Abstract

LADA, EMILY KATE. A Wavelet-Based Procedure for Process Fault Detection.
(Under the direction of Dr. J. C. Lu.)

The objective of this research is to develop a procedure for detecting faults in a particular process by analyzing data generated by the process on a compressed wavelet scale. In order to compress the data, three different methods are compared for reducing the number of wavelet coefficients needed to obtain an accurate representation of the data. One method is ultimately chosen to compress all data used in this study. This method, which involves minimizing the relative reconstruction error, effectively balances model parsimony against reconstruction error. The compressed data from seven in-control runs is used to define a standard, or baseline, by which to compare new data sets. The compressed baseline process is defined by forming the union set of the top, or coarsest, coefficient positions selected by the relative reconstruction error method for each of the seven runs. In order to determine if a new, compressed data set differs significantly from the baseline, a variant of Hotelling's T^2 -statistic for two-sample problems is formulated. This statistic is

tested by applying it to four induced-fault data sets, as well as to 21 in-control data sets. The results provide substantial evidence of the statistic's effectiveness in detecting process faults from compressed data. It is also shown that traditional bootstrapping methods cannot be implemented in this case.

A WAVELET-BASED PROCEDURE FOR PROCESS FAULT DETECTION

by

EMILY KATE LADA

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

OPERATIONS RESEARCH

Raleigh

2000

APPROVED BY:

Chair of Advisory Committee

Biography

Emily Lada was born December 18, 1974 in Durham, North Carolina. She was raised by her parents Thomas and Kathleen along with her brothers Matthew and Aaron and her sister Julie. She graduated from Athens Drive High School in 1993.

In the fall of 1993 she entered the University of North Carolina at Chapel Hill. While attending UNC, she was inducted into several honor societies including Phi Beta Kappa and Pi Mu Epsilon. She graduated with highest distinction in May 1997, receiving a Bachelor of Arts degree in Mathematics with a minor in Music.

In the fall of 1997 she began graduate study at North Carolina State University. After completion of her Master of Science in Operations Research, she plans to pursue her Ph.D. in Operations Research.

Acknowledgments

I thank Dr. J. C. Lu for introducing me to the fascinating world of wavelets and for encouraging me to continue my studies as a Ph.D. student. I thank Dr. James R. Wilson for the time and effort he has spent helping me with my research. Without his guidance I could not have completed this project. I also thank Dr. Charles E. Smith for serving on my advisory committee. Finally, I thank my family for the love and support they have given me throughout my entire school career.

Contents

1	Introduction	1
1.1	Problem Statement and Research Objectives	2
1.2	Organization of the Thesis	2
2	Literature Review	3
2.1	Introduction to Wavelets	3
2.1.1	What is a Wavelet?	4
2.1.2	Wavelet Basics	6
3	A Wavelet-Based Procedure for Process Fault Detection	10
3.1	Description of the Data	10
3.2	A Procedure for Detecting Process Faults	12
3.2.1	Model Fitting	13
3.2.2	Determination of the Baseline Process	25
3.2.3	Definition of the Performance Statistic	25
3.2.4	Application of the Statistic	36
4	Conclusions and Future Research	39
4.1	Conclusions	39
4.2	Future Research	48
	Appendices	54
	Appendix A. Computation of the Probability the Bootstrap Sample Co- variance Matrix is Nonsingular	54

List of Tables

3.1	Number of Coefficients Selected for Standard Runs	16
3.2	Coefficients selected to be in the Baseline	26
3.3	The 19 elements of $\hat{\hat{\beta}}$	36
3.4	F -Statistic Values for the 4 Fault Runs	37
3.5	F -Statistic Values for the 21 In-control Runs	38
4.1	Results of the Shapiro-Wilk Test for Normality	41

List of Figures

2.1	Sample Wavelet Coefficient Plot	9
3.1	Data for Nominal Runs 1–4	12
3.2	Data for Nominal Runs 5–7	13
3.3	Data for Four Induced-Fault Runs	14
3.4	Method Comparison for Run 5	17
3.5	Method Comparison for Run 1	19
3.6	Method Comparison for Run 2	20
3.7	Method Comparison for Run 3	21
3.8	Method Comparison for Run 4	22
3.9	Method Comparison for Run 6	23
3.10	Method Comparison for Run 7	24
3.11	Empirical Distribution of the Sample Statistic Compared to the F Distribution	34
3.12	Cumulative Distribution Function of the Sample Statistic Compared to the Corresponding F C.d.f.	35
4.1	Normal Probability Plots for the First 4 Coordinates	42
4.2	Normal Probability Plots for Coordinates 5–8	43
4.3	Normal Probability Plots for Coordinates 9–12	44
4.4	Normal Probability Plots for Coordinates 13–16	45
4.5	Normal Probability Plots for Coordinates 17–19	46
4.6	Normal Probability Plots for Coordinates 17–19 with $m = 10$	47

Chapter 1

Introduction

One of the problems faced by many industries today is an influx of data. With the onset of the computer age, it has become relatively easy to collect information, and as a result data is plentiful. In order to make room for all this information, it is becoming increasingly necessary to compress data before storing it. The following question then naturally arises: is it possible to study and draw accurate conclusions about the data while it is in a compressed form? If it is possible, then it would not be necessary to monopolize computer memory with huge data sets while an analysis is conducted. Furthermore, there are many process-control situations that require quick decisions. A huge amount of data can slow down this decision-making process since additional time is required to manipulate and study a large data set. The use of compressed data for analysis would therefore improve the speed and accuracy of many decision-making processes.

1.1 Problem Statement and Research Objectives

In this study, quadrupole mass spectrometry samples of a rapid thermal chemical vapor deposition process are analyzed on a compressed wavelet scale to determine if it is possible to recognize a specific pattern in the data. In order to compress the data, we compare three different methods that reduce the number of wavelet coefficients needed to obtain an accurate representation of the data. One method is ultimately selected to compress the data used in this study. The compressed data are then used to define a standard, or baseline, by which to compare new data sets and thereby determine if there is a problem with the process used to generate the new data. If the pattern of the new data set does not match the pattern of the baseline to some degree, then the process generating the new data set is not in control. In order to determine if a new, compressed data set differs significantly from the baseline, we formulate a variant of Hotelling's T^2 -statistic for two-sample problems. Essentially, the goal is to detect faults in the process by only examining the set of wavelet coefficients associated with each data set.

1.2 Organization of the Thesis

The remainder of the thesis is organized as follows: Chapter 2 contains an introduction to wavelets and their basic properties. In Chapter 3 there is a description of the data used in this study as well as a detailed outline of the procedure developed to detect process faults on a compressed wavelet scale, including the determination of the baseline process, the definition of the performance statistic, and the application of the statistic. Finally, Chapter 4 contains concluding remarks and suggestions for future research.

Chapter 2

Literature Review

2.1 Introduction to Wavelets

Since the development of wavelet analysis in the late 1980s, there has been tremendous interest in wavelets and their possible applications. This interest is not only expressed in the mathematical world, but in numerous other fields as well, including statistics, physics, computer science, and econometrics. Wavelet analysis can be applied to a number of diverse disciplines, and this is one of the reasons for its growing popularity.

