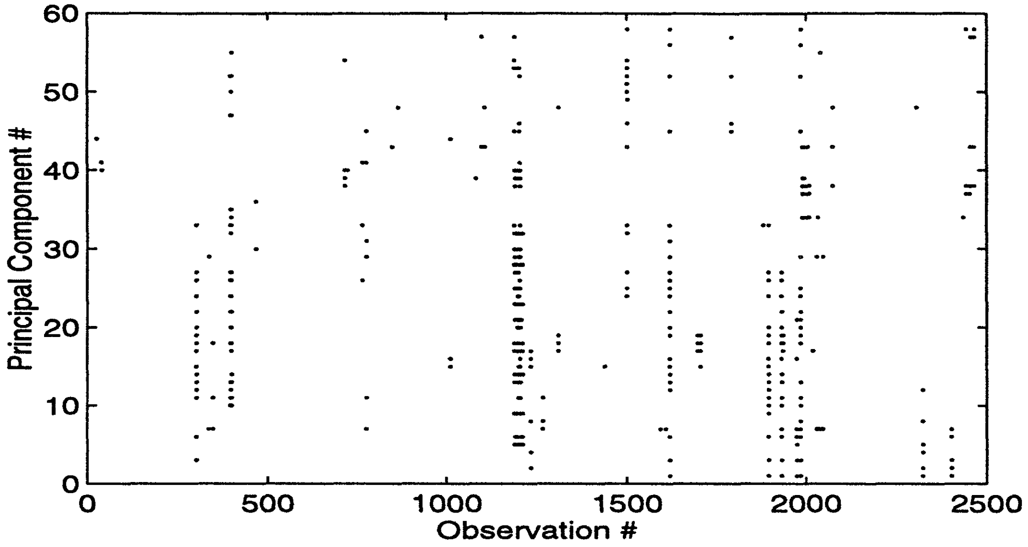

Figure 6-5: Input principal component values greater than two standard deviations from the mean for observations with $d_i^{*2} \geq 300$.

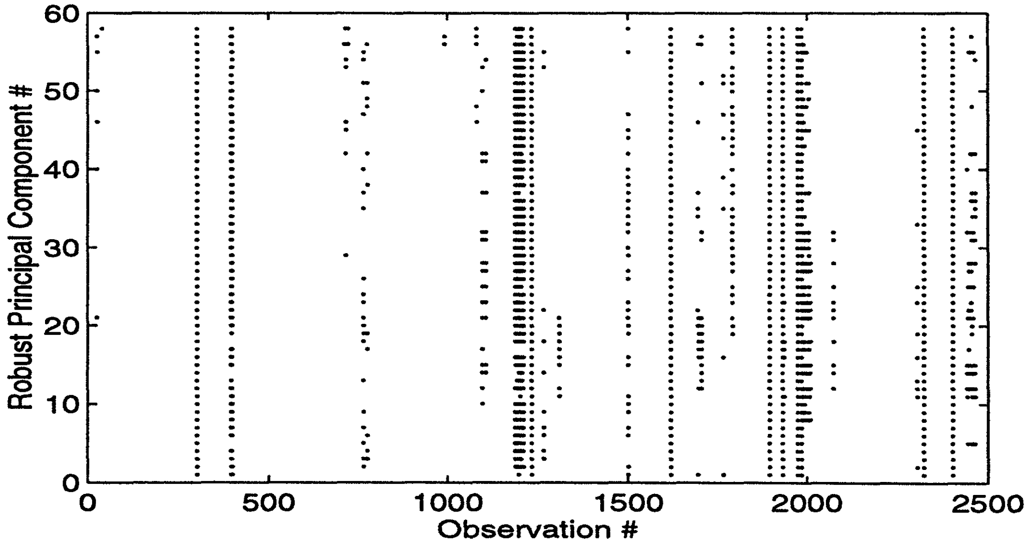

Figure 6-6: Input robust principal component values greater than two robust standard deviations from the mean for observations with $d_i^{*2} \geq 300$.

many more of the robust principal components than regular principal components. The reason for this is that the regular principal components take into account the variation due to the gross outliers.

Since there are only 2 output variables, the 2-dimensional scatter plot shown in Figure 6-7 can represent the data exactly. The main population is located at (0,0), and the gross outliers seem be located only at certain values for each variable. Observations 1620, 1933, and 1986 are clustered near 0 for output variable 1 and near 22 for output variable 2.

