Abstract

With the rapid growth of data-acquisition technology and computing resources, a plethora of data can now be collected at high frequency. Because a large number of characteristics or variables are collected, interdependency among variables is expected and hence the variables are correlated. As a result, multivariate statistical process control is receiving increased attention. This thesis addresses multivariate quality control techniques that are capable of detecting covariance structure change as well as providing information about the real nature of the change occurring in the process. Eigenspace analysis is especially advantageous in data rich manufacturing processes because of its capability of reducing the data dimension.

The eigenspace and Cholesky matrices are decompositions of the sample covariance matrix obtained from multiple samples. Detection strategies using the eigenspace and Cholesky matrices compute second order statistics and use this information to detect subtle changes in the process. Probability distributions of these matrices are discussed. In particular, the precise distribution of the Cholesky matrix is derived using Bartlett’s decomposition result for a Wishart distribution matrix. Asymptotic properties regarding the distribution of these matrices are studied in the context of consistency of an estimator. The eigenfactor, a column vector of the eigenspace matrix, can then be treated as a random vector and confidence intervals can be established from the given distribution.

In data rich environments, when high correlation exists among measurements, dominant eigenfactors start emerging from the data. Therefore, a process monitoring strategy using only the dominant eigenfactors is desirable and practical. The applications of eigenfactor analysis in semiconductor manufacturing and the automotive industry are demonstrated.
Acknowledgments

It is a nice feeling to be able to write acknowledgments, remembering all the people that have been part of this long journey and have touched my life in so many ways. It has been a wonderful experience; I will cherish it for years to come.

I want to thank my advisor, Professor Duane Boning, for giving me the opportunity to work with him when I was at a cross road in my career. His support, insight and faith in me were the keys to the success of this work. He never ceases to amaze me with his ability to keep things in perspective and his wealth of knowledge in several areas.

I would like to thank my co-advisor, Professor Roy Welsch, from whom I have learnt so much in the field of applied statistics. His feedback and constructive criticism have taught me not to settle for anything less than what is best. Working as a TA by his side has allowed me to appreciate that teaching is a two-way action: to teach and to be taught.

I am grateful to Professor Eric Feron and Professor John Deyst for serving on my committee and taking time to read my thesis.

The support of several friends at MTL has made the process an enjoyable experience. I will always remember the many stimulating conversations, either research related or just related to everyday life, I had with Aaron. I would also like to thank him for keeping the computers in the office running smoothly during my thesis writing. I am also thankful to Dave and Brian Goodlin for several great discussions on endpoint detection, Taber for showing me the meaning of “productivity” and the art of managing time to get 48 hours a day, and my extra-curricular activity companions Sandeep, Stuckey, Minh and Angie. Thanks also to our younger office mates Joe, Allan, Mike and Karen (who is already thinking of taking over my cubicle) for making our office more lively with their enthusiasm, Brian Lee, Tamba and Tae for their support, and in particular, Vikas for keeping me under control during my defense.

Throughout my graduate studies, my closest friends Joey and Yu-Huai have provided me a shelter away from my worries. Their friendship and companionship have allowed me to take breaks and recharge myself as needed. I am grateful to all my MIT softball teammates for sharing the championships and glory over the past two years in the Greater Boston area and New England tournaments, especially to my friend Ken for being my sports-buddy for so many years.

My deepest thanks to Maria for being a source of both mental and emotional support during the toughest time in my graduate work. Her belief in me has propelled me to derive the results and complete my thesis. I could not have done without her patience and encouragement, which have been vital during my thesis writing and defense.

I would like to thank my parents whose constant support and understanding have paved the way for me to get where I stand today. They have opened my eyes by providing me with the opportunity to experience different cultures and to adapt to new environments. A special thanks to my brother and sisters for putting up with me and for always sharing with me all the wonderful experiences that life offers.

Finally, I would like to thank the financial support of LFM/SIMA Remote Diagnostic Program, LFM Variation Reduction Research Program and LFM GM Dimensional Analysis Project.
# Table of Contents

**Chapter 1. Introduction** .................................................................13
  1.1 Problem Statement and Motivation .............................................. 13
  1.2 Thesis Outline ........................................................................ 14

**Chapter 2. Background Information on Multivariate Analysis** ........17
  2.1 Control Chart and its Statistical Basis ........................................ 18
  2.2 Multivariate Quality Control: $\chi^2$ and Hotelling’s $T^2$ statistic . 20
    2.2.1 Examples of Univariate Control Limits and
    Multivariate Control Limits ......................................................... 22
  2.3 Aspects of Multivariate Data ...................................................... 26
  2.4 Principal Components Analysis .................................................. 27
    2.4.1 Data Reduction and Information Extraction .......................... 29
  2.5 SPC with Principal Components Analysis ................................... 31
  2.6 Linear Regression Analysis Tools ................................................ 34
    2.6.1 Linear Least Squares Regression ........................................ 35
    2.6.2 Principal Components Regression ....................................... 36
    2.6.3 Partial Least Squares ......................................................... 37
    2.6.4 Ridge Regression ............................................................... 38

**Chapter 3. Second Order Statistical Detection: An Eigenspace Method** 41
  3.1 Introduction ................................................................................. 41
  3.2 First Order Statistical and Second Order Statistical Detection Methods . 42
    3.2.1 First Order Statistical Detection Methods ........................... 42
    3.2.2 Second Order Statistical Detection Methods ....................... 44
  3.3 Weaknesses and Strengths in Different Detection Methods ............ 45
    3.3.1 Univariate SPC ................................................................. 45
    3.3.2 Multivariate SPC First Order Detection Methods ($T^2$) ........ 46
    3.3.3 PCA and $T^2$ methods ...................................................... 47
    3.3.4 Generalized Covariance .................................................... 49
  3.4 Motivation Behind Second Order Statistical Detection Methods ....... 50
    3.4.1 Example 1 ........................................................................ 51
    3.4.2 Example 2 ........................................................................ 53
  3.5 Eigenspace Detection Method ...................................................... 56
    3.5.1 Distribution of Eigenspace Matrix $E$ .................................... 60
    3.5.2 Asymptotic Properties on Distribution of the Eigenspace Matrix $E$
Chapter 4. Monte Carlo Simulation of Eigenspace and Cholesky Detection Strategy .................................................................79

4.1 Introduction .............................................................................79

4.2 Simulated Distributions of Eigenspace Matrix and Cholesky Decomposition Matrix .........................................................82
  4.2.1 Example 1: N=20,000, n=50 and k=100 times .................84
  4.2.2 Example 2: N=20,000, n=500 and k=500 times .............88
  4.2.3 Example 3: N=20,000, n=500 and k=500 times (8-variate) 94

4.3 Approximation of Distribution of E and M for Large n ............99
  4.3.1 Example 1: p=2, n=3, k=100 times ..............................100
  4.3.2 Example 2: p=2, n=100, k=100 times ..........................102
  4.3.3 Example 3: p=10, n=500, k=500 times ......................102

4.4 Sensitivity Issues: Estimation of Dominant Eigenfactors for Large p ...........................................................105
  4.4.1 Example 1: .................................................................106
  4.4.2 Example 2: .................................................................110

4.5 Oracle Data Simulation ..........................................................112
  4.5.1 Example 1: .................................................................113
  4.5.2 Example 2: .................................................................118

Chapter 5. Application of Eigenspace Analysis ....................................................125

5.1 Eigenspace Analysis on Optical Emission Spectra (OES) ..........125
  5.1.1 Optical Emission Spectra Experiment Setup ....................126
  5.1.2 Endpoint Detection ......................................................128
  5.1.3 Motivation for Application of Eigenspace Analysis to Low Open Area OES ....................................................129
  5.1.4 Low Open Area OES Endpoint Detection Results ...........131
  5.1.5 Control Limit Establishment for the Eigenspace Detection Method ..........................................................138

5.2 Eigenspace Analysis on Automotive Body-In-White Data ...........142
  5.2.1 Vision-system Measurements ........................................143
  5.2.2 Out of Control Detection Using PCA and T^2 Technique on the Left Door Data ....................................................145
  5.2.3 Eigenspace Analysis ......................................................148

Chapter 6. Conclusions and Future Work .......................................................153

6.1 Summary .................................................................153

6.2 Directions for Future Research ............................................154

References ..........................................................................................159
List of Figures

Chapter 1.

Chapter 2.

Figure 2-1: A typical control chart ..............................................19
Figure 2-2: Multivariate statistical analysis vs. univariate statistical
analysis. .................................................................22
Figure 2-3: Control limits for multivariate and univariate methods ........24
Figure 2-4: Graphical interpretation of PCA .................................31
Figure 2-5: Graphical interpretation of $T^2$ and Q statistics ...............33

Chapter 3.

Figure 3-1: $T^2$ statistic drawback ..............................................47
Figure 3-2: Drawback of $T^2$ and $T^2$ with PCA ..........................49
Figure 3-3: Two different covariance matrices in 2-D ..................50
Figure 3-4: $T^2$ values for example 1 ........................................52
Figure 3-5: A second order static detection method .......................53
Figure 3-6: Generalized variance for example 2 ...........................54
Figure 3-7: A second order statical detection method ....................55
Figure 3-8: Possibilities of two different population in 2-D ............56
Figure 3-9: Selection procedure of a unique eigenvector ...............60

Chapter 4.

Figure 4-1: Eigenfactor used to detect variance volume and
eigenvector angle change ..............................................81
Figure 4-2: Normal quantile plot for 100 samples and Q-Q plot for
eigenfactor elements ..............................................85
Figure 4-3: Q-Q plot for Cholesky decomposition matrix $M$ ...........86
Figure 4-4: Joint Distributions of eigenfactor 1 in $E$ .......................87
Figure 4-5: Joint Distributions of column one of $M$ ........................87
Figure 4-6: Normal plot and Q-Q plots for eigenfactor elements ............. 89
Figure 4-7: Q-Q plot for Cholesky Decomposition Matrix $M$ .................. 90
Figure 4-8: Joint distribution of eigenfactor 1 ................................. 91
Figure 4-9: Joint Distribution of Cholesky decomposition matrix $M$
column 1 ................................................................. 92
Figure 4-10: Normal plot and Q-Q plots for eigenspace .................. 95
Figure 4-11: Q-Q plot for Cholesky Decomposition Matrix $M$ ............. 96
Figure 4-12: Normal Q-Q plot of predicted and sampling distribution of
$E$ and $M$ ........................................................................ 98
Figure 4-13: Normal Quantile plot for $t_{ii}$ .................................... 100
Figure 4-14: Normal Quantile plot ................................................. 102
Figure 4-15: Normal Quantile Plot of $t_{ii}$ (high dimension) ............. 103
Figure 4-16: Normalization of d and total standard deviation ............ 108
Figure 4-17: Normalized 2-norm and total standard deviation as
function of $p$ .................................................................. 109
Figure 4-18: Confidence interval for individual eigenfactor ............ 113
Figure 4-19: The $T^2$ statistic detection .......................................... 114
Figure 4-20: Generalized variance detection using 100 samples window .. 115
Figure 4-21: Eigenspace detection ................................................. 116
Figure 4-22: Cholesky detection .................................................... 117
Figure 4-23: PCA with $T^2$ and Q plots ........................................... 119
Figure 4-24: Generalized variance with 100-sample window .......... 120
Figure 4-25: Eigenspace detection using two columns with
100-sample window ......................................................... 121
Figure 4-26: Cholesky detection using the first column with
100-sample window ......................................................... 121