One area in which wavelets have become an especially important tool is data compression. With the onset of the computer age, the amount of data requiring storage is increasing rapidly. Wavelets are being used to transform various types of data sets into a compressed form. The transformation must be chosen, however, so as to minimize the loss when the data is later reconstructed. One organization that is currently utilizing wavelets for data compression is the Federal Bureau of Investigation. The FBI has approximately thirty million sets of fingerprints that

need to be stored, each requiring about 0.6 megabytes of storage. By means of a wavelet representation, only thirty kilobytes of storage space are required for an adequate representation of a fingerprint. This yields an impressive compression ratio of 20 to 1.

Another field where wavelets are being used extensively is signal processing. In signal processing, the goal is to transmit the encoded signal so that it can be reconstructed with only minimal loss. In addition, signal processing involves the removal of random noise that is contained in most signals. Wavelets have been used successfully to “smooth” the noisy data points and obtain an estimate of the underlying function.

The specific applications of wavelets cited above are only a few of the abundant ways wavelets are being used today. Even though it is still a relatively new field, wavelet analysis is developing quickly. As the amount of research increases, it is certain that the number of areas making use of wavelets will also grow.

2.1.1 What is a Wavelet?

In simple terms, a wavelet is a building block function, similar to the sine and cosine functions. In wavelet analysis, linear combinations of dilated and translated wavelet functions are used to estimate other functions, thus making it similar to the Fourier decomposition of functions. Fourier’s result states that any continuous function $f(x)$ on $[0, 2\pi]$ can be expressed as an infinite sum of dilated sine and cosine functions:

$$a_0 + \sum_{k=1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)); \quad (2.1)$$

and this sum converges to $f(x)$ for all $x \in [0, 2\pi]$ (Apostol 1974, Theorem 11.6).

The coefficients a_0, a_k , and b_k are calculated by:

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x) dx$$

$$a_k = 1/\pi \int_0^{2\pi} f(x) \cos(kx) dx$$

$$b_k = 1/\pi \int_0^{2\pi} f(x) \sin(kx) dx$$

for $k = 1, 2, \dots$. Fourier analysis is often used in signal processing. First, a function in the time domain is transformed into a function in the frequency domain using the above formulas. The signal can then be analyzed for its frequency content.

Despite the popularity of Fourier analysis, the use of wavelets in applications where the Fourier transform would normally be applied, such as signal processing, is increasing. There are several advantages to using wavelet analysis over Fourier analysis. First, the Fourier transform has a single set of basis functions, the sine and cosine functions. Wavelet analysis, on the other hand, has access to an infinite set of basis functions (Graps 1995). Therefore, it is often possible with wavelet analysis to obtain information about a function that may not be possible to obtain using Fourier analysis, such as a particularly sharp spike or jump.

A second difference between Fourier methods and wavelet analysis is that Fourier basis functions are localized in frequency, but not time. A small change in the frequency of the sine and cosine functions will result in changes everywhere in the time domain. Wavelet functions, however, damp down to zero, making them local in both frequency and time. This property makes wavelet functions excellent representors of functions, like signals, that may be changing over time and

that may have discontinuities, sharp spikes, and other nonsmooth features. Furthermore, even if it is possible to represent a particular function with nonsmooth features using the Fourier transform, it will take considerably more sine and cosine functions to approximate the function than wavelet functions. As it will be shown later, this plays a role in the use of wavelets for data compression.

2.1.2 Wavelet Basics

One of the most important types of wavelets is the orthogonal wavelet, of which there are many kinds. The first known orthogonal wavelet is the Haar wavelet, discovered in 1910. Other examples of orthogonal wavelets include daubelets, symmlets, and coiflets. Wavelets are grouped according to families. Mother wavelets, denoted by $\psi(t)$, integrate to zero and father wavelets, $\phi(t)$, integrate to one. Dilations and translations of the mother function and father function define an orthogonal basis (Graps 1995), as shown in the following formulas:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k), \quad (2.2)$$

$$\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k). \quad (2.3)$$

The scale index j indicates the wavelet's width. The scale factor 2^j is also called the dilation factor. As j gets larger, the functions $\psi_{j,k}(t)$ and $\phi_{j,k}(t)$ get smaller in magnitude and more spread out. The location index k gives the position of the wavelet. The translation parameter $2^j k$ is linked to the scale parameter 2^j in the sense that as the function gets wider (i.e. as j increases), the translation steps also

get correspondingly larger (Bruce and Gao 1996).

A sum of dilated and translated wavelet functions generated from father and mother wavelets can be used to estimate other functions. For example, the wavelet approximation to a continuous time signal $f(t)$ is given by

$$f(t) \approx \sum_{k=1}^K s_{J,k} \phi_{J,k}(t) + \sum_{j=1}^J \sum_{k=1}^{K_j} d_{j,k} \psi_{j,k}(t), \quad (2.4)$$

where J is the number of multiresolution components and K_j is the number of coefficients in the j th component. In the above representation, $\phi_{J,k}(t)$ and $\psi_{j,k}(t)$ are the scaled and translated wavelet functions. The father wavelets, $\phi_{J,k}(t)$, are good for representing the low frequency parts of the signal, or the underlying smoothness of the data. The mother wavelets, $\psi_{j,k}(t)$, are excellent representors of the high frequency parts of the signal, or the fine scale deviations from the overall smooth behavior of the data. Each wavelet function in the sum is multiplied by a wavelet transform coefficient, $s_{J,k}, d_{J,k}, \dots, d_{1,k}$. The formulas for these wavelet coefficients are given approximately by:

$$s_{J,k} \approx \int \phi_{J,k}(t) f(t) dt, \quad k = 1, \dots, K, \quad (2.5)$$

$$d_{j,k} \approx \int \psi_{j,k}(t) f(t) dt, \quad j = 1, 2, \dots, J. \quad (2.6)$$

The magnitude of a wavelet coefficient gives a measure of the contribution of the corresponding wavelet function to the approximating sum (Bruce and Gao 1996).

Given a signal $f(t)$, the discrete wavelet transform can be used to calculate the wavelet coefficients. In S-PLUS, the following code calculates the discrete wavelet

transform for a signal called run1:

```
run1.dwt ← dwt(run1)
plot(run1.dwt)
```

Figure 2.1 on the following page shows the output from the above code. In the plot, the wavelet coefficients are plotted as vertical lines from the horizontal axis. The fine scale coefficients $d_{1,k}$ are plotted in the second row. The coarse scale coefficients $d_{J,k}$ and $s_{J,k}$ are plotted in the bottom two rows. There are a total of n coefficients, where n is the number of samples from the continuous time signal $f(t)$. As mentioned earlier, 2^j is a measure of the scale, or width, of $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$. At the finest scale $2^1 = 2$, therefore there are $n/2$ coefficients at the $d_{1,k}$ level. Similarly, there are $n/4$ at the next finest scale $d_{2,k}$. Continuing this pattern until the coarsest scale, there are $n/2^J$ coefficients for both $d_{J,k}$ and $s_{J,k}$. In the above, if n is not divisible by 2^j , then there are approximately $n/2^j$ wavelet coefficients at scale 2^j , where the total number of coefficients is still n .

In practical terms, many functions can be estimated by using only a few wavelet coefficients (or, equivalently, by the summation of only a few wavelet functions). One of the important aspects of wavelet analysis is determining the number of coefficients needed to obtain a reasonable estimate of a function. Once this number has been determined, only that certain number of coefficients must be saved, and from those coefficients the function can be re-created. In this way, wavelet analysis can be used for data compression.