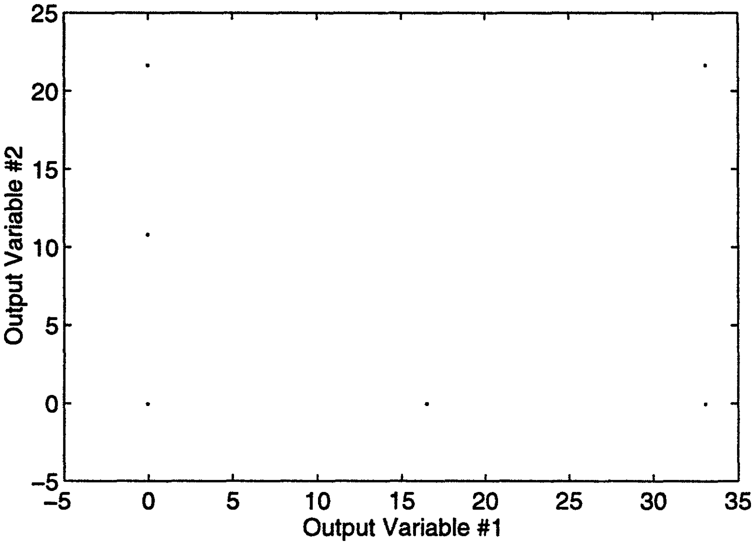

Figure 6-7: Scatter plot of the two output variables.

Figure 6-8 gives a closer view of the main population. This plot shows two groups of outliers plus a positive correlation between the two variables. The group of outliers at the lower left of the figure turned out to be gross outliers in the input measurements, while the group of outliers to the upper right of the main cloud of points were from a single batch. It turned out that the outlying batch in the output was not the same as the outlying batch in the input.

6.3.3 Interpretation of Outliers

The interpretation of the outliers combines engineering knowledge of the process with the information found during outlier characterization. Several hypotheses which may need to be investigated include the following:

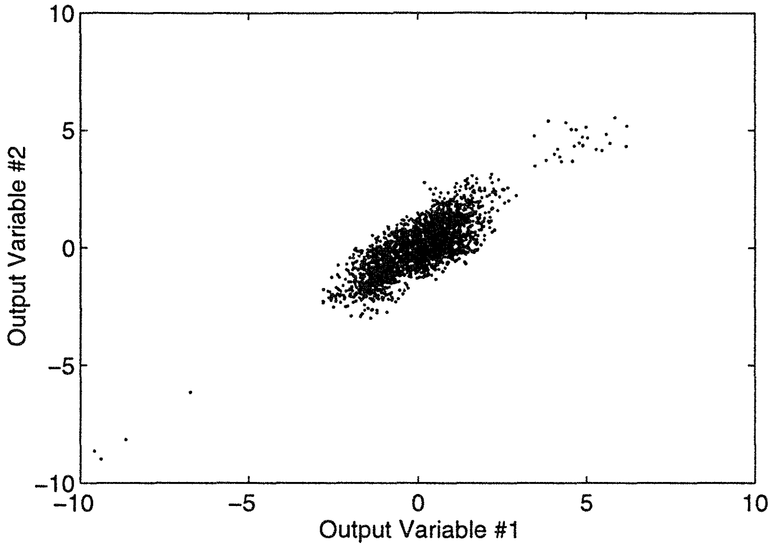

Figure 6-8: The main population in the scatter plot of the two output variables.

1. Observations 1620, 1933, and 1986 have the same cause since they had extreme measurements on the same input variables and output variables.

2. The batch with extreme input measurements did not have extreme output measurements. Thus, the input variables where this batch had extreme measurements are unrelated to the output variables.

3. The gross outliers which had extreme values for principal components but not for measurement variables represent random "noise".

6.3.4 Treatment of Outliers

The treatment of the outliers depends on the subsequent data analysis. Some possibilities include the following:

1. Input-to-output modelling—gross outliers and outlying boats may be removed or modelled separately.

2. Time-series analysis—gross outliers may be replaced with estimates.

3. Inherent variation modelling—gross outliers and outlying boats should be removed.

6.4 Analysis of Other Features

6.4.1 Feature Detection

Figure 6-9 shows the first and second robust principal components for the input data. The main feature of the data is distinct clustering of points.