Chapter 5.
Figure 5-1: Optical emission spectroscopy experiment setup .......... 126
Figure 5-2: Time evolution of spectral lines in an oxide
plasma etch process. ......................................................... 127
Figure 5-3: Two spectral lines showing different behavior as endpoint is reached. ........................................128
Figure 5-4: Plasma etch endpoint is reached when the intended etch layer (oxide) is completed removed ............129
Figure 5-5: Euclidean distance of $(E_1 - E_{ep1})$ in run 4 ..................................................132
Figure 5-6: Euclidean distance of $(E_1 - E_{ep1})$ in run 5 ..................................................133
Figure 5-7: Euclidean norm using characterization data from within the run (run 3) ....................................134
Figure 5-8: Euclidean norm of run 3 using characterization data from run 2 ..................................................134
Figure 5-9: Euclidean norm of run 4 using characterization data from run 2 ..................................................135
Figure 5-10: Euclidean norm of run 5 using characterization data from run 2 .................................................136
Figure 5-11: Euclidean distance for $n=30$ samples run 2 .................................................................137
Figure 5-12: Euclidean distance for $n=30$ samples for run 4 .................................................................138
Figure 5-13: Univariate control limits on multivariate normal data, applied to a selected eigenfactor ...............139
Figure 5-14: Euclidean norm of run 5 with control limit .................................................................141
Figure 5-15: Euclidean norm of run 5 using an overlapping sliding window of 50 samples .........................142
Figure 5-16: Possible vision-system measurements on a door ...............................................................144
Figure 5-17: Scree Plot for the left door data ........................................................................................145
Figure 5-18: The $T^2$ plot for the left door data ....................................................................................146
Figure 5-19: Scores and loading plots of record 50 (an out of control sample) ........................................147
Figure 5-20: Plot of standardized variable 20 across records ...............................................................148
Figure 5-21: The 2-norm of the difference between the sampling and population eigenfactor 1 ..................149
Figure 5-22: Individual contribution across variables ............................................................................150
Figure 5-23: Eigenspace analysis using window size of 50 samples ......................................................151
Chapter 6.

Figure 6-1: Multiway PCA ..................................................156
List of Tables

Chapter 1.

Chapter 2.

Chapter 3.

Chapter 4.
  Table 4-1: Properties of $t_{ij}$ from simulated distribution ....................104
  Table 4-2: Mean and variance from example 3 ($p=10$ and $n=500$) ............104
  Table 4-3: Summary of statistics as function of $p$ for eigenfactor 1 ............106
  Table 4-4: Summary of statistics as function of $p$ for eigenfactor 2 ............107
  Table 4-5: Summary of statistics as function of $p$ for eigenfactor 1 ............110
  Table 4-6: Summary of statistics as function of $p$ for eigenfactor 2 ............110

Chapter 5.

Chapter 6.
Chapter 1

Introduction

1.1 Problem Statement and Motivation

In large and complex manufacturing systems, statistical methods are used to monitor whether the processes remain in control. This thesis reviews and discusses both conventional methods and new approaches that can be used to monitor manufacturing processes for the purpose of fault detection and diagnosis. On-line statistical process control (SPC) is the primary tool traditionally used to improve process performance and reduce variation on key parameters. With faster sensors and computers, massive amounts of real-time equipment signals and process variables can be collected at high frequency. Due to the large number of process variables collected, these variables are often correlated. Consequently, multivariate statistical methods which provide simultaneous scrutiny of several variables are needed for monitoring and diagnosis purposes in modern manufacturing systems. Thus, multivariate statistical techniques have received increased attention in recent research. Furthermore, data reduction strategies such as projection methods are needed to reduce the dimensionality of the process variables in data rich environments.

SPC has strong ties with input-output modeling approaches such as response surface methods (RSM). In order to build models for prediction, it is important to make sure that all the experiment runs are under statistical process control. Once one is confident in the prediction model, deviation of production measurement from the prediction could indicate process drift or other disturbances. In the case of process drift, adaptive modeling could be
used to include effects introduced by slowly varying processes. The purpose of RSM is to identify the source of product quality variation, in other words, to discover which in-line data contributes to end-of-line data variation. One can use experimental design approaches with projection methods to build statistical models between in-line data and end-of-line data. Both partial least squares (PLS) and principal components regression (PCR) are parametric regression techniques, and assume there is only one functional form to characterize the whole system; thus PLS and PCA can be labeled as “global modeling methods.” These global modeling methods impose strong model assumptions that restrict the potential complexity of the fitted models, thereby losing local information provided by the sample. Finally, the RSM model can be used for optimization to minimize quality variation.

The goal of this thesis is to develop a multivariate statistical process control methodology that is capable of localized modeling. The eigenspace detection strategy computes the localized model from the test data and compares that with the model characterized using the training data. In essence, this approach allows us to compare the subspace spanned by the test data with an existing subspace. Moreover, the eigenspace analysis enables us to detect covariance and other subtle changes that are occurring in the process. Finally, this detection strategy inherits nice properties such as data compression and information extraction from the projection methods and factor analysis, and it is efficient when used in data rich environments; i.e. using a few eigenfactors is often sufficient to detect abnormality in the process.

1.2 Thesis Outline

Chapter 1 briefly overviews the state of the art in manufacturing process control strategies. Traditional multivariate analysis tools are mentioned and the goal of the thesis is
defined.

Chapter 2 covers background information on statistical process control and multivariate quality control methods. Both univariate and multivariate process control are discussed. Moreover, an example with correlation between the variables is provided to demonstrate the need for a multivariate process control strategy. The motivation for an additional multivariate monitoring and detection strategy is discussed. Conventional data characterization and prediction tools are reviewed.

Key definitions and problem statements are provided in Chapter 3. Disadvantages and issues related to traditional statistical process control methods described in Chapter 2 are addressed. A new multivariate statistical detection method is developed and its purpose is discussed. Mathematical properties such as probability distributions and asymptotic behavior are derived.

Chapter 4 provides oracle/synthetic data simulations using the new multivariate detection methods. Several abnormalities are induced in the oracle data and it is desirable that those changes be detected. The results from the eigenspace analysis are compared with those obtained from the traditional detection methods. Moreover, discussion on approximation to facilitate the use of the eigenspace detection technique is provided. Sensitivity issues regarding estimation of a reduced set of eigenfactors are also addressed.

The focus of Chapter 5 is on applications using the new detection strategy. In particular, these applications include semiconductor manufacturing and the automotive industry. Results from traditional multivariate detection methods and the newly developed multivariate detection methods are compared.
Chapter 6 provides a summary of the thesis and suggestions for future research in massive data environments.
Chapter 2

Background Information on Multivariate Analysis

In any production process, it is certain that there will be some degree of “inherent or natural variability.” However, other kinds of variability may occasionally be present. This variability usually arises from three sources: machine errors, operator error, or defective raw materials. Such variability is generally large when compared to the natural variability (background noise), and it usually represents an unacceptable level of process performance. These kinds of variability are referred to as “assignable causes,” and a process that is operating in the presence of assignable causes is said to be out of control.

In complex manufacturing processes, Statistical Process Control (SPC) [Mon91] has become very important due to its ability to achieve tight process control over the critical process steps. The objective of SPC is to monitor the performance of a process over time in order to detect any costly process shifts or other non-random disturbances. Historically, SPC has been used with process measurements in order to uncover equipment and process problems. The essential SPC problem-solving tool is the control chart to monitor if the manufacturing processing remains in a stable condition. With Hotelling’s $T^2$ statistic [Alt84], [MK95], multivariate statistical process control based on the $T^2$ statistic extends traditional univariate Shewhart, CUSUM and EWMA control charts [Mon91]. By dealing with all the variables simultaneously, multivariate methods not only can extract information on the directionality of the process variations, but also can reduce the noise level
through averaging.

### 2.1 Control Chart and its Statistical Basis

There is a strong tie between control charts and hypothesis testing. In essence, the control chart is a test of the hypothesis that the process is in a state of statistical control. A point on a chart within the control limits is equivalent to failing to reject the hypothesis of statistical control, and a point plotting outside the control limits is equivalent to rejecting the hypothesis that the process is in statistical control. Similar to hypothesis testing, probability of type I and II errors can be established in the context of control charts. The type I error of the control chart is to conclude the process is out of control when it is really in control, and the type II error is to conclude the process is in control when it is really out of control. A typical control chart is shown in Figure 2-1, which displays a quality characteristic that has been measured or computed from a sample. In this case, the sample characteristic is mean-centered at 55 with standard deviation of 1. The control limits are chosen to be $\pm 3\sigma$ (where $\sigma = 1$), so with probability of 99.73% a sample falls within the control limits; in other words, if the process is indeed under control on average 27 false alarms (or type I error) out of 10,000 samples are generated.
Though control charts are mainly used for monitoring purpose after a process has been characterized as in the state of control, control charts can also be used to improve the process capability. It is found in general that most processes do not operate in a state of statistical control. Therefore, the use of control charts will identify assignable causes and if these causes can be eliminated from the process, variability will be reduced and the process will be improved.

It is standard practice to control both the mean and variation of a quality characteristic. We can then design two control charts; one monitors the central tendency of the process and is called the $\bar{x}$ chart (see [Mon91]). The other chart monitors the variability of the process. Two common control charts serve this purpose, the control chart for the standard deviation (S chart), or the control chart for the range (R chart) [Mon91]. The $\bar{x}$ and R or S
control charts are called variables control charts and are among the most important and useful on-line statistical process control techniques.

When there are several quality characteristics, separate $\bar{x}$ chart and R (or S) charts are maintained for each quality characteristic. However, when there are thousands of quality characteristics to keep track of, the task of maintaining all the control charts can be cumbersome. Moreover, the information extracted from an individual control chart can sometimes be misleading because the correlation among quality characteristics is ignored.

2.2 Multivariate Quality Control: $\chi^2$ and Hotelling’s $T^2$ statistic

Because of rapid sensor advancement and modern manufacturing systems’ complexity, more and more process measurements can now be collected at a high frequency. As a result, multivariate statistical methods are very much desired. One of the key messages of multivariate analysis is that several correlated variables must be analyzed jointly. One such example can be found in the automotive industry where correlation exists among different measurements taken from the rigid body of an automobile.

By dealing with all of the variables simultaneously, multivariate quality control methods not only can extract information on individual characteristics, but also can keep track of correlation structure among variables. Univariate control chart monitoring does not take into account that variables are not independent of each other and their correlation information can be very important for understanding process behavior. In contrast, multivariate analysis takes advantage of the correlation information and analyzes the data jointly.