Figure 2.1: Sample Wavelet Coefficient Plot

Chapter 3

A Wavelet-Based Procedure for Process Fault Detection

3.1 Description of the Data

The data used in this study are quadrupole mass spectrometry (QMS) samples of a rapid thermal chemical vapor deposition (RTCVD) process. RTCVD is a process developed to obtain accurate wafer temperature measurements and across-wafer temperature uniformity. QMS data was collected from a series of process and diagnostic experiments involving polycrystalline Si deposition from 10% SiH₄/Ar in a single-wafer RTCVD tool ranging in temperature from 625°C to 725°C. The data consists of seven nominal and four induced-fault RTCVD process runs. Figure 3.1 displays the data for the first four nominal runs, Figure 3.2 displays the data for the last three nominal runs, and Figure 3.3 contains graphs of the four induced-fault data sets. The seven nominal runs represent what the data for a standard, in-control run should look like. For a more in-depth description of the process used

to generate the data used in this study, see Rying, Gyurcsik, Lu, Bilbro, Parson, and Sorrell (1997) and Tedder, Rubloff, Cohaghan, and Parsons (1996).

In a process such as this one, it is important to make timely process control decisions. The abundance of data, however, can often hinder the process engineer's ability to make such decisions. The use of compressed data would possibly speed-up the process engineer's ability to detect faults and make accurate decisions to correct the process. From the plots in Figures 3.1 and 3.2, it is clear that a wavelet transformation can be used to compress the data. The sharp spikes and steep drops in the data can be easily captured by the wavelet basis functions. Once the data has been compressed, a procedure that detects faults in the process from the compressed data can then be developed.

Figure 3.1: Data for Nominal Runs 1–4

3.2 A Procedure for Detecting Process Faults

The procedure for detecting process faults from compressed data is broken down into the following steps:

- model fitting;
- determination of the baseline process;
- definition of the performance statistic; and
- application of the performance statistic to identify in-control and out-of-control situations.

Figure 3.2: Data for Nominal Runs 5–7

3.2.1 Model Fitting

As described earlier, the data used in this study consists of seven sets of standard runs and four sets of induced-fault runs. Before any analysis can be done, all seven standard runs must first be compressed using wavelets in order to determine a baseline process to which new data sets can be compared. Each of the seven runs is first fitted using the S-PLUS discrete wavelet transform function (dwt), as

Figure 3.3: Data for 4 Induced-Fault Runs

described in Bruce and Gao (1996). The S-PLUS default wavelet, the symmlet “s8”, is used to fit all data. The “s8” wavelet is an excellent overall choice for many functions since it is orthogonal, smooth, nearly symmetric, and nonzero on a relatively short interval. After the dwt function has been applied to the data, a set of 128 wavelet coefficients is associated with each data set, corresponding to the 128 data points in each run. In order to compress the data, it must now be determined which of the 128 coefficients for each data set are the “most important” and which ones can be set to zero. In other words, those coefficients that help to capture the overall trend of the data are left alone while all other coefficients are set to zero.

One of the challenges of wavelet analysis is determining the best number of coefficients to represent the data accurately. In the model selection process for this study, three different procedures are compared. The first is the built-in S-PLUS function `waveshrink`. The `waveshrink` function is actually intended to be used for noise removal purposes more so than for data compression. It is used in this study for the sake of comparison since it is the simplest way to compress data in S-PLUS. The main premise behind the `waveshrink` function is to shrink the detail coefficients at the smallest scales to zero. The coefficients at the coarsest scales are left untouched.

While the `waveshrink` function does yield an accurate estimate of the function, two other methods for data compression are tested to determine if it is possible to improve the compression. Both methods involve the calculation of the reconstruction error (RE), which is used to evaluate the approximation accuracy in the wavelet reconstruction. The reconstruction error is given by

$$\text{RE} = \left[\sum_{i=1}^n (g_i - \hat{g}_i)^2 \right]^{1/2} / \left(\sum_{i=1}^n g_i^2 \right)^{1/2}, \quad (3.1)$$

where g_i and $\hat{g}_i$ are the data and the wavelet model predictions, respectively, at the i th position, and $n=128$ is the number of data points.

The first process that utilizes the reconstruction error minimizes the approximate minimum description length (AMDL), found in Saito (1994) and Antoniadis, Gijbelo and Gregoire (1997). The approximate minimum description length is given by

$$\text{AMDL}(C) = 1.5 \log_2 n + 0.5n \log_2 \left[\sum_{i=1}^n (g_i - \hat{g}_{i,C})^2 \right], \quad (3.2)$$

where C is the number of largest wavelet coefficients and $\hat{g}_{i,C}$ is the prediction from the wavelet model with C coefficients. In the second process that utilizes the reconstruction error, the number of terms, or coefficients, used in the model is taken into account. The relative reconstruction error (RRE) balances model parsimony against reconstruction error. The relative reconstruction error is defined as

$$\text{RRE}(C) = \left[\sum_{i=1}^n (g_i - \hat{g}_{i,C})^2 \right]^{1/2} / \left(\sum_{i=1}^n g_i^2 \right)^{1/2} + C/n \quad (3.3)$$

(Lu 1999). This second method involves minimizing the sum of two components: the reconstruction error and the ratio of the number of wavelet coefficients to the size of the data set. Although this method of model selection criterion is not well studied yet, it does appear to work well in most data compression cases.

In order to see the comparison between the waveshrink, AMDL, and RRE methods visually, the following table shows the number of coefficients selected using all three methods for the seven standard runs.

Table 3.1: Number of Coefficients Selected for Standard Runs

Run	1	2	3	4	5	6	7
Waveshrink	55	65	50	66	77	59	68
AMDL	86	86	93	86	77	90	84
RRE	11	14	10	11	10	12	10

The following figure shows the RE, RRE, and AMDL values plotted against the number of coefficients for the 128 data points of Run 5 from Figure 3.2. Because of the irregular pattern of the data, the reconstruction using a small number of coefficients (less than 10) is not stable, as exhibited by the first plot of the reconstruction error. The third plot of the relative reconstruction error confirms this.

The minimum value of RRE occurs when the number of coefficients is 10. Note also from the second plot that the AMDL achieves its minimum at 77 coefficients, a considerably higher number than that obtained by minimizing RRE. The fourth plot is a reconstruction of the data using the RRE method. From this plot it is evident that an accurate representation of the data can be obtained from just 10 wavelet coefficients.

Figure 3.4: Method Comparison for Run 5

Similar plots for the other six runs are displayed in Figures 3.5 to 3.10. From these plots, as well as from Table 3.1, it is clear that the RRE method for com-

pression results in a smaller number of selected wavelet coefficients than both the Waveshrink and AMDL methods while still keeping the reconstruction error small. For these reasons, the compression obtained from the RRE method is used for all data sets in this study.