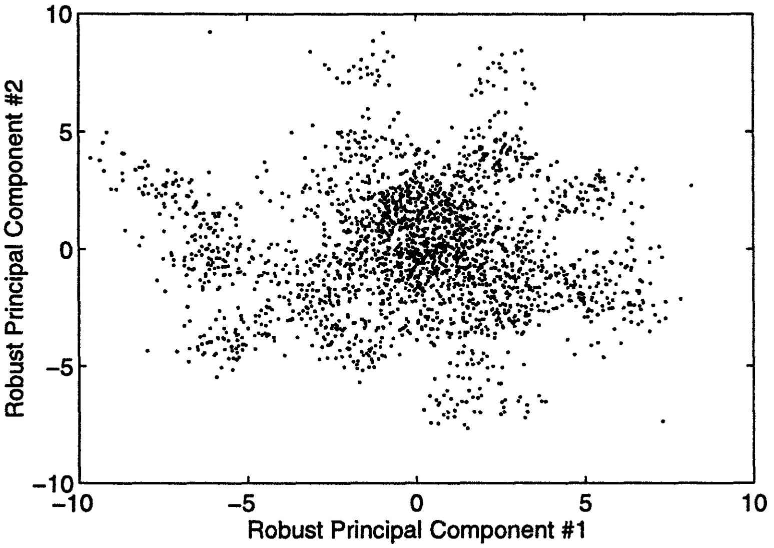

Figure 6-9: First two robust principal components of the input data. (Gross outliers are off the graph.)

In contrast to the scatter plot of the first two robust principal components is the scatter plot of the first two standard principal components shown in Figure 6-10. The clusters are not clearly distinguishable in this representation of the data.

6.4.2 Feature Characterization and Interpretation

Figure 6-11 shows the first robust principal component values for each observation. The clustering in Figure 6-9 clearly reflects variation between batches. This variation seems to have a regular pattern in which 3 or 4 batches drift before being reset. The variation between batches was evident in almost all of the input and output measurement variables.

6.4.3 Treatment of Features

Different data analyses will require different treatments of the variation between clusters. For input-output modelling, the clusters should be left intact. On the other hand, inherent

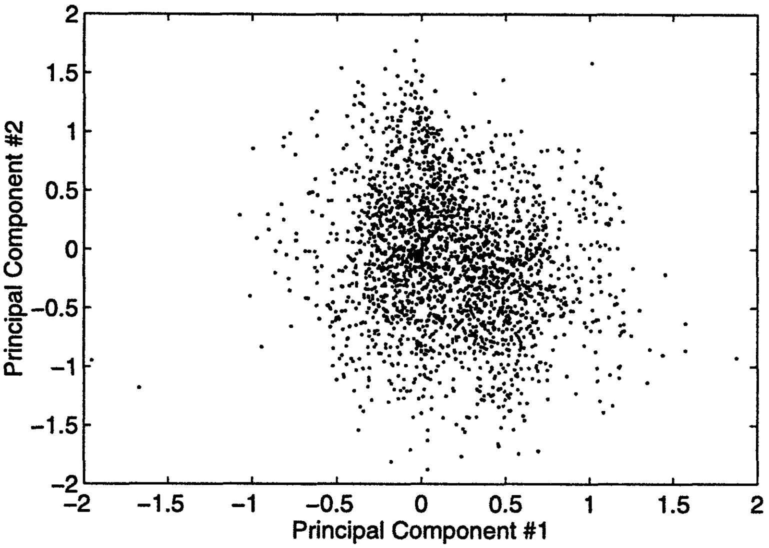

Figure 6-10: First two standard principal components of the input data. (Gross outliers are off the graph.)