The difficulty with using independent univariate control charts can be illustrated in Figure 2-2. Here we have two quality variables ($x_1$ and $x_2$). Suppose that, when the pro-
cess is in a state of statistical control where only natural variation is present, $x_1$ and $x_2$ follow a multivariate normal distribution and are somehow correlated as illustrated in the joint plot of $x_1$ versus $x_2$ in Figure 2-2. The ellipse represents a contour for the in-control process with 95% confidence limits; both dots (•) and $x$ represent observations from the process. The same observations are also plotted in Figure 2-2 as individual Shewhart charts on $x_1$ and $x_2$ with their corresponding upper (UCL) and lower (LCL) control limits (roughly 95% confidence limits). Note that by inspection of each of the individual Shewhart charts the process appears to be in a state of statistical control, and none of the individual observations gives any indication of a problem. However, a customer could complain about the performance of the product corresponding to the $x$ points, as the product is in fact different than expected. If only univariate charts were used, one would not detect the problem. The true situation is only revealed in the multivariate $x_1$ and $x_2$ plot where it is seen that the $x$ observations are outside the joint confidence region (with the corresponding covariance structure) and are thus different from the normal in-control population of products.
2.2.1 Examples of Univariate Control Limits and Multivariate Control Limits

In this section, we illustrate the advantage of the multivariate over univariate method through examples. Let $p$ be the number of quality characteristics/variables. We start with two process variables $x_1$ and $x_2$, which translates into a two-dimensional plot making graphical interpretation plausible. Here, we sample both $x_1$ and $x_2$ coming from a multivariate normal distribution of mean zero and covariance matrix $\Sigma_0 = \begin{bmatrix} 9 & 4 \\ 4 & 4 \end{bmatrix}$; the correla-
ation between $x_1$ and $x_2$ is $\rho = \frac{4}{\sqrt{9} \times \sqrt{4}} = \frac{2}{3}$. Control charts of these two quality characteristics independently can be very misleading. Consider the special case where variables are independent of each other; the confidence intervals of individual variables ignoring the covariance structure are

$$0 - Z_{\alpha/2} \frac{\sigma_{x_1}}{\sigma_{x_1}} \leq x_1 \leq 0 + Z_{\alpha/2} \frac{\sigma_{x_1}}{\sigma_{x_1}}$$

$$0 - Z_{\alpha/2} \frac{\sigma_{x_2}}{\sigma_{x_2}} \leq x_2 \leq 0 + Z_{\alpha/2} \frac{\sigma_{x_2}}{\sigma_{x_2}}$$

where $\alpha$ is the probability of type I error and $z_{\alpha/2}(\cdot)$ is the percentage point of the standard normal distribution such that $\text{Prob}(z \geq Z_{\alpha/2}) = \frac{\alpha}{2}$. Since the observations on the $x_1$ are independent of those on $x_2$, the probability of all intervals containing their respective $x_i$ can be assessed using the product rule for independent events and

$$\text{Prob(both z-intervals contain the } x_i \text{'s)} = (1 - \alpha)(1 - \alpha) = (1 - \alpha)^2$$

If $\alpha=0.05$, then this probability is $(1-.05)^2=0.9025$; and the type I error ([WM93]) under the independence assumption is now $\alpha' = 1 - (1 - \alpha)^2 = 0.0975$. The type I error has become 0.0975 instead of 0.05. One can see that the distortion in using univariate control intervals applied to multivariate data continues to increase as the number of quality variables increases. Therefore, the number of false alarms (type I error) can be much too frequent since $\alpha' = 1 - (1 - \alpha)^p$ as $p$ increases; for $\alpha=0.05$ and $p=10$, we have type I error $\alpha' = 1 - (1 - 0.05)^{10} = 0.40$. In order to rectify such problems but still use univariate
charts, one needs to increase the control limits by using Bonferroni limits [JW98]. The Bonferroni limits are chosen to be large so that the test will reduce the false alarms; this however could decrease the power of our test. Figure 2-3 shows the limits for the univariate scenario and multivariate scenario; the dotted lines are the univariate Bonferroni limits, and the solid lines are the regular univariate control limits with type I error equal to 0.05 for each variable. The thicker lines are the multivariate limits with overall type I error equal to 0.05, and the control limits are calculated from \( x'\Sigma_0x \geq \chi^2_{2, 0.05} = 5.99 \).

**Figure 2-3: Control limits for multivariate and univariate methods**

We have simulated 1000 samples from the given covariance matrix ten times and, the average number of false alarms provided by the multivariate limits is \( 47.3 \pm 6.70 \) (where
6.70 is the standard deviation), which is close to 50 expected from 5% type I error. The average number of false alarms given by the regular univariate control limits is 84.6 ± 9.51, and the false alarms given by Bonferroni control limits are 40.6 ± 6.45. So the number of false alarms is reduced by 44.09% going from the regular control limits to multivariate control limits. This difference is even more significant when we have five quality variables (p=5). For the following covariance matrix, the 5% multivariate control limits give us on the average 52.3 ± 7.27 false alarms in 1000 samples. The regular individual control limits with 5% on each variable produce on the average 199 ± 10.27 false alarms in 1000 samples, so using multivariate limits reduces the false alarms by 73.7%.

Bonferroni limits (α = \( \frac{0.05}{p} = \frac{0.05}{5} = 0.01 \)) produce on the average 49.9 ± 6.10 false alarms per 1000 samples.

Though Bonferroni limits reduce the number of false alarms, we can show that the power detection using Bonferroni limits may be reduced significantly when given an alternative hypothesis. Graphically, this can be seen in Figure 2-3; sample points denoted by \( x \) in region B are out of control samples not detected using Bonferroni limits. Now we need to establish an alternative hypothesis so that we can examine type II error (\( \beta \)), which is one minus the power of a test (\( \beta = 1 - \text{Power} \)). Assume that observations could be coming from another population with different mean but the same covariance structure, i.e.

\[
\begin{bmatrix}
9 & 4 & -3 & 2 & 6 \\
4 & 4 & -1 & 2 & 3 \\
-3 & -1 & 10 & 4 & -4 \\
2 & 2 & 4 & 8 & 5 \\
6 & 3 & -4 & 5 & 16
\end{bmatrix}
\]

with \( \mu = \begin{bmatrix} 2\sigma_1 \\ -2\sigma_2 \end{bmatrix} \) and the same covariance matrix \( \begin{bmatrix} 9 & 4 \\ 4 & 4 \end{bmatrix} \). This can be thought of as mean
drift from \( \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \). We can now find the type II error for each test. By simulating 1000 samples with the drifted mean and covariance matrix ten times, we find the average number of not detected out-of-control samples to be 5.3±2.45 samples (out of 1000 samples) for the multivariate test. The average number of type II errors for regular control limits is 121.5 ± 12.32 samples, and the number given by Bonferroni limits is 253.7 ± 19.12 samples. Therefore, the average number of type II errors when using Bonferroni limits is almost 50 times that of a full multivariate test. Note that the average number of type II errors given above strongly depends on the alternative hypothesis; however, the key factor for type II errors depends on the size of the area in region C and region B in Figure 2-3. Since the area in region B is much larger than that of C, type II errors for the Bonferroni limit test would be larger than that of a multivariate test for almost any alternative hypothesis (with the exception that one could construct an unusual probability density function such that it has very high probability in region C and near zero probability in region B).

### 2.3 Aspects of Multivariate Data

Throughout the thesis, we are going to be concerned with multivariate datasets. These datasets can frequently be arranged and displayed in various ways. Graphical representations and array arrangements are important tools in multivariate data analysis.

Usually, a multivariate dataset is analyzed using a two dimensional array, which results in a matrix form. We will use the notation \( x_{ij} \) to indicate the particular value of the \( i \)-th row and \( j \)-th column. Let \( p \) be the number of variables or characteristics to be recorded and \( n \) be the number of measurements collected on \( p \) variables. A multivariate dataset can then be presented by an \( n \times p \) matrix, where a single observation of all variables constitutes a row, and all \( n \) observations of a single variable are in the format of a column.
Therefore, the following matrix $\mathbf{X}$ contains the data consisting of $n$ observations on $p$ variables.

$$
\mathbf{X} =
\begin{bmatrix}
x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\
x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np}
\end{bmatrix}
$$

We also use the notation $\mathbf{X}_i$ to represent the $i$-th observation of all variables, i.e.

$$
\mathbf{X}_i = \begin{bmatrix}x_{i1} & x_{i2} & \ldots & x_{ip}\end{bmatrix}^T. \text{ As a result, the data matrix } \mathbf{X} \text{ can be written as}
$$

$$
\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_i & \ldots & \mathbf{X}_n\end{bmatrix}^T
$$

### 2.4 Principal Components Analysis

Principal components analysis (PCA) is used to explain the variance-covariance structure through a few linear combinations of the original variables. Principal components analysis is also known as a projection method and its key objectives are data reduction and interpretation, see [JW98] and [ShS96]. In many instances, it is found that the data can be adequately explained just using a few factors, often far fewer than the number of original variables. Moreover, there is almost as much information in the few principal components as there is in all of the original variables (although the definition of information can be subjective). Thus the data overload often experienced in data rich environments can be solved by observing the first few principal components with no significant loss of information. It is often found that PCA provides combinations of variables that are useful indicators of particular events or stages in the process. Because the presence of noise almost
always exists in a process, some signal processing or averaging is very much desirable. Hence, these combinations of variables from PCA are often a more robust description of process conditions or events than individual variables.

In massive datasets, analysis of principal components often uncovers relationships that could not be previously foreseen and thereby allows interpretations that would not ordinarily be found. A good example is that when PCA is performed on some stock market data, one can identify the first principal component as the general market index (average of all companies) and the second principal component can be the industry component that shows contrast among different industries.

Algebraically, PCA relies on eigenvector decomposition of the covariance or correlation matrix from variables of interest. Let \( x \) be a random vector with \( p \) variables, i.e.

\[
x = \begin{bmatrix} x_1 & x_2 & \ldots & x_j & \ldots & x_p \end{bmatrix}^T.
\]

The random vector \( x \) has zero mean and a covariance matrix \( \Sigma \) with eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \), so that \( \Sigma = \mathbf{V} \Lambda \mathbf{V}^T \) where \( \mathbf{V} \) is the eigenvectors matrix and \( \Lambda \) is a diagonal matrix whose elements are the eigenvalues. Consider a new variable formed from a linear combinations of \( x_i \)

\[
z_1 = v_1^T x \tag{Eq. 2-1}
\]

Then the variance of \( z_1 \) is just \( \text{var}(z_1) = v_1^T \Sigma v_1 \). The first principal component is the linear combination which maximizes the variance of \( z_1 \), i.e. the first principal component maximizes \( \text{var}(z_1) \). Since the \( \text{var}(z_1) \) can always be increased by multiplying \( v_1 \) by some constant, it is then constrained that the coefficients of \( v_1 \) be unit length. To summarize, the first principal component is defined

\[
\max \text{var}(z_1) = v_1^T \Sigma v_1,
\]

subject to \( v_1^T v_1 = 1 \).
The solution to this problem can be solved using Lagrange multipliers and $v_1$ is the eigenvector associated with the largest eigenvalue $\lambda_1$ of the covariance matrix $\Sigma$ (see [ShS96], [JW98]). The rest of the principal components can then be found as the eigenvectors of the covariance matrix $\Sigma$ with eigenvalues in descending order. Therefore, in order to compute the principal components, we need to know the covariance matrix. In real life, the true covariance matrix of a population is often unknown, so a sample covariance matrix ($S$) computed from the data matrix $X$ is used to estimate the principal components.