Figure 3.5: Method Comparison for Run 1

Figure 3.6: Method Comparison for Run 2

Figure 3.7: Method Comparison for Run 3

Figure 3.8: Method Comparison for Run 4

Figure 3.9: Method Comparison for Run 6

Figure 3.10: Method Comparison for Run 7

3.2.2 Determination of the Baseline Process

After the seven standard runs were compressed using the RRE method, we determined a compressed baseline process by which to compare new data sets. In order to decide which wavelet coefficient positions to include in the definition of the baseline, we formed the union set of the top, or coarsest, coefficient positions selected by the RRE method for each of the seven runs. The main premise behind forming the union of the coarsest coefficient positions is that the largest coefficients capture the most important features of each set of data. In forming the baseline, we seek to combine the most important features of all seven runs; and therefore we formed the union of all the top coefficient positions. The resulting union set contains nineteen coefficient positions. Table 3.2 lists these nineteen positions, as well as the number of runs for which each position is found in the set of top coefficients, as selected by the RRE method (frequency).

3.2.3 Definition of the Performance Statistic

Now that the coefficient positions for the baseline process have been established, it is necessary to define a performance statistic to determine whether a new run is in control or not. First, however, it is necessary to collect additional data sets that are considered to be in control. Since 19 parameters are being estimated, there must be more than 19 runs in the original set of in-control standard runs. Therefore, 14 additional data sets were obtained. These 14 sets of data were collected in the same manner as the original seven, and therefore should be similar in structure. The new data sets were transformed using a wavelet transformation, and the values of the same 19 coefficient positions for each run were determined. The 1×19

Table 3.2: Coefficients selected to be in the Baseline

Coefficient	Frequency
s5.1	7
s5.2	4
s5.4	1
d5.1	7
d5.2	7
d5.3	7
d4.1	7
d4.3	4
d4.4	7
d4.5	4
d3.6	4
d3.7	4
d3.8	2
d2.13	5
d2.14	3
d1.23	1
d1.26	1
d1.27	2
d1.28	1

dimensional vectors representing the 19 coefficients for each of the standard runs will be denoted by $\{\hat{\beta}_i^\circ, 1 \leq i \leq n^\circ = 21\}$. It is assumed that the $\{\hat{\beta}_i^\circ\}$ are independent and identically distributed random variables with some mean β° and covariance matrix Σ° .

When a new set of data is obtained, it will be transformed using the discrete wavelet transform, in the same manner that the original 21 standard runs were, and the values from the 19 coefficient positions used to define the baseline process will be placed in a vector, $\hat{\beta}$. It is assumed $\hat{\beta}$ is distributed with mean β and covariance matrix Σ° , the same covariance matrix as for the $\{\hat{\beta}_i^\circ\}$. The question then becomes whether the population mean of a new data set is the same as the population mean β° from which were drawn the estimated wavelet coefficient vectors $\{\hat{\beta}_i^\circ\}$ of the 21 in-control runs. The null hypothesis can be written as

$$H_0 : \beta = \beta^\circ. \quad (3.4)$$

Before testing this null hypothesis independent samples of size $m = 5$ are taken *without replacement* from the set

$$\{\hat{\beta}_i^\circ : 1 \leq i \leq n^\circ = 21\}. \quad (3.5)$$

In other words, 5 different sets of 19 coefficient values are selected at a time from the set of 21 1×19 dimensional vectors corresponding to the in-control runs. This process is independently replicated $K = 40$ times. Note that the values of m and K are chosen arbitrarily. There are in general $\binom{n^\circ}{m}$ ways to choose m vectors from a set of n° vectors. So m must be chosen small enough to guarantee that there is a sufficient number of different samples of size m , and K must be less than the

number of unique samples of size m that can be selected from a population of size n° . For $m = 5$, there are $\binom{21}{5} = 20,349$ ways to choose a sample of 5 different vectors from the set of 21. So for $K = 40$, there is a high probability that the 40 samples of size 5 will be unique. The uniqueness of the samples becomes important when calculating the sample covariance matrix.

Let $\hat{\hat{\beta}}_k$ denote the sample mean of all $m = 5$ values in the k th sample from (3.5), $k = 1, 2, \dots, K = 40$. Since each sample of size m is taken without replacement from the original data set (3.5) of estimated wavelet coefficients for n° in-control processes, $\hat{\hat{\beta}}_k$ is the average of m quantities that in general are assumed to be i.i.d. random vectors with mean β° and covariance matrix Σ° ; thus, invoking the multivariate Central Limit Theorem (Anderson 1984), we see that $\hat{\hat{\beta}}_k$ is approximately normal with mean vector β° and covariance matrix

$$\text{Cov}(\hat{\hat{\beta}}_k) = \frac{\text{Cov}(\hat{\beta}_i^\circ)}{m} \left(1 - \frac{m}{n^\circ}\right) = \frac{\Sigma^\circ}{m} \left(1 - \frac{m}{n^\circ}\right), \quad (3.6)$$

where the quantity $(1 - \frac{m}{n^\circ})$ is the finite population correction for sample size m and population size n° (Cochran 1977). Since we are sampling from a finite population of size $n^\circ = 21$, we must multiply the covariance matrix of the mean vector of the sample, $\text{Cov}(\hat{\hat{\beta}}_k)$, by the finite population correction. If the sampling fraction m/n° is small, the population correction is close to unity and therefore the size of the population has little effect on the standard error of the sample mean. Otherwise, if the correction is ignored, the covariance matrix of the sample mean will be overestimated.

We observe that the $\{\hat{\hat{\beta}}_k : 1 \leq k \leq K\}$ are i.i.d. since they are obtained via K independent replications of the basic operation of taking a sample of size m from

(3.5). The following overall sample statistics are calculated:

$$\widehat{\widehat{\beta}} = \frac{1}{K} \sum_{k=1}^K \widehat{\beta}_k \quad (3.7)$$

and

$$\widehat{\widehat{S}} = \frac{1}{K-1} \sum_{k=1}^K (\widehat{\beta}_k - \widehat{\widehat{\beta}})(\widehat{\beta}_k - \widehat{\widehat{\beta}})^T \quad (3.8)$$

where

$$E[\widehat{\widehat{\beta}}] = E[\widehat{\beta}] = \beta^\circ \text{ and } E[\widehat{\widehat{S}}] = \text{Cov}(\widehat{\beta}_k) = \frac{\Sigma^\circ}{m} \left(1 - \frac{m}{n^\circ}\right).$$

So $\widehat{\widehat{S}}$ is an unbiased estimator of $\text{Cov}(\widehat{\beta}_k)$, and $\widehat{\widehat{\beta}}$ is approximately normally distributed with mean β° (the same mean as for the original 21 in-control runs) and covariance matrix

$$\text{Cov}(\widehat{\widehat{\beta}}) = \text{Cov}(\widehat{\beta}_k)/K = \frac{\Sigma^\circ}{mK} \left(1 - \frac{m}{n^\circ}\right).$$

Recall that in the previous section it was decided that 14 additional data sets needed to be used since 19 parameters are being estimated. In order for the matrix $\widehat{\widehat{S}}$ to be nonsingular, a necessary and sufficient condition is that the corresponding random sample $\{\widehat{\beta}_k : k = 1, \dots, K\}$ must contain $p = 19$ linearly independent vectors and must satisfy $K \geq p+1$ (Anderson 1984, Theorem 7.5.2). Starting from an initial data set consisting of 7 in-control runs, we augmented this data set with 14 additional in-control runs to obtain the original sample (3.5); and we specifically designed our scheme for generating the random sample $\{\widehat{\beta}_k : 1 \leq k \leq K\}$ to ensure

it is “practically certain” that $\{\hat{\hat{\beta}}_k : 1 \leq k \leq K\}$ will contain p linearly independent vectors and thus $\hat{\hat{S}}$ will be nonsingular.