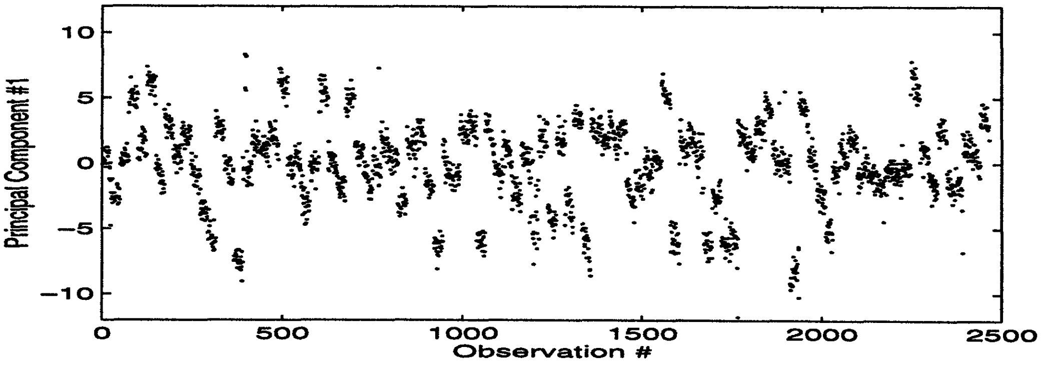

Figure 6-11: First robust principal components of the input data. (Gross outliers are off the graph.)

variation modelling would be more focused on the variation within boats than the variation between boats. To highlight the variation within boats, the mean can be estimated for each boat individually and subtracted out.

6.5 Discussion

The exploratory analysis of the batch process data showed that most of the variation in the data was due to gross outliers and variation between the batches. Based on this information, possible objectives for further analysis include the following:

1. Determine the causes of the gross outliers.

2. Explain the main sources of variation between the batches.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

7.1.1 Outliers

Comparisons of visual displays of standard and robust normalized distance from the mean for each observation showed substantial differences between the statistics based on the standard estimate of the mean, standard deviation, and correlation matrix and statistics based on robust estimates. These differences were caused by gross outliers in the data. Thus, the robust normalized distance from the mean is better suited to initial outlier analysis than the standard normalized distance from the mean.

Three different types of observations were effectively identified as outliers using the robust normalized distance from the mean. The first type had extreme measurements on several variables. The second type of outlier was extreme on only one or two variables. Finally, the last type of outlier had no extreme measurements but had extreme values for the principal components with the lowest variances. This kind of outlier was often due to differences between two highly correlated variables.

7.1.2 Missing Data

Filling out the data set with maximum likelihood estimates based on the recorded values proved effective for detecting the main features in the data. Maximum likelihood methods are the treatment of choice for missing data whenever the amount of missing data is not excessive. Removing complete variables or observations from the data should be considered only if they have an excessive amount of missing data.

The missing data definitely complicated the data analysis. Several of the causes for missing data encountered during the research included:

1. Product removed from line at earlier processing step.

2. Information about product written over on the disk.

3. Measurement equipment did not operate correctly.

While little can be done to correct the first cause, eliminating the last 2 causes for missing data will improve the quality of the data analysis.

7.1.3 Other Features

Both the standard principal components and principal components based on robust estimates of the mean and correlation matrix proved useful for the initial analysis of manufacturing data. The standard principal components showed the dominant variation in the data set—gross outliers in the batch process data, clustering in the assembly process data, and large excursions in the web process data. On the other hand, the robust principal components showed variation in the main population which was obscured by the outliers in the data—clustering in the batch process and drifting in the web process. The standard principal components and the robust principal components were nearly identical for the assembly data because the clustering dominated the main population variation as well as the variation of the main population plus outliers.

7.2 Future Work

One primary area for future work is to develop tools for taking into consideration the time-series nature of manufacturing data when dealing with outliers and missing data. For missing data this means basing estimates of missing values on previous and future observations as well as the current observation. For outliers, this means developing methods for characterizing drifts and periodicities in the presence of outliers and vice versa.

Bibliography

[1] Barnett, V., and Lewis, T. *Outliers in Statistical Data*, Third Edition, John Wiley, Chichester, England, 1994.

[2] Krzanowski, W.J., and Marriott, F.H.C. *Multivariate Analysis*, Part I, Edward Arnold, London, 1994.

[3] Leon-Garcia, A. *Probability and Random Processes for Electrical Engineering*, Addison-Wesley, Reading, Massachusetts, 1994.

[4] Little, Roderick J.A. Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values *Applied Statistics*, **37**, No. 1, pp. 23-38, 1988.

[5] Little, R.J.A., and Rubin, D.B. *Statistical Analysis With Missing Data*, John Wiley and Sons, New York, 1987.

[6] MacGregor, J.F., and Kourti, T. Statistical Process Control of Multivariate Processes. *Control Eng. Practice.* **3**, No. 3, pp. 403-414, 1995.