An alternate approach to obtain principal components is to use singular value decomposition in the given data matrix $X=U\Sigma V^T=TV^T=t_1v_1^T + t_2v_2^T + \ldots + t_pv_p^T$, where $t_i$, which is also known as the score, is the projection of the given data matrix onto the $i$-th principal component. In this scenario, PCA decomposes the data matrix as the sum of the inner product of vector $t_i$ and $v_i$. With this formulation, $v_1$ can be shown to capture the largest amount of variation from $X$ and each subsequent eigenvector captures the greatest possible amount of variance remaining after subtracting $t_1v_1^T$ from $X$.

2.4.1 Data Reduction and Information Extraction

One graphical interpretation of principal components analysis is that it can be thought of as a coordinate transformation where the transformation allows principal components to be orthonormal to each other. Hence the principal components are uncorrelated to each other. This transformation is especially useful when one is dealing with a multivariate nominal distribution since uncorrelatedness is equivalent to independence for normal random variables. Furthermore, such a transformation allows us to interpret the data using the correlation structure. Consider the example of stock data, where we monitor weekly six different stocks, three from the oil industry and three chemical companies. We might be
able to summarize the data just using two principal components, one can be called the oil industry component and the other called the chemical industry component. It is generally found that massive data contains redundant information because of highly correlated variables. Thus, the data can be compressed in such a way that the information is retained in the reduced dimension. In real practice, PCA also helps to eliminate noise from the process, so PCA serves as a useful tool for noise filtering.

Graphical interpretation of principal components analysis is found in Figure 2-4. In this example there are two normal random variables ($x_1$ and $x_2$) measured on a collection of samples. When plotted in two dimensions, it is apparent that the samples are correlated and can be enclosed by an ellipse. It is also apparent that the samples vary more along one axis (semi-major axis) of the ellipse than along the other (semi-minor axis). From the correlation between the two variables, it seems that the knowledge of one variable provides substantial (and perhaps sufficient) information about the other variable. Therefore, monitoring the first principal component could give us most of the information about what is going on in the process, where by information in this context we mean the total variance. Furthermore, the second principal component can be thought of as a noise factor in the process, and one may chose to ignore or neglect it in comparison to the first principal component.
2.5 SPC with Principal Components Analysis

With hundreds or thousands of measured variables, most recent multivariate SPC methods have been focused on multivariate statistical projection methods such as principal components analysis (PCA) and partial least squares (PLS). The advantages of projection methods are the same as those discussed in PCA. The key advantages include data reduction, information extraction and noise filtering. These projection methods examine the behavior of the process data in the projection spaces defined by the reduced order model, and provide a test statistic to detect abnormal deviations only in the space spanned by a subset of principal components or latent variables. Therefore, projection methods must be used with caution so that these methods can keep track of unusual variation inside the model as well as unusual variation outside the model (where a model is defined by the
number of principal components retained). The projection methods are especially useful when the data is ill conditioned since the filtering throws out the ill conditioned part of the data. Ill conditioning occurs when there is exact or almost linear dependency among the variables; more details of such condition will be discussed in Section 2.6.1.

The multivariate $T^2$ statistic can then be combined with PCA to produce just one control chart for easily detecting out of control sample points on a reduced dimension provided by the PCA model. Let us assume that $k$ out of $p$ principal components are kept for the PCA model. Because of the special mapping of PCA, each PC is orthogonal to every other. Therefore, $T^2$ is computed as the sum of normalized squared scores from the $k$ principal components and it is a measure of the variation from each sample within the PCA model from the $k$ principal components. It is calculated based on the following formula

$$T^2 = \sum_{i=1}^{k} \left( \frac{t_i}{s_i} \right)^2$$

where $s_i$ is the standard deviation associated with $t_i$.

However, in order to identify the underlying causal variables for a given deviation, one needs to go back to the loadings or eigenvectors of the covariance matrix. First, from the $T^2$ value of the out of control point, we can find the contribution from each score by plotting the normalized scores from $T^2 = \sum_{i=1}^{k} \left( \frac{t_i}{s_i} \right)^2$, where $k$ is the number of principal components kept in the model and $s_i$ is the standard deviation associated with the $i$-th principal component (see [KM96]). Control limits such as Bonferroni limits can then be used on the chart as rough guidelines for detecting large $\frac{t_i}{s_i}$ values. Once the dominant scores are determined, one can then identify the key contributing variables on those
scores. From principal components analysis, the scores are given by the following formula, where $v_i$ is the eigenvector corresponding the $i$-th principal component and $X$ is the mean-centered data matrix.

$$t_i = Xv_i \quad \text{(Eq. 2-2)}$$

The above equation provides the contribution of each variable $x_j$ to the scores of the $i$-th principal component as $v_{ij}x_j$.

Aside from tracking a $T^2$ statistic within the space spanned by PCA model, one must also pay attention to the residual between the actual sample and its projection onto the PCA model. The Q statistic does this [WRV90]; it is simply the sum of squares of the error:

$$Q = xx^T - xVV^Tx^T = x(I - VV^T)x^T \quad \text{(Eq. 2-3)}$$

The Q statistic indicates how well each sample conforms to the PCA model.

**Figure 2-5: Graphical interpretation of $T^2$ and Q statistics**
Figure 2-5 provides graphical interpretation of principal components analysis, $T^2$ and $Q$ statistics. Though the data resides in a 3-D environment, most of the data, except one sample (point $x$), lie in a plane formed by the vectors of PC #1 and PC #2. As a result, a PCA model with two principal components adequately describes the process/data. The geometrical interpretation of $T^2$ and $Q$ is also shown in the figure. In this case, $T^2$ is a squared statistical distance within the projection plane (see o point). On the other hand, $Q$ is a measure of the variation of the data outside of the principal components defined by the PCA model. From the figure, $Q$ is the squared statistical distance of the $x$ point (see Eq. 2-3) off the plane containing the ellipse. Also note, a point can have a small $T^2$ value because its projection is well within the covariance structure, yet its $Q$ value can be large as for the point $x$ in Figure 2-5.

2.6 Linear Regression Analysis Tools

In many design of experiment setups, we wish to investigate the relationship between a process variable and a quality variable. In some cases, the two variables are linked by an exact straight-line relationship. In other cases, there might exist a functional relationship which is too complicated to grasp or to describe in simple terms (see [Bro91] and [DS81]). In this scenario we often approximate the complicated functional relationship by some simple mathematical function, such as linear functions, over some limited ranges of the variables involved. The variables in regression analysis are distinguished as predictor/independent variables and response/dependent variables. In this section, we briefly give some background information on some popular regression tools. While this thesis focuses on correlation structures within a set of input or output data (rather than between input and output), regression analysis is often employed in a overall quality control methodology.
Our detection methods make heavy use of PCA and eigenspace methods, and some background is provided in this section on related regression methods so that the reader may understand the increasing importance of such eigenspace approaches in emerging data rich quality control environments.

The linear regression equations express the dependent variables as a function of the independent variables in the following way:

\[ Y = X\beta + \varepsilon \]  \hspace{1cm} (Eq. 2-4)

where \( Y \) denotes the matrix of response variables and \( X \) is the matrix of predictor variables. The error, \( \varepsilon \), is treated as a random variable whose behavior is characterized by a set of distribution assumptions.

### 2.6.1 Linear Least Squares Regression

The first approach we review is linear least squares regression. The objective is to select a set of coefficients \( \beta \) such that the Euclidean norm (also known as 2-norm) of the discrepancies \( \varepsilon = Y - X\beta \) is minimized. In other words, let \( S(\beta) \) be the sum of squared differences, \( S(\beta) = (Y - X\beta)^T(Y - X\beta) \). Then \( \beta \) is chosen by searching through all possible \( \beta \) to minimize \( S(\beta) \); this optimization is also known as the least squares criterion and its estimate is known as the least squares estimate ([DS81], [FF93]). Solution to this optimization can be solved using the normal equation. Let \( b \) be the least squares estimate of \( \beta \). We then can use the fact that the error vector \( (e = Y - Xb) \) is orthogonal to the vector subspace spanned by \( X \). Therefore, the solution is given by

\[ 0 = X^T e = X^T (Y - Xb) \Rightarrow X^T Y = X^T X b \Rightarrow b = (X^T X)^{-1} X^T Y \]  \hspace{1cm} (Eq. 2-5)

Least squares methods can be modified easily to weighted least squares, where weights are placed on different measurements. Such a weighting matrix is desirable when
the variances across different measurements are not the same, i.e. some prior information on the measurement can be included in the weighting matrix.

The least squares estimate requires that $X^TX$ be invertible; this might not be always the case. In the case when a column of the predictor matrix $X$ can be expressed as a linear combination of the other columns of $X$, $X^TX$ becomes singular and its determinant is zero. When dependencies hold only approximately, the matrix of $X^TX$ becomes ill conditioned giving rise to what is known as the multicollinearity problem (see [DS81] and [Wel00]). With massive amount of data, it is very possible that redundancy or high correlation exits among variables, hence multicollinearity becomes a serious issue in data rich environments.

2.6.2 Principal Components Regression

As its name suggests, the foundation of Principal Components Regression (PCR) is based on principal components. PCR is one of many regression tools that overcome the problem of multicollinearity. Multicollinearity occurs when linear dependencies exist among process variables or when there is not enough variation in some process variable. If that is the case, such variables should be left out since they contain no information about the process. One advantage of using principal components is that because all PCs are orthogonal to each other, multicollinearity is not an issue. Moreover, in PCA the variance equals information; so principal components with small variances can be filtered out and only a subset of the PCs are used as predictors in the matrix $X$. Using the loading matrix $V$ as a coordinate transformation, the resulting equation for PCR is

$$Y = X\beta + \epsilon = XVV^T\beta + \epsilon = Z\gamma + \epsilon$$

(Eq. 2-6)

where $Z$ is the projection of $X$ onto $V$ and is called the scores matrix, and $\gamma$ can be com-
puted using the least squares estimate equation in Eq. 2-5, i.e. \( \hat{Y} = (Z^T Z)^{-1} Z^T Y \).

The analysis above is done around principal components, namely \( Z \)'s. Though we can reduce the number of PCs in the analysis, all the original \( X \) variables are still present and none is eliminated by the above procedure. Ideally, one would like to eliminate those input variables in \( X \) which do not contribute to the model, so some sort of variables selection procedure can be done prior to the regression procedure.