To test the null hypothesis in (3.4), we compute the following test statistic

$$D = \sqrt{\frac{mKn^\circ}{mKn^\circ + n^\circ - m}}(\hat{\beta} - \hat{\hat{\beta}}), \quad (3.9)$$

where $\hat{\beta}$ is a 1×19 dimensional vector containing the 19 coefficient values for a new set of data. Note that under H_0 ,

$$E[D] = 0$$

and

$$\begin{aligned} \text{Cov}(D) &= \frac{mKn^\circ}{mKn^\circ + n^\circ - m} \left\{ \text{Cov}(\hat{\beta}) + \text{Cov}(\hat{\hat{\beta}}) \right\} \\ &= \frac{mKn^\circ}{mKn^\circ + n^\circ - m} \left\{ \Sigma^\circ + \frac{\Sigma^\circ}{mK} \left(1 - \frac{m}{n^\circ}\right) \right\} = \Sigma^\circ. \end{aligned}$$

By multiplying the difference $(\hat{\beta} - \hat{\hat{\beta}})$ by $\sqrt{\frac{mKn^\circ}{mKn^\circ + n^\circ - m}}$, we obtain a test statistic D that has the same covariance matrix as $\{\hat{\beta}_i^\circ\}$ and $\hat{\beta}$.

The statistic D can now be used to calculate a two-sample version of Hotelling’s T^2 -statistic. In general, the T^2 -statistic has the form

$$T^2 = YS^{-1}Y^T, \quad (3.10)$$

where Y is distributed according to $N_p(0, \Sigma)$ and the distribution of S is p -dimensional Wishart with n degrees of freedom (that is, nS is distributed independently as $\sum_{\alpha=1}^n Z_\alpha Z_\alpha^T$ with the Z_α independent, each with distribution $N_p(0, \Sigma)$).

Since we know $\hat{\hat{\beta}}$ is approximately normal, the difference $(\hat{\beta} - \hat{\hat{\beta}})$ should also be approximately normal since the distribution of $\hat{\beta}$ is not likely to be extremely different from the distribution of $\hat{\hat{\beta}}$, unless the process used to generate the data is severely out of control (in which case, one would not need to conduct a statistical test to recognize there is something wrong with the process). So, if we take $Y = D$, the following T^2 -statistic can be calculated:

$$T^2 = D[W]^{-1}D^T$$

where $E[W] = \text{Cov}(D)$. Since $E[\hat{\hat{S}}] = \frac{\Sigma^\circ}{m}(1 - m/n^\circ)$, if we take $W = (\frac{m}{1-m/n^\circ})\hat{\hat{S}}$, then $E[W] = \frac{m}{1-m/n^\circ}E[\hat{\hat{S}}] = \Sigma^\circ = \text{Cov}(D)$. Furthermore, since the $\hat{\hat{\beta}}_k$ are independent, identically distributed random variables that are approximately normal and have mean β° and covariance matrix $\frac{\Sigma^\circ}{m}(1 - m/n^\circ)$, we see that $\hat{\hat{S}}$ is Wishart; and now our T^2 -statistic can be written as

$$\begin{aligned} T^2 &= \sqrt{\frac{mKn^\circ}{mKn^\circ + n^\circ - m}}(\hat{\beta} - \hat{\hat{\beta}}) \left[\frac{m}{(1 - m/n^\circ)}\hat{\hat{S}} \right]^{-1} \sqrt{\frac{mKn^\circ}{mKn^\circ + n^\circ - m}}(\hat{\beta} - \hat{\hat{\beta}})^T \\ &= \left(\frac{mKn^\circ}{mKn^\circ + n^\circ - m} \right) \left(\frac{1 - m/n^\circ}{m} \right) (\hat{\beta} - \hat{\hat{\beta}}) \hat{\hat{S}}^{-1} (\hat{\beta} - \hat{\hat{\beta}})^T \\ &= \left(\frac{Kn^\circ(1 - m/n^\circ)}{mKn^\circ + n^\circ - m} \right) (\hat{\beta} - \hat{\hat{\beta}}) \hat{\hat{S}}^{-1} (\hat{\beta} - \hat{\hat{\beta}})^T, \end{aligned} \tag{3.11}$$

where T^2 has the T^2 -distribution with $K - 1$ degrees of freedom. Furthermore, consider a T^2 -distribution with n degrees of freedom. Then we have the following theorem from multivariate statistics (Anderson 1984):

Theorem 1 Let $T^2 = Y^T S^{-1} Y$, where Y is distributed according to $N(0, \Sigma)$ and nS is independently distributed as $\sum_{\alpha=1}^n Z_\alpha Z_\alpha^T$ with the Z_α independent,

each with distribution $N(0, \Sigma)$. Then $(T^2/n) [(n-p+1)/p]$ is distributed as an F -distribution with p and $n - p + 1$ degrees of freedom.

Applying this theorem to our example, $n = K - 1, p = 19$, and

$$\frac{T^2}{K-1} \frac{(K-1) - p + 1}{p} \quad (3.12)$$

has an F -distribution with $p = 19$ and $K - p = 40 - 19 = 21$ degrees of freedom.

Note that in the calculation of the T^2 -statistic in (3.11), it is not assumed that $\hat{\beta}$ is normal. Recall that each $\hat{\beta}$ vector is a sample of wavelet coefficients from a new set of data. It can be difficult to determine the exact distribution of wavelet coefficients, and therefore assuming the $\hat{\beta}$ vectors are samples from a normal distribution would be incorrect. However, providing that randomization is properly employed, certain statistical procedures, such as the above F test, can be used as approximations to exact tests based on specific distributional assumptions (such as normality for $\hat{\beta}$ and $\hat{\hat{\beta}}$) (Box, Hunter, and Hunter 1978). By taking samples of size $m = 5$ without replacement from the set of coefficient vectors $\{\hat{\beta}_i^\circ\}$ and then averaging them to form each $\hat{\hat{\beta}}_k$ vector, $k = 1, \dots, 40$, randomization is introduced. Therefore, the assumption of exact normality of the $\hat{\beta}$ vectors is not required.

In order to verify that the statistic calculated in Equation (3.12) does approximately follow an F -distribution, an empirical distribution was formed in the following way. First $\hat{\hat{\beta}}$ and $\hat{\hat{S}}$ were calculated according to Equations (3.7) and (3.8), respectively, for a sample of $K = 40$ new $\hat{\hat{\beta}}_k$ vectors. Next, an in-control $\hat{\beta}$ was randomly sampled from the original set of 21 observations, specified in Equation (3.5). The corresponding sample statistic value was then computed using slightly adjusted versions of Equations (3.11) and (3.12). Since $\hat{\beta}$ is sampled from

the original 21 runs, we have to take into account the finite population correction, thus making the covariance matrix of $\hat{\beta}$ equal to $\Sigma^\circ(1 - 1/n^\circ)$. This results in the slightly altered test statistic

$$T^2 = \left(\frac{1}{(1 - 1/n^\circ) + \frac{(1-m/n^\circ)}{mK}} \right) \left(\frac{1 - m/n^\circ}{m} \right) (\hat{\beta} - \hat{\hat{\beta}}) \hat{\hat{S}}^{-1} (\hat{\beta} - \hat{\hat{\beta}})^T. \quad (3.13)$$

This process was repeated 1000 times to obtain the empirical distribution of the sample statistic (3.12). In Figure 3.11, a histogram of the 1000 sample statistic values is shown. The actual p.d.f. of the F -distribution with $p = 19$ and $K - p = 21$ degrees of freedom is superimposed on the histogram. In Figure 3.12 the cumulative distribution function (c.d.f.) of the empirical distribution is plotted against the c.d.f. of the F -distribution. From these figures it is clear that the sample statistic values do in fact closely follow an F -distribution with 19 and 21 degrees of freedom—especially in the middle. The fit is also fairly good in the upper tail, which is the part of the distribution used to detect significant departures from an in-control condition. Hence, it is possible to conduct a significance test, as outlined above, without making any assumptions about the underlying distribution of the $\hat{\beta}_i^\circ$ or $\hat{\beta}$.