### 2.6.3 Partial Least Squares

Partial Least Squares (PLS) has also been used in ill conditioned problems encountered in massive data environments. While PCA finds factors through the predictor variables only, PLS finds factors from both the predictor variables and response variables. Because PCA finds factors that capture the greatest variance in the predictor variables only, those factors may or may not have any correlation with the response variables. Because the purpose of regression is to find a set of variables that best predict the response variables, it is desirable then to find factors that have strong correlation with the response variables. Therefore, PLS finds factors that not only capture variance of predictor variables but also achieve correlation to the response variables. That is why PLS is described as a covariance maximizing technique ([FF93], [LS95]). PLS is a technique that is widely used in chemometrics applications.

There are several ways to compute PLS, however, the most instructive method is known as NIPLS for Non-Iterative Partial Least Squares. Instead of one set of loadings in PCR, there are two sets of loadings used in PLS, one for the input matrix and one for the response variables. The algorithm is described in the following steps:

1. Initialize: Pick \( Y_0 = Y, X_0 = X \)
2. For \( i = 1 \) to \( p \) do the following

3. Find \( i \)-th covariance vector of \( X_0 \) and \( Y_0 \) data by computing

\[
w_i = \frac{X_{i-1}^T Y_{i-1}}{\|X_{i-1}^T Y_{i-1}\|}
\]

4. Find \( i \)-th scores of \( X \) data by computing

\[
t_i = X_{i-1} w_i
\]

5. Estimate the \( i \)-th input loading

\[
p_i = \frac{X_{i-1}^T t_i}{\|t_i\|}
\]

6. Compute the \( i \)-th response loadings

\[
q_i = \frac{Y_{i-1}^T t_i}{\|t_i\|}
\]

7. Set \( X_i = X_{i-1} - t_i p_i^T \) and \( Y_i = Y_{i-1} - t_i q_i \)

8. End of loop.

From the vectors \( w_i, t_i \) and \( p_i \) found above, matrices of \( W, T \) and \( P \) are formed by

\[
W = \begin{bmatrix} w_1 & w_2 & \ldots & w_p \end{bmatrix}, \quad T = \begin{bmatrix} t_1 & t_2 & \ldots & t_p \end{bmatrix}, \quad P = \begin{bmatrix} p_1 & p_2 & \ldots & p_p \end{bmatrix}
\]

and the PLS estimate of \( \beta \) is

\[
\hat{\beta} = W (P^T W)^{-1} (T^T T)^{-1} T^T Y
\]

### 2.6.4 Ridge Regression

Ridge regression is intended to overcome multicollinearity situations where correlations between the various predictors in the model cause the \( X^T X \) matrix to be close to singular, giving rise to unstable parameter estimation. The parameter estimates in this case may either have the wrong sign or be too large in magnitude for practical consideration [DS81].
Let the data \( X \) be mean-centered (so one does not include the intercept term). The estimates of the coefficients in Eq. 2-4 are taken to be the solution of a penalized least squares criterion with the penalty being proportional to the squared norm of the coefficients \( \beta \):

\[
b = \arg\min_{\beta} E[(Y - X\beta)^2 + \gamma \beta^T \beta]
\]  

(Eq. 2-7)

The solution to the problem is

\[
b(\gamma) = (X^T X + \gamma I)^{-1} X^T Y
\]

The only difference between this and the solution to the least squares estimate is the additive term \( \gamma I \). This term stabilizes \( X^T X \), which then becomes invertible. \( \gamma \) is a positive number and in real applications the interesting values of \( \gamma \) are in the range (0,1).

We can examine ridge regression from the principal components perspective [Wel00]. In order to do so, we need to express \( X^T X \) in terms of principal components

\[
S = \frac{X^T X}{n-1} = V\Lambda V^T \Rightarrow X^T X = (n-1)V\Lambda V^T
\]

where \( n \) is the number of observations. Then \( (X^T X + \gamma I)^{-1} \) can be expressed as the following

\[
(X^T X + \gamma I)^{-1} = ((n-1)V\Lambda V^T + \gamma VV^T)^{-1} = V[(n-1)\Lambda + \gamma I]^{-1}V^T
\]  

(Eq. 2-8)

Eq. 2-8 provides some insight into the singular values of \( (X^T X + \gamma I)^{-1} \) in terms of the singular values of \( (X^T X)^{-1} \). Through ridge regression, the \( i \)-th singular value has been modified from \( \frac{1}{(n-1)\lambda_i} \) to

\[
\frac{1}{(n-1)\lambda_i + \gamma}
\]  

(Eq. 2-9)
From Eq. 2-9, if $\lambda_i$ is large, then $(n-1)\lambda_i$ dominates over $\gamma$ and \( \frac{1}{(n-1)\lambda_i + \gamma} \) becomes \( \frac{1}{(n-1)\lambda_i} \). When $\lambda_i$ is small (almost zero), then $\gamma$ dominates over $(n-1)\lambda_i$ and $\frac{1}{(n-1)\lambda_i + \gamma}$ becomes $\frac{1}{\gamma}$. Finally, the key difference between PCR and ridge regression is that PCR removes principal components with small eigenvalues (near zero), whereas ridge regression compensates the eigenvalues of those principal components by $\gamma$. 
Chapter 3

Second Order Statistical Detection: An Eigenspace Method

3.1 Introduction

The goal of this chapter is to introduce two new multivariate detection methods, an eigenspace and a Cholesky decomposition detection method, as well as their fundamental properties. Before introducing these new multivariate detection methods, we need to establish certain terminology that will be used throughout the thesis. These terms are sometimes defined differently according to the field of discipline, and we must be clear on our usage so that confusion will not arise in the following sections.

We then revisit all SPC methods mentioned in the previous chapter and discuss advantages and disadvantages associated with each method. Graphical and concrete examples are provided to enhance the understanding of what each detection method can do and can not do. Moreover, because the focus is on second order statistics, most of the examples provided in this chapter have to do with covariance/correlation shift and the goal is to detect such change in the examples. Motivation for a new multivariate detection method is then addressed. The new eigenspace detection method capable of detecting subtle covariance structure changes is then defined and specific multivariate examples are provided to illustrate the advantages of this new method.

Mathematical properties of the eigenspace detection method are derived. In particular, these properties include the distribution of the test statistic provided by the eigenspace
detection method and key asymptotic properties on this distribution. These properties are important, as control limits must be established for event detection using the new eigenspace detection method, and these limits of course depend strongly on the distribution of the new statistic.

3.2 First Order Statistical and Second Order Statistical Detection Methods

Almost all modern SPC methods are based on hypothesis testing. In practice, an independent normally distributed assumption is placed on the data. Moreover, data collected is placed in the matrix form $X$ defined in Section 2.3. Basic descriptive statistics such as sample mean and sample variance can then be computed from the random samples in $X$. Some of the descriptive statistics use only the first moment and are called first order statistics. Other descriptive statistics compute the second moment from the sample and are called second order statistics. Finally, these sample statistics are used in the hypothesis testing to arrive at a conclusion as to whether or not these data samples could be coming from a known distribution.

3.2.1 First Order Statistical Detection Methods

Before we discuss the definition of first order statistical detection methods, let us establish terminology for single sample and multiple sample detection methods. Single sample detection collects one sample at a time and a test statistic can then be computed from the sample; that statistic is then used to conclude if the sample conforms with the past data. Multiple sample detection methods use sample statistics computed from several samples for a hypothesis test.
In this section, we provide detailed discussion to distinguish a multivariate first order detection method from a multivariate second order detection method. First order detection methods extract only a first order statistic or first moment from a future sample or samples and use that first order statistic or moment to make inferences about some known parameters. However, the known parameters can include higher order statistics computed from the training or past data. We want to emphasize that the order of a statistical detection method is defined based on the statistic derived from a test sample or samples, and does not directly utilize the trial or historical samples. In other words, first order statistical methods compare the first moment from the test samples to the subspace spanned by the training data and determine if the test samples could be generated from the training data population. Note that first order statistical detection methods can be either single sample or multiple sample detection methods.

The sample mean control chart is used extensively and is an example of a first order detection method. Basically, independent identically distributed test random samples of size \( m \) are collected, \( X = [X_1, X_2, \ldots, X_i, \ldots, X_m]^T \), and the sample mean \( \bar{X} \) is computed. Hypothesis testing can then be performed, \( H_0 : \bar{X} = \mu_0 \) and its alternative hypothesis is \( H_1 : \bar{X} \neq \mu_0 \). We are trying to decide the probability that \( \bar{X} \) can be generated by a population whose mean is \( \mu_0 \). Bear in mind that \( \mu_0 \) is found from the historical data and is characterized as the population mean. Many hypothesis testing problems can be solved using the likelihood ratio test (see [JW98]). However, there is another more intuitive way of solving hypothesis testing problems, and it uses the concept of a statistical distance measure. In the univariate scenario, \( t = \frac{(\bar{X} - \mu_0)}{s} \) has a student’s t-distribution and is the statistical distance from the sample to the test value \( \mu_0 \) weighted by sample standard deviation \( s \).
By analogy, $T^2$ is a generalized statistical squared distance from $x$ to the test vector $\mu_0$, defined to be

$$
T^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)
$$

The $T^2$ statistic is also known as Hotelling’s $T^2$. Since it is a distance measure, if $T^2$ is too large, then $x$ is too far from $\mu_0$, hence the null hypothesis is rejected. Thus a standard $T^2$ statistic is a single sample first order detection method where $T^2$ is calculated based on a single new data point.

### 3.2.2 Second Order Statistical Detection Methods

In order to compute second order statistics, multiple samples (more than one sample) must be collected from the test data. Hence, second order statistics or moments are extracted from these samples and inferences can then be carried out using hypothesis testing. A well known second order statistic widely used in univariate SPC is the S (standard deviation) chart. In practice, the R (range) chart is often used, especially when the sample size is relatively small. Moreover, an estimate of standard deviation can be computed from the sample range $R$ [Mon91]. In multivariate data, individual S or R charts can be monitored when the dimension of the data is small. However, the problem becomes non-tractable when the data dimension increases rapidly as in modern data rich environments.

A widely used measure of multivariate dispersion is the sample generalized variance. The generalized variance is defined to be the determinant of the covariance matrix. This measure provides a way of compressing all information provided by variances and covariances into a single number. In essence the generalized variance measure is proportional to
the square of the volume spanned by the vectors in the covariance matrix [JW98], [Alt84], and can be thought of as a multidimensional variance *volume*.

Although the generalized variance has some intuitively pleasing geometrical interpretations, its weakness is similar to all descriptive summary statistics - lost information. In matrix algebra, several different matrices can generate the same determinant (that is, different covariance structures may generate the same generalized variance), geometrical interpretation of this problem will be illustrated in Section 3.3.4. Mathematically, however, we can see that the following three covariance matrices all have the same determinant:

\[
S_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}
\]

Although each of these covariance matrices has the same generalized variance, they possess distinctly different covariance structures. In particular, there is positive correlation between variables in $S_1$ and the correlation coefficient of $S_2$ is the same in magnitude as in $S_1$, but the variables in $S_2$ are negatively correlated. The variables in $S_3$ are independent of each other. Therefore, different correlation structures are not detected by the generalized variance. A better second order statistical detection method is desired which compares in more detail the subspace spanned by the test samples with the subspace generated by the training data, and this is the focus of this thesis.

### 3.3 Weaknesses and Strengths in Different Detection Methods

#### 3.3.1 Univariate SPC

The weakness of univariate SPC applied on multivariate data is discussed in
Section 2.2.1, and an example is illustrated in that section. Although univariate SPC is easy to use and monitor, it should only be used when the dimension of the data is small. Moreover, the correlation information is essential in multivariate analysis, yet univariate SPC does not use that information.

3.3.2 Multivariate SPC First Order Detection Methods ($T^2$)

The $T^2$ has been used extensively due to its attractive single value detection scheme using the generalized distance measure. The advantage of using the $T^2$ statistic is that it compresses all individual control charts on $x_i$ into a single control chart as well as keeps track of the correlation information among variables. The power of $T^2$ based detection can be boosted when multiple samples are collected; this is a direct result of the sample mean. As $n$ increases, the variance of sample mean decreases as $\frac{1}{n}$. Therefore, one can always resolve two populations with different means using a $T^2$ detection method with large enough $n$.

There are some computational issues related to $T^2$ methods. When data dimension increases, the multicollinearity issue can not be overlooked. Collinearity causes the covariance matrix to be singular, hence $T^2$ can not be computed. There are some ways to work around the problem: one can compute the Moore-Penrose generalized inverse, which is also known as the pseudo inverse. The covariance matrix also becomes non-invertible when the number of sequential samples collected exceeds the data dimension.

As with most of the data compression methods, $T^2$ gains from compression but also suffers from compression, i.e. because of compression some key information is lost. The generalized distance loses information on directionality, as depicted in Figure 3-1. In this example $T^2$ can not distinguish the difference between the out-of-control point on the left
and those on the right; they all have a large $T^2$ value.

Figure 3-1: $T^2$ statistic drawback

Therefore, $T^2$ is suitable if the purpose is only to detect out-of-control events. Second order statistical methods might be desirable when the lost information such as directionality can be used for both event detection and classification, or for process improvement purposes.

3.3.3 PCA and $T^2$ methods

PCA provides great advantage for data compression: instead of dealing with hundreds of variables, we are now dealing with a few principal components. However, the $T^2$ statistic only tracks the data in the projection hyperplane; one must also track the Q statistic in order to detect if the PCA model no longer describes the process. PCA and $T^2$ suffer the same problem as $T^2$ does; it is a type of statistical distance measure, so it can not resolve
the differences in directionality. Moreover, a $T^2$ based on PCA is not able to detect certain covariance structure changes. An example of this scenario is described in Figure 3-2. Here we have two populations; one of the populations has more variations in all direction hence a larger confidence ellipse volume. The other population has smaller variation, therefore a smaller ellipse. Furthermore, the smaller ellipse lies completely within the larger ellipse. Both populations are mean-centered in the same place. In this scenario, let us suppose that at the beginning all the sample points were coming from population 1, but due to maintenance or personnel shift sample points are now coming from the smaller region denoted as population 2. It is desirable to detect such a change since this information could lead us to improve the process capability.

Although the $T^2$ statistic can not detect the change described in Figure 3-2, the generalized covariance could be used to detect this type of change. Thus it is possible that some combination of statistical detection methods can give acceptable detection of some types of shift or covariance structure changes. Our purpose in this thesis is to provide single-statistic detection methods that enable both covariance structure change detection and classification.
3.3.4 Generalized Covariance

The main advantage of generalized covariance methods is that they collapse all the information of a matrix into a single number, so it is very easy to monitor this single number. However, generalized covariance is invariant under rotation, i.e. multiplication of the covariance matrix by a rotational matrix $U$ whose determinant is $1$. We then have

$$\det(\Sigma) = \det(U\Sigma) = \det(U)\det(\Sigma) = \det(\Sigma) \quad \text{(Eq. 3-1)}$$

Figure 3-3 shows problems of this nature in two dimensions. Variables in one of the populations are positively correlated, while the other populations show negative correlation or no correlation. All populations have the same volume. The corresponding mathematical example is presented in Section 3.2.2. Note that although the generalized covariance can not detect a rotational change in the covariance, the $T^2$ method would
detect such a rotational change as an out of control event.

Figure 3-3: Two different covariance matrices in 2-D

\[
\begin{align*}
S_1 &= \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \\
S_2 &= \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix} \\
S_3 &= \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}
\end{align*}
\]

3.4 Motivation Behind Second Order Statistical Detection Methods

In this section, we provide a number of multivariate dataset examples that include small mean shift and covariance structure changes. Traditional first order and second order statistical detection methods are applied to these datasets and their Monte Carlo simulation results are presented. We also provide simulation results using a new second order statistical method and compare that result with those obtained from conventional methods. To simplify the simulation, we are using only two variables, so that principal components analysis need not be used in these examples. Alternatively, all the results can be interpreted and extended to datasets of higher dimension, in that we could compress down to two principal components, and interpret our two variables in these examples as the princi-
pal components.

In all the examples, we generate a randomized covariance matrix by first generating a symmetric matrix $\mathbf{M}$ with the off diagonal elements satisfying the covariance definition, namely $\sigma_{ik} = \rho_{ik} \sigma_{ii} / \sigma_{kk}$, where $-1 \leq \rho_{ik} \leq 1$. The diagonal elements of $\mathbf{M}$ are also randomly generated. The square of this matrix $\mathbf{M}$ is then calculated to ensure that all the eigenvalues are positive. Now because $\mathbf{M}^2$ is a positive semi definite matrix, a unique positive semi definite square root of $\mathbf{M}^2$ can be found. We then use this matrix $\sqrt{\mathbf{M}^2}$ as our base covariance matrix. The reason we call it the base of our covariance matrix is that we will modify the magnitude of its eigenvalues yet maintain all its eigenvectors in the following examples.

### 3.4.1 Example 1

In this example, the data of size 2000 samples is originally generated from an population of zero mean and covariance matrix $\Sigma_0 = \begin{bmatrix} 32.27 & -14.61 \\ -14.61 & 27.08 \end{bmatrix}$, but for the next 1000 samples the covariance structure has changed to $\Sigma_{\text{new}} = \begin{bmatrix} 21.52 & -9.74 \\ -9.74 & 18.05 \end{bmatrix}$. We actually keep the same correlation coefficients between the two variables, but the eigenvalues of the new covariance matrix are two-thirds of their original values. Figure 3-4 provides the $T^2$ value for each sample based on the sample mean and sample covariance matrix characterized from the first 2000 samples. In that plot, no significant irregular pattern in $T^2$ values stands out visually and most of the sample points are within the control limits (control limits will be discussed in Section 4.5 and are omitted from this example).
An innovative second order statistic detection method to be introduced and discussed in detail in Section 3.5 is then applied to this dataset. First the distribution of the test statistic is established from the first 2000 samples shown in Figure 3-5. The dotted points in Figure 3-5 represent the sampling distribution of the test statistic using the first 2000 samples. Then the test statistic from the new data (the next 1000 samples) is computed and plotted (the x in Figure 3-5). When the test statistic computed from the new data is compared to the sampling distribution (shown in Figure 3-5), we can see a significant difference between the test statistic distribution (from the historical data) and the sample test statistic (from the new test data).
3.4.2 Example 2

In this example, we have a dataset where the first 20,000 samples are generated from a given covariance matrix $\Sigma = \begin{bmatrix} 55.25 & -1.57 \\ -1.57 & 18.50 \end{bmatrix}$. Then we use a new covariance matrix $\Sigma_{new} = \begin{bmatrix} 18.50 & 1.57 \\ 1.57 & 55.25 \end{bmatrix}$ which is a rotation of 90-degrees of the original covariance matrix to generate the next 10,000 samples. The generalized variance detection method is then applied to find any change in the process. The generalized variance is computed using the data within a window. In this example, we have chosen a non-overlapping successive window size of 100 samples. By non-overlapping, we mean samples in a given window would not be found in any other windows. Therefore, there are 200 samples of sample generalized variance from the original population (20,000 samples generated by the original covariance matrix) and 100 samples of sample generalized variance from the new population generated from the new covariance matrix (consisting of 10,000 samples). Figure 3-6
shows side by side the 200 samples of sample generalized variance from the original population and the 100 samples of sample generalized variance from the new population. From the plot, there is no significant difference between the first 200 samples and the last 100 samples; as expected the generalized variance can not differentiate this type of change.

**Figure 3-6: Generalized variance for example 2**

However, if we include directional information in our detection method, we should be able to detect the change in this dataset. Figure 3-7 shows values of the test statistic from the new second order detection method. The method also uses 100 sample non-overlapping successive windows, and the test statistic is computed from the data within the window. The detail of this detection method will be discussed in the next section. We see there is a clear change between the values of the first 200 samples and those of the last 100 sam-
The above examples provide illustration of conventional methods failing to catch changes in certain scenarios. Control limits are not provided in the example since the changes can be detected visually.

In real life manufacturing, sometimes we need to detect subtle changes. In fact, in many scenarios, the difference between the two populations can not be resolved using single sample detection methods. That is why we use the sample mean instead of a single sample: since the sample mean has much smaller variance and as the number of samples increases, eventually the two populations can be resolved as long as they are not identical.

Unlike in the univariate case where the distributions are all aligned in the same axis, multivariate distributions can have all sort of orientations. Therefore, one should take advan-
tage of this in multivariate detection and explore the orientation of the covariance; such additional information should help us when we are trying to detect subtle changes. Figure 3-8 shows several possible orientations of two different populations in a 2-D environment. The possible scenarios for multivariate detection presented in the figure are more complex than those in the univariate case.

**Figure 3-8: Possibilities of two different population in 2-D**

3.5 Eigenspace Detection Method

In this section we present a new second order detection method, which we term the “eigenspace detection method” that takes into account the directional change in the population. We provide fundamental properties on the eigenspace distribution and discuss the consistency issues of the method.

As the name suggests, the eigenspace detection method requires eigen-decomposition
of the covariance matrix. Because every sample covariance matrix \( S \) is real and symmetric, there is a real orthogonal matrix \( V \) and a real diagonal matrix \( \Lambda \), such that \( S = V \Lambda V^T \). Furthermore, \( S \) has a spectral decomposition, and one can write \( S \) as

\[
S = \sum_{i=1}^{p} \lambda_i v_i v_i^T = \sum_{i=1}^{p} (\sqrt{\lambda_i} v_i)(\sqrt{\lambda_i} v_i)^T
\]

(Eq. 3-2)

Where \( p \) is the number of variables, \( \lambda_i \) is an eigenvalue and a diagonal element of \( \Lambda \) and \( v_i \) is an eigenvector in \( V \). The above equation resembles very much how the sample covariance matrix is computed. Let each column vector \( X_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T \) represent a \( p \)-variate random vector with density function \( f(X_i) = f(x_1, x_2, \ldots, x_p) \); note the subscript \( i \) is omitted in the distribution because all \( X_i \) have identical distribution. If all column vectors \( X_1, X_2, X_3, \ldots, X_n \) form \( n \) independent identically distributed observations, and we write

\[
X = [X_1, X_2, \ldots, X_i, \ldots, X_n]^T,
\]

then its mean corrected matrix is

\[
X_m = [X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}]^T.
\]

We then can express the sample variance-covariance matrix as the following

\[
S = \frac{X_m^T X_m}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X})
\]

(Eq. 3-3)

The matrix summation from the above equation is very similar to that in Eq. 3-2.