Figure 3.11: Empirical Distribution of the Sample Statistic Compared to the F Distribution

Figure 3.12: Cumulative Distribution Function of the Sample Statistic (Step Function) Compared to the Corresponding F C.d.f. (Dashed Curve)

3.2.4 Application of the Statistic

The F statistic obtained in Equation (3.12) in the previous section can be used to determine if a new set of data is in control or not. Recall that in Figure 3.3 the plots of four induced-fault data sets are shown. These four runs will be used to test the effectiveness of the statistic derived in the previous section. First, each of these fault runs is transformed using wavelets; and the values for the same 19 coefficient positions used to define the baseline process in Section 3.2.2 are placed in a 1×19 dimensional vector, $\hat{\beta}$. After applying the technique outlined in Section 3.2.3, we obtained the value for the sample statistic $\hat{\hat{\beta}}$, defined in Equation (3.7), and it is shown in Table 3.3.

Table 3.3: The 19 elements of $\hat{\hat{\beta}}$

8.6083
2.9913
0.1398
-0.9096
0.6110
-0.4951
-0.3804
-0.1216
-0.2223
-0.1249
0.0510
-0.0264
-0.0471
-0.0228
0.0094
0.0196
0.0100
-0.0349
0.0043

The sample covariance matrix $\hat{\hat{S}}$, defined in (3.8), was also calculated. Both $\hat{\hat{\beta}}$ and $\hat{\hat{S}}$ are then used to calculate for each of the four fault datasets the T^2 -statistic defined in (3.11). This T^2 value is used to determine an F -statistic with $p = 19$ and $K - p = 21$ degrees of freedom, as outlined in equation (3.12). Table 3.4 lists the value of the F -statistic for each of the four fault runs, as well as the P -value when the statistic value is compared to an F -distribution with 19 and 21 degrees of freedom. Note that the P -values for all four runs are either 0 or extremely close to 0. This means that the probability of seeing an F -value greater than or equal to any of the four values in Table 3.4 is 0 when the null hypothesis $H_0 : \beta = \beta^\circ$ is true. In other words, under the null hypothesis the four fault runs are considered to be rare events. Therefore, these four runs are classified as out-of-control, and steps would need to be taken to correct the process generating the data.

Table 3.4: F -Statistic Values for the 4 Fault Runs

Run	F -Value	P -value
Fault 1	8.476612	4.7294×10^{-6}
Fault 2	94.71474	3.3307×10^{-16}
Fault 3	5192.069	0
Fault 4	205.1892	0

To further test the statistic derived in Section 3.2.3, the F -values and P -values for the 21 original, in-control runs are calculated. The results are in Table 3.5. If in fact the sample statistic values do follow an F -distribution with 19 and 21 degrees of freedom, then the P -values for the 21 in-control runs should be uniformly distributed on (0,1). From Table 3.5, it appears that the P -values for the 21 runs are uniformly distributed between 0 and 1. To verify this statistically, a Kolmogorov-Smirnov test was performed. The Kolmogorov-Smirnov statistic, de-

Table 3.5: F -Statistic Values for the 21 In-control Runs

Run	F-Value	P -value
Run 1	1.184815	0.351288
Run 2	0.524459	0.918744
Run 3	0.604390	0.862763
Run 4	1.152553	0.374146
Run 5	0.821726	0.664396
Run 6	1.660176	0.130432
Run 7	0.810794	0.675034
Run 8	2.578866	0.018960
Run 9	3.515237	0.003231
Run 10	0.630564	0.841669
Run 11	0.947632	0.544215
Run 12	0.712667	0.769062
Run 13	0.642125	0.831989
Run 14	0.903888	0.585163
Run 15	1.065073	0.441716
Run 16	0.689684	0.790193
Run 17	0.724572	0.757930
Run 18	0.677874	0.800840
Run 19	0.917106	0.572667
Run 20	0.831845	0.654552
Run 21	0.549827	0.902493

fined in this case to be the greatest absolute vertical distance between the empirical distribution function of the P -values and the Uniform(0,1) distribution, is 0.2585. The P -value for the K-S statistic is 0.1002, which provides further evidence that the P -values for the 21 in-control runs are randomly distributed between 0 and 1.

Chapter 4

Conclusions and Future Research

4.1 Conclusions

From the results obtained in the previous chapter, it is evident that the proposed method for detecting process faults from data that is compressed using wavelets is successful. However, the validity of the procedure depends on the $\{\hat{\hat{\beta}}_k\}$ being normally distributed. Recall that the $\{\hat{\beta}_i^\circ\}$ are vectors containing 19 coefficients from the wavelet transformation of the i th data set, $i = 1, \dots, 21$. As mentioned in the previous chapter, it can be difficult to determine the distribution of the coefficients from a wavelet transformation, and hence, the exact distribution of the $\{\hat{\beta}_i^\circ\}$ is not known. It is assumed the $\{\hat{\beta}_i^\circ\}$ are samples from some distribution with mean β° and covariance matrix Σ° . In the procedure described in Section 3.2.3, the Central Limit Theorem is used to conclude that $\hat{\hat{\beta}}_k$, the sample mean of $m = 5$ values in the k th sample (taken without replacement) from the set

$$\{\hat{\beta}_i^\circ : 1 \leq i \leq n^\circ = 21\}, \quad (4.1)$$

is a sample from a normal distribution with mean β° and covariance matrix $\frac{\Sigma^\circ}{m}(1 - m/n^\circ)$. Clearly, if the distribution of the $\{\hat{\beta}_i^\circ\}$ is known to be normal, then the Central Limit Theorem would not be required to conclude the $\{\hat{\hat{\beta}}_k\}$ are normal. However, since the distribution of the $\{\hat{\beta}_i^\circ\}$ is not known, the Central Limit Theorem is invoked to conclude that when batches of size $m = 5$ are taken and averaged, the result, $\hat{\hat{\beta}}_k$, is approximately normal.

In order to verify that the $\{\hat{\hat{\beta}}_k\}$ calculated in this study are in fact approximately normal, the Shapiro-Wilk test statistic (Royston 1982) was calculated for each of the 19 coordinates in the $K = 40$ vectors $\{\hat{\hat{\beta}}_k\}$. The results are shown in Table 4.1. Coefficient positions 4, 16, 17, and 18 have very small P -values, suggesting a departure from normality. However, the majority of the marginal distributions of the 19 coefficient positions do appear to be approximately normal, as evidenced by their high P -values. This provides strong evidence that the joint distribution of the 19 coordinates is normal.