The generalized variance is defined as the determinant of a covariance matrix. Furthermore, from the eigenvalue decomposition, we know that the determinant of a matrix is equal to the product of its eigenvalues. Therefore, the generalized variance only tracks the overall product of the eigenvalues; individual eigenvalues are not tracked in the generalized variance. That is why detection fails for generalized variance when we have permuta-
tion of eigenvalues and their product remains unchanged, even though the covariance
structure has changed.

We now present a second order detection method that not only provides information
on volume change but also identifies when there is an orientation shift in the covariance
structure. Let $\lambda_i$ be arranged in descending order, just like the ordering of singular values,
and let $v_i$ be *almost* the eigenvector associated with $\lambda_i$ except that $v_i$ is selected uniquely.
We will come back to the selection of $v_i$ in a moment. We assume the eigenvalues are not
repeated, i.e. all eigenvalues have multiplicity of one. We introduce a new term: the product
of the square root of the singular value and the eigenvector

$$\sqrt{\lambda_i} v_i, i = 1, 2, \ldots, p$$

is called the *eigenfactor* and the matrix containing all the eigenfactors is the *eigenspace
matrix*. Detection using the eigenspace matrix is termed the eigenspace detection method.

Because this is a second order detection method, a window of samples must be collected
before diagnosis can be performed. We can also rewrite the spectral decomposition as

$$S = \sum_{i=1}^{p} \lambda_i v_i v_i^T = \sum_{i=1}^{p} (\sqrt{\lambda_i} v_i)(\sqrt{\lambda_i} v_i)^T = (V \Lambda^{1/2})(V \Lambda^{1/2})^T = EE^T$$

So instead of tracking all eigenfactors $\sqrt{\lambda_i} v_i, i = 1, 2, \ldots, p$ individually, we can just track
the eigenspace matrix $E$.

We now provide a selection procedure for eigenvectors such that $E$ becomes a unique
decomposition matrix from the sample covariance matrix $S$. From its previous definition
$E = V \Lambda^{1/2}$, there are still two possibilities that we can pick when selecting an eigenvector,
i.e. if $v$ is an eigenvector of $S$ then $-v$ is also an eigenvector of $S$. Note that for any square
matrix any scalar multiple of an eigenvector is also an eigenvector. Because of this ambi-
guity associated with the eigenvectors, we provide here a procedure so we will get a unique eigenvector for each positive eigenvalue. One way to do so is to find a hyperplane in $\mathbb{R}^p$; then given any eigenvector, we can pick the eigenvector that lies on one side of the hyperplane. The following is a formal definition of how to pick the unique eigenvector

**Definition:** Unique eigenvector: Given an eigenvector, we pick a hyperplane, say $x_1=0$.

The orthonormal vector associated with this hyperplane is

$$n_\perp = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then all the eigenvectors picked must lie on one side of the hyperplane (either $x_1 \geq 0$ or $x_1 \leq 0$). For example, we could choose the eigenvector whose angle with the normal vector is less than $90^\circ$ (this corresponds to the case $x_1 \geq 0$). Figure 3-9 presents graphically the selection of two unique eigenvectors in the two dimensional case; both $V_1$ and $V_2$ are on the right side of the hyperplane ($x_1=0$).
Such a selection of an eigenvector enables us to obtain a unique eigenvector. This eigenvector is called the unique eigenvector and $E$ whose columns consist of unique eigenvectors is called the unique eigenspace matrix. For the following discussion, the term eigenspace matrix is used interchangeably with the term unique eigenspace matrix.

3.5.1 Distribution of Eigenspace Matrix $E$

We now provide some fundamental properties of the distribution of the sample eigenspace matrix $E$. In order to establish the distribution for matrix $E$, we must first study properties of the distribution of the sample covariance matrix $S$. The distribution of $S$ is called the Wishart distribution after its discoverer ([GN00], [Arn81]); it is defined as the sum of independent products of multivariate normal random vectors, as in Eq. 3-3. We shall think of the Wishart distribution as a generalized extension of the $\chi^2$-distribution into
a multivariate scenario.

**Definition:** Let \( Z = [Z_1 \ Z_2 \ ... \ Z_n]^T \) such that the \( Z_i \) are independently distributed as \( N_p(0, \Sigma) \) (this is a \( p \)-variate normal distribution). Let \( W = Z^T Z \). Then \( W \) is said to have a Wishart distribution with \( n \) degrees of freedom (d.o.f.). Note that \( W \) is a \( p \times p \) matrix and positive definite.

From the above definition, we then summarize the sampling distribution results below.

Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from a \( p \)-variate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \). Then the following statement is true:

- \((n-1)S\) where \( S \) is defined in Eq. 3-3, is distributed as a Wishart random matrix with \( n-1 \) d.o.f., i.e. \((n-1)S \sim W_p(n-1, \Sigma)\).

More properties associated with the Wishart distribution are now stated. The proof of these properties is either provided or pointed out with references. These properties together with a number of matrix algebra properties will later be used to prove an important theorem, which is then used to derive the distribution for the eigenspace matrix \( E \). In the following theorems and corollaries, the \((n-1)\) term from \((n-1)S\) has been suppressed for simplicity.

**Theorem 3.5.1.** Let \( S \sim W_p(n, \Sigma) \) and let \( A \) be any \( p \times p \) nonsingular matrix. Then \( ASA^T \sim W_p(n, A\Sigma A^T) \).

**Proof:** Found in [GN00].

**Corollary 3.5.2.** Let \( S \sim W_p(n, \Sigma) \) and \( \Sigma^{-1} = A^T A \). Then \( ASA^T \sim W_p(n, I_p) \).

**Proof:** From \( \Sigma^{-1} = A^T A \), we have \( \Sigma = A^{-1} A^T \). Note \( A \) is nonsingular so its inverse exits.
Then from Theorem 3.5.1, we have

$$ASA^T \sim W_p(n, A\Sigma A^T) \rightarrow ASA^T \sim W_p(n, AA^{-1}A^{-T}A^T) \rightarrow ASA^T \sim W_p(n, I)$$

The above theorem and corollary provide transformation properties related to the Wishart distribution. In particular, for any positive definite matrix $S$ with a Wishart distribution, we can always transform $S$ into a Wishart distribution with identity matrix as the key parameter through some matrix $A$.

The following decomposition theorem of the Wishart distribution is known as Bartlett’s decomposition and is quite important in multivariate analysis.

**Theorem 3.5.3.** Let $S \sim W_p(n, I)$ and $S=TT^T$ where $T$ is a lower triangular matrix with $t_{ii}>0$. Then $t_{ij}$, $1 \leq j \leq i \leq p$ are independently distributed, with $t_{ii}^2 \sim \chi^2_{n-i+1}$, $1 \leq i \leq p$ and $t_{ij} \sim N(0,1)$, $1 \leq j < i \leq p$.

**Proof:** Found in [GN00].

Note that the restriction of $t_{ii}>0$ in this theorem turns out to remove the ambiguity of eigenvectors of a positive definite matrix $S$, since the distribution of the eigenspace matrix $E$ depends on $T$.

The following matrix theory of a positive definite matrix sheds some light on relationships between different square root decompositions of a given positive definite matrix $A$.

**Theorem 3.5.4.** Suppose $A$ is a positive definite matrix, where $A=C_1C_1^T$ and $A=C_2C_2^T$. Then $C_2=C_1V$, where $V$ is a unitary matrix. In fact, we can show that any solution $C$ to $A=CC^T$ is of the form $C=A^{1/2}V$ with $V$ unitary.

**Proof:** In order to be a positive definite matrix $A$ must be an Hermitian matrix by defini-
tion (if $A$ is real, then Hermitian means symmetric). We first prove the second statement.

The second statement can be rephrased as $A = CC^T \iff C = A^{1/2}V$. Let $C=A^{1/2}V$ with $V$ unitary, we then have

$$CC^T = (A^{1/2}V)(A^{1/2}V)^T = A^{1/2}VV^TA^{1/2} = A \quad \text{(Eq. 3-6)}$$

Hence, we have shown $C = A^{1/2}V \rightarrow A = CC^T$. To prove the other direction, we need to show that if $A=CC^T$, then $V$ is unitary. We can write $V=A^{-1/2}C$, then

$$VV^T = A^{-1/2}CC^TA^{-1/2} = A^{-1/2}AA^{-1/2}. \quad \text{Because } A \text{ is positive definite, } A \text{ has a unique positive definite square root matrix and } A = A^{1/2}A^{1/2} \text{ (see [HJ85])}. \text{ Therefore, } VV^T = A^{-1/2}AA^{-1/2} = I, \text{ and } V \text{ is indeed unitary.}$$

Next we show the relationship between $C_1$ and $C_2$. From the second statement we know $C_1$ is of the form $C_1=A^{1/2}V_1$ with $V_1$ unitary and similarly $C_2=A^{1/2}V_2$ with $V_2$ unitary. From $C_1=A^{1/2}V_1$, we can find $A^{1/2}=C_1V_1^T$, so

$$C_2 = A^{1/2}V_2 = C_1V_1^TV_2 = C_1V \quad \text{(Eq. 3-7)}$$

and $V$ is a unitary matrix, this follows because if $V_1$ is a unitary matrix then its transpose $V_1^T$ is also a unitary matrix and the product of two unitary matrices is another unitary matrix, i.e. $(V_1V_2)(V_1V_2)^T = V_1V_2V_2^TV_1^T = I$.

From Bartlett’s decomposition theorem and the matrix algebra property, we can say something about the distribution of other decompositions of $S$ with identity matrix as its
key parameter. Let $C$ denote a decomposition matrix of $S$ such that $C = TU$ and $U$ is a given unitary matrix, i.e. $S = CC^T = TT^T$. The distribution of $C$ is related to $T$ by the unitary matrix $U$.

We are now ready to prove the key theorem that establishes the distribution of $E$ as a function of Bartlett’s decomposition matrix $T$. We want to point out that although $S = EE^T$, $E$ is not just any square root matrix of $S$. $E$ is precisely defined in the eigenspace matrix. Combining all the previous theorems and corollary, we can say something about the distribution of $E$ in general when $S$ is a Wishart distribution with $\Sigma$ as the key parameter.