To further test the normality of the $\{\hat{\hat{\beta}}_k\}$ vectors, one normal probability plot for each of the 19 coordinates in the $K = 40$ vectors $\{\hat{\hat{\beta}}_k\}$ is graphed. If the plots appear to be linear, then the marginal distributions of the 19 individual coordinates are approximately normal; and hence there is strong evidence that the joint distribution of the 19 coordinates is normal (i.e. the $\hat{\hat{\beta}}_k$ are normal). The following five figures contain the normal probability plots of the 19 coordinate positions across the sampled $\hat{\hat{\beta}}_k$ vectors. The normal probability plots agree with the results of the Shapiro-Wilk Test. Most of these plots are close to being linear. A few of them, such as the plots for the seventeenth, eighteenth, and nineteenth coordinates, exhibit more of a nonlinear pattern. The linearity of these plots would improve if a larger sample size m were taken. The plots for these same three

Table 4.1: Results of the Shapiro-Wilk Test for Normality

Coefficient	Test Statistic	P -value
1	0.964585	0.328223
2	0.958640	0.212767
3	0.963281	0.299376
4	0.757117	4.0733×10^{-8}
5	0.989243	0.978931
6	0.985483	0.925783
7	0.981626	0.833249
8	0.964693	0.330712
9	0.981467	0.828751
10	0.963822	0.311104
11	0.971289	0.507970
12	0.972663	0.550445
13	0.977929	0.720630
14	0.960465	0.243920
15	0.954961	0.160303
16	0.801332	8.31863×10^{-7}
17	0.805744	1.13739×10^{-6}
18	0.913233	4.93825×10^{-3}
19	0.932421	2.50503×10^{-2}

coordinates for a sample size of $m = 10$ are shown in Figure 4.6. The linearity of the plots for these three coordinates is vastly improved by slightly increasing the sample size m . It is clear, however, that even with the small sample size of five, most of the marginal distributions are approximately normal, and therefore the $\{\hat{\beta}_k\}$ are also approximately normal.

While the test described above is often a sufficient test for multivariate normality, it is not always the case that normality of the marginal distributions implies that the joint distribution is multivariate normal. Take for example the following case from Stewart and Ord (1997). Suppose x and y have a standardized binormal distribution. Whenever the signs of x and y differ, the sign of x is changed to

Figure 4.1: Normal Probability Plots for the First 4 Coordinates

Figure 4.2: Normal Probability Plots for Coordinates 5–8

Figure 4.3: Normal Probability Plots for Coordinates 9–2

Figure 4.4: Normal Probability Plots for Coordinates 13–16

Figure 4.5: Normal Probability Plots for Coordinates 17–19

Figure 4.6: Normal Probability Plots for Coordinates 17–19, sample size = 10

give a new variate $x' = x \operatorname{sgn}(xy)$. Notice that x' and y are marginally normally distributed, but that their joint distribution is not binormal since $\operatorname{sgn}(x') = \operatorname{sgn}(y)$. There are more rigorous and in-depth tests for multivariate normality that could be conducted to further validate the normality assumption of the $\hat{\hat{\beta}}_k$ vectors. One such test that involves calculating a multivariate Shapiro-Wilk statistic is outlined in Tew and Wilson (1988).

4.2 Future Research

It can be argued that, despite the excellent results obtained in Chapter 3, the suggested procedure to detect faults in compressed data is not accurate because of the normality assumption placed on the $\hat{\hat{\beta}}_k$ and $(\hat{\beta} - \hat{\hat{\beta}})$ vectors. Perhaps a procedure that does not require any distribution assumptions, such as bootstrapping, should be applied. The following steps outline a possible bootstrapping procedure to detect faults from compressed data:

STEP 1: Mark the original sample of 21 vectors, each containing 19 elements, as

$$\beta_1, \beta_2, \dots, \beta_{21}. \quad (4.2)$$

Generate a uniformly distributed random number R from $[0,1]$. Let $k = 20R + 1$. Pick up the k th sample β_k from our original 21 samples as the new bootstrap sample β_k^* . Repeating this 21 times, we obtain the bootstrap sample

$$\{\beta_1^*, \beta_2^*, \dots, \beta_{21}^*\}. \quad (4.3)$$

STEP 2: Estimate the mean β and covariance matrix Σ of the underlying distribution that the set $\beta_1, \beta_2, \dots, \beta_{21}$ is a sample from by calculating the current bootstrap sample mean $\bar{\beta}^*$ and covariance matrix $\hat{\Sigma}^*$.

STEP 3: Randomly choose a β_k from the original data set in (4.2) as a new vector β and calculate the following statistic:

$$T^* = (\beta - \bar{\beta}^*)^T (\hat{\Sigma}^*)^{-1} (\beta - \bar{\beta}^*). \quad (4.4)$$

STEP 4: Repeating Steps 1 and 3 $N \geq 1000$ times, N bootstrap estimates $T_1^*, T_2^*, \dots, T_N^*$ are obtained.

STEP 5: Sorting $T_1^*, T_2^*, \dots, T_N^*$ from smallest to largest, the bootstrap distribution of T^* is obtained.

STEP 6: Let T_L^* and T_U^* be the $\alpha/2$ and $1 - \alpha/2$ percentiles, respectively, and $\hat{\bar{\beta}}$ be the mean value of the bootstrap distribution of T^* .

STEP 7: Let $\bar{\beta} = \frac{1}{21} \sum_{i=1}^{21} \beta_i$ be the sample mean and $\hat{\Sigma} = \frac{1}{21} \sum_{i=1}^{21} (\beta_i - \bar{\beta})(\beta_i - \bar{\beta})$ be the original sample covariance matrix.

STEP 8: Decision Rule: when a new β results in a statistic $T = (\beta - \bar{\beta})^T \hat{\Sigma}^{-1} (\beta - \bar{\beta})$ that is outside of the interval (T_L^*, T_U^*) , then that new β is out of control.

While the above bootstrapping procedure does solve the problem of having to assume the $\hat{\beta}_k$ from the procedure outlined in Chapter 3 are normal, there is one major problem with the procedure. The probability that the covariance matrix calculated in Step 2 will be nonsingular is very low. The sample covariance matrix $\hat{\Sigma}^*$ of the bootstrap sample specified in Equation (4.3) will be singular if there

are not at least $p=19$ linearly independent vectors in the bootstrap sample. Furthermore, assuming there are 21 different vectors in the original sample, specified in Equation (4.2), and the bootstrap sample is taken with replacement from the original sample, the probability that the bootstrap sample will contain at least 19 different vectors (that is, $\hat{\Sigma}^*$ is invertible) is about 3.526×10^{-4} (see Appendix A). Since this probability is definitely less than 1, there will be problems executing the bootstrap procedure since it is highly unlikely that the sample covariance matrix calculated in Step 2 will be invertible.

This is an example of a bootstrap scheme that cannot be implemented. A p -dimensional vector is being estimated ($p = 19$ in this case). The original sample size, n , must be substantially larger than p to ensure that there will be at least p different vectors in each bootstrap sample taken with replacement from the original n vectors. As the sample size becomes much larger than p , the probability that the sample covariance matrix is invertible approaches 1. So one solution to the bootstrap method problem is to obtain more data sets. In this study, only 21 original data sets were used. In order to implement the above bootstrapping method, much more than 21 data sets will have to be used. However, even if the sample size is considerably larger than p , the probability that the covariance matrix is nonsingular will not be exactly 1, so it cannot be guaranteed that this bootstrapping procedure will work every time.