Given a sample covariance matrix $S = EE^T$ with a Wishart distribution $S \sim W_p(n, \Sigma)$, we then can find a square root matrix $A$ of $\Sigma^{-1}$ to transform $S$ to $S_{\text{new}}$ such that $S_{\text{new}}$ has a Wishart distribution with identity matrix (i.e. $S_{\text{new}} = ASA^T \sim W_p(n, I)$). Now we can apply Bartlett’s decomposition theorem on $S_{\text{new}}$ and obtain the distribution of the lower triangular matrix $T$ associated with $S_{\text{new}} = TT^T$. Therefore, we have the following equation relating the eigenspace matrix $E$ to $T$

$$S_{\text{new}} = TT^T = ASA^T = AEE^TA^T = (AE)(AE)^T \quad (\text{Eq. 3-8})$$

From the matrix Theorem 3.5.4, the matrix $T$ is related to $AE$ through a unitary matrix $U$, i.e.

$$AE = TU \Rightarrow E = A^{-1}TU \quad (\text{Eq. 3-9})$$

Note $A$ is nonsingular hence its inverse exists, see Eq. 3.5.1 Finally, the distribution of the eigenspace matrix $E$ is related to the distribution of $T$ linearly from the above equation. Because $U$ is just a unitary matrix, it is bounded. Furthermore, because $A^{-1}$ is just a decomposition (square root) matrix of $\Sigma = A^{-1}A^{-T}$, $A^{-1}$ is also bounded. As a result, the
asymptotic properties in the distribution of $E$ solely depend on those of $T$. Hence, the remainder of the section is focused on deriving key asymptotic properties of $T$, and those properties can be applied to the distribution of $E$. Moreover, the elements of $E (e_{ij})$ are linear combinations of the elements in $T (t_{ij})$.

Because there are many possible choices of $A$ and $U$ in Eq. 3-9, we still do not have a clear understanding regarding the distribution of $E$. We now want to provide a choice of $A$ and $U$ and discuss the advantage of using such distribution as the predicted distribution of the eigenspace matrix $E$. The population covariance matrix $\Sigma$ can be decomposed uniquely to its eigenspace matrix $F$ such that $\Sigma=FF^T$. From $\Sigma=FF^T$, we have $\Sigma^{-1}=(FF^T)^{-1}=(F^{-1})^TF^{-1}$. Therefore, we can pick $A=F^{-1}$, $U=I$, and Eq. 3-9 becomes

$$E = A^{-1}TI = (F^{-1})^T = FT$$  \hspace{1cm} (Eq. 3-10)

With such choice of $A$ and $U$, Eq. 3-10 becomes a very attractive choice to study the distribution of $E$. We want to be able to say that $E$ converges to $F$ asymptotically in probability. This asymptotic result can be proved directly from the definition of $E$, as there is a unique eigenspace decomposition of both $S$ and $\Sigma$ into $E$ and $F$ respectively. However, it is much more interesting to look at the distribution of $E$ through Eq. 3-10; in particular we can show $T$ converges to $I$ asymptotically. We now want to prove with the decomposition matrix $E$ defined in Eq. 3-10, that $S=EE^T$ has a Wishart distribution $S \sim W_p(n, \Sigma)$. We have the following theorem:

**Theorem 3.5.5.** Let $S_{new} \sim W_p(n, I)$, where $S_{new}=TT^T$ and $T$ is a lower triangular matrix. Suppose $E=FT$ where $F$ is the eigenspace matrix of a positive definite matrix $\Sigma$ (i.e.
\( \Sigma = FF^T \), then \( S = EE^T \) has a Wishart distribution \( S \sim W_p(n, \Sigma) \).

**Proof:** We just need to use the transformation theorem (Theorem 3.5.1). If \( E = FT \), then

\[
S = EE^T = (FT)(FT)^T = F(TT^T)F^T = FS_{new}F^T
\]  
(Eq. 3-11)

Thus, \( S \sim W_p(n, FF^T) = W_p(n, \Sigma) \).

The above theorem provides a possible distribution of the eigenspace matrix \( E \); this distribution will be used to compare with the sampling distribution of the eigenspace matrix \( E \) in the next chapter. Although the precise distribution of the eigenspace matrix \( E \) could not be obtained (only up to a functional form is obtained), we can derive the precise distribution of the Cholesky decomposition matrix \( M \) of a sampling covariance matrix \( S \).

The Cholesky decomposition of \( S \) factors \( S \) into a product of a unique lower triangular matrix \( M \) and its transpose, i.e. \( S = MM^T \). Such decomposition matrix \( M \) is called the Cholesky decomposition matrix of \( S \). Moreover, the matrix \( M \) can also be used for detecting covariance structure change and examples will be presented in the next chapter. The following theorem provides the distribution of the Cholesky decomposition matrix \( M \) of a sample covariance matrix \( S \).

**Theorem 3.5.6.** Let a sample covariance matrix \( S \) be decomposed to its Cholesky matrix \( M \) by \( S = MM^T \), and the population covariance matrix \( \Sigma \) also be decomposed to its Cholesky matrix \( (\Sigma = LL^T) \). Then the distribution of a sample Cholesky decomposition matrix \( M \) of \( S \) has the distribution \( M = LT \), where \( T \) is a lower triangular matrix with \( t_{ii} > 0 \). Then the elements \( t_{ij}, 1 \leq j \leq i \leq p \) are independently distributed with \( t_{ii}^2 \sim \chi^2_{n-i+1}, 1 \leq i \leq p \).
and \( t_{ij} \sim \text{N}(0,1), \quad 1 \leq j < i \leq p \).

**Proof:** The proof again requires the transformation Eq. 3.5.1 From Eq. 3-9, \( E \) is replaced by \( M \) and we pick \( A^{-1} = L \Rightarrow A = L^{-1} \). We then have

\[
M = A^{-1} TU = LTU ~~~~ (Eq. 3-12)
\]

Since both \( L \) and \( T \) are lower triangular matrix, their product must be a lower triangular matrix. Use the fact that the Cholesky decomposition matrix \( M \) is a unique lower triangular matrix decomposition of \( S \), any other lower triangular matrix decomposition of \( S \) must be equal to \( M \). Therefore, the unitary matrix \( U \) in Eq. 3-12 must be the identity matrix \( I \).

The distribution of \( M \) is then established through \( M = LT \).

Just like \( E \), each element of \( M \) is a linear of combination of \( T \); and the columns of \( M \) can be expressed as \( M = [M_1 \ M_2 \ldots \ M_p] = [LT_1 \ LT_2 \ldots \ LT_p] \). Hence \( M_i \) can be written as

\[
M_i = LT_i ~~~~ (Eq. 3-13)
\]

Thus the column vector \( M_i \) has useful properties easily derived through matrix algebra. In particular, its mean is \( E\{M_i\} = LE\{T_i\} \), and its covariance matrix is \( L\Sigma_T L^T \) (where \( \Sigma_T \) is the covariance matrix of \( T_i \)). Monte Carlo simulation results of \( M \) will be shown in the next chapter.

**3.5.2 Asymptotic Properties on Distribution of the Eigenspace Matrix \( E \) and the Cholesky Decomposition Matrix \( M \)**

In this section, we turn our attention to the consistency of an estimator. The definition of consistency is provided, consistency in turn is discussed in a limiting property of an estimator [FW87]. Moreover, sufficient conditions on consistency used in actual practice
are provided.

Asymptotic properties on the distribution of \( \mathbf{E} \) are discussed in the context of consistency in this section. In particular, issues related to the consistency of the estimator \( \mathbf{E} \) from its true value are addressed. Let \( \Sigma \) be the population covariance matrix, and \( \mathbf{S} \) be the sample covariance matrix. Then \( \Sigma \) can be decomposed to an unique eigenspace matrix \( \mathbf{F} \) such that \( \Sigma = \mathbf{F} \mathbf{F}^T \). Similarly \( \mathbf{S} \) has the following eigenspace matrix \( \mathbf{E} \) such that \( \mathbf{S} = \mathbf{E} \mathbf{E}^T \).

We want to study the asymptotic properties of the eigenspace matrix \( \mathbf{E} \) given the knowledge that the sample covariance matrix \( \mathbf{S} \) converges in probability to \( \Sigma \). In other words, we want to show the sample eigenspace matrix \( \mathbf{E} \) converges in probability to the population eigenspace matrix \( \mathbf{F} \). This is equivalent to showing that the variances of the elements of \( \mathbf{E} \) go to zero as the sample size goes to infinity, since the eigenspace matrix is unique.

The concept of converging in probability can be generalized in the following definition of consistency.

**Definition:** The statistic \( \hat{\theta} \) is a consistent estimator of the parameter \( \theta \) if and only if for each positive constant \( c \),

\[
\lim_{n \to \infty} Pr(|\hat{\theta} - \theta| \geq c) = 0
\]

or, equivalently, if and only if

\[
\lim_{n \to \infty} Pr(|\hat{\theta} - \theta| < c) = 1
\]

In actual practice, we can often find out if an estimator is consistent by using the following sufficient conditions (though not necessary conditions).
Theorem 3.5.7.

The statistic $\hat{\theta}$ is a consistent estimator of the parameter $\theta$ if

$\hat{\theta}$ is unbiased and

$$\lim_{n \to \infty} \text{var}(\hat{\theta}) = 0$$

With the sufficient conditions of consistency, we want to show that the sample covariance matrix $S$ is a consistent estimator of $\Sigma$. Since each element of $S=(s_{ij})$ is an unbiased estimator of each element of $\Sigma=(\sigma_{ij})$ (see [Pap84]), it remains to be shown that the variance of each element of the sample covariance matrix $S$ goes to zero as $n$ gets large. In order to do so, we need the following theorem regarding the variance of $(n-1)S$ as a Wishart distribution:

Theorem 3.5.8.

Let elements of $W=(w_{ij}) \sim W_p(n, \Sigma)$, then

$$E\{w_{ij}\} = n\sigma_{ij},$$

$$\text{Cov}(w_{ij}, w_{kl}) = n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$$

Proof: The proof can be found in [GN00].

Now we use this theorem to show that the variance of each element of the sample covariance matrix $S$ goes to zero as $n$ gets large.

Theorem 3.5.9.

Let elements of $(n-1)S=(n-1)(s_{ij}) \sim W_p(n-1, \Sigma)$, then

$$\lim_{n \to \infty} \text{var}(s_{ij}) = 0.$$
Proof: From Theorem 3.5.8, we know that

\[ \text{var}((n-1)s_{ij}) = (n-1)(\sigma_{ii}^2 + \sigma_{ij}^2) \]

Therefore, as \( n \) goes to infinity, we have

\[ \lim_{n \to \infty} \text{var}(s_{ij}) = 0 \]

In particular, for the diagonal elements of \( S \) we have \( \text{var}(s_{ii}) = \frac{(\sigma_{ii}^2 + \sigma_{ij}^2)}{n-1} = \frac{2\sigma_{ii}^2}{n-1} \) and we often express \( \sigma_{ii} = \sigma^2 \). So this collapses to what we know in the univariate scenario: that \( s_{ii} \) is a consistent estimator of \( \sigma^2 \) and

\[ \text{var}(s_{ii}) = \frac{2\sigma_{ii}^2}{n-1} = \frac{2\sigma^4}{n-1} \]

Therefore, the sample covariance matrix \( S \) is a consistent estimator of \( \Sigma