Since the bootstrapping procedure does not impose any specifications on underlying distributions, it would be worthwhile to research a solution to the singularity problem that arises in Step 2 so that the method could be implemented successfully. Furthermore, it would be interesting to compare the results of the bootstrap procedure to those obtained in Chapter 3 and see the differences, if any, the im-

position of normality on the $\widehat{\widehat{\beta}}_k$ and $(\hat{\beta} - \widehat{\widehat{\beta}})$ vectors makes.

Bibliography

- Anderson, T.W. 1984. *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.
- Antoniadis, A., I. Gijbels, and G. Gregoire. 1997. Model selection using wavelet decomposition and applications. *Biometrika* 84(4):751-763.
- Apostol, T.M. 1974. *Mathematical analysis*. 2d ed. Reading, Massachusetts: Addison-Wesley.
- Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley.
- Bruce, Andrew, and Hong-Ye Gao. 1996. *Applied wavelet analysis with S-PLUS*. New York: Springer.
- Cochran, William G. 1977. *Sampling Techniques*. New York: John Wiley & Sons.
- Das Gupta, Somesh. 1971. Nonsingularity of the sample covariance matrix. *Sankhya A* 33:475-478.
- Graps, Amara. 1995. An introduction to wavelets. *IEEE Computational Science and Engineering* 2(2).
- Lu, J. C. 1999. Personal Communication.
- Neuts, Marcel F. 1973. *Probability*. Boston: Allyn and Bacon, Inc.
- Royston, J. P. 1982. An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Statist.* 31(2):115-124.
- Rying, E.A., R.S. Gyurcsik, J.C. Lu, G. Bilbro, G. Parson, and F. Y. Sorrell. 1997. Wavelet analysis of mass spectrometry signals for transient event detection and run-to-run process control. In *Proceedings of the Second International*

- Symposium on Process Control, Diagnostics, and Modeling in Semiconductor Manufacturing*, ed. M. Meyyappan, D. Economou, and S. W. Bulter, 37-44.
- Saito, N. 1994. *Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion*. In *Wavelets in Geophysics*, ed. E. Foufoula-Georgiou and P. Kumar, 299-324. New York: Academic Press.
- Stewart, Alan, and J. Keith Ord. 1987. *Kendall's Advanced theory of statistics*. Vol. 1, *Distribution Theory*. New York: Oxford University Press.
- Tedder, L.L., G.W. Rubloff, B.F. Cohaghan, and G.N. Parsons. 1996. Dynamic rate and thickness metrology during poly-Si rapid thermal chemical vapor deposition from SiH_4 using real time *in situ* mass spectrometry. *J. of Vac. Sci. and Tech. A* 14(2):267-270.
- Tew, J.D., and J.R. Wilson. 1992. Validation of simulation analysis methods for the Schruben-Margolin correlation-induction strategy. *Operations Research* 40:87-103.
- Venables, W.N., and B.D. Ripley. 1994. *Modern applied statistics with S-PLUS*. New York: Springer.

Appendices

Appendix A. Computation of the Probability the Bootstrap Sample Covariance Matrix is Nonsingular

The probability $\hat{\Sigma}^*$ is nonsingular is equal to the probability there are at least $p = 19$ linearly independent vectors in the bootstrap sample $\{\beta_k^*, k = 1, 2, \dots, n\}$, which is less than or equal to the probability there are at least $p = 19$ different vectors in the bootstrap sample. So if we compute the probability that there are at least $p = 19$ different vectors in the bootstrap sample, we will obtain an upperbound for the probability $\hat{\Sigma}^*$ is nonsingular.

STEP 1. Given an original sample $\{\beta_1, \beta_2, \dots, \beta_n\}$ consisting of n distinct objects, the number of different objects x that could be selected to fill the n slots defining the bootstrap sample of size n is given by

$$\binom{n}{x} = \frac{n!}{(n-x)!x!},$$

which is the number of combinations (without repetitions) of n objects taken x at

a time (Neuts 1973).

STEP 2. After selecting x different objects, we see that the number of different ways we can select the (unordered) composition of the bootstrap sample of size n so that there is at least one occurrence of each of the x different objects is

$$\binom{n-1}{x-1} = \frac{(n-1)!}{(n-x)!(x-1)!}.$$

The above formula is the number of combinations with repetitions of x symbols taken n at a time which contain at least one of each of the x symbols, as described in Theorem 2.1.5 of Neuts (1973). Note that at this point we have simply determined the number of copies of each of the x symbols that will go into the bootstrap sample of size n ; we have not yet determined how to assign all of these copies to the n slots defining the final bootstrap sample.

STEP 3. Now for $j = 1, \dots, \binom{n-1}{x-1}$, let $k_i(j)$ be the number of occurrences of the i th symbol in the j th combination with repetitions described in Step 2 so that we have

$$\sum_{i=1}^x k_i(j) = n$$

and

$$k_i(j) \geq 1 \text{ for } i = 1, \dots, x \text{ and for } j = 1, \dots, \binom{n-1}{x-1}.$$

Thus, the x -tuple $[k_1(j) k_2(j) \dots k_x(j)]$ defines the number of copies of each of the x different objects that will make up the unordered set of objects in the final bootstrap sample.

STEP 4. Now for x different objects, of which we have $k_i(j)$, $j = 1, \dots, \binom{n-1}{x-1}$, identical copies of the object of type i for $i = 1, \dots, x$, the number of distinct ways in which the $n = \sum_{i=1}^x k_i(j)$ objects can be ordered in the n slots defining the final bootstrap sample is given by the multinomial coefficient

$$\frac{n!}{[k_1(j)]! [k_2(j)]! \dots [k_x(j)]!},$$

as outlined in Theorem 2.1.9 of Neuts (1973). In other words, given x different symbols, there are

$$\sum_{j=1}^{\binom{n-1}{x-1}} \frac{n!}{\prod_{i=1}^x [k_i(j)]!}$$

different ways of assigning these symbols to n slots so that each symbol occurs at least once.

STEP 5. When we start with an original sample consisting of n different objects, the number of different bootstrap samples of size n that we can take from this original sample is n^n . Putting the results of Steps 1 – 4 above together, we obtain the final formula for the exact probability that from an original sample of size n we can randomly sample a bootstrap sample that contains at least p different objects. This probability is

$$\frac{\sum_{x=p}^n \binom{n}{x} \sum_{j=1}^{\binom{n-1}{x-1}} \frac{n!}{\prod_{i=1}^x [k_i(j)]!}}{n^n},$$

which is less than or equal to

$$\frac{\sum_{x=p}^n \binom{n}{x} \binom{n-1}{x-1} n!}{n^n}.$$

STEP 6. For $n = 21$ and $p = 19$, the probability of sampling with replacement at least p different vectors from a sample of n distinct vectors is less than or equal to

$$\frac{\sum_{x=19}^{21} \binom{21}{x} \binom{20}{x-1} 21!}{21^{21}} = 0.0003526.$$

Therefore, an upperbound for the probability the bootstrap sample covariance matrix $\hat{\Sigma}^*$ is nonsingular is 3.526×10^{-4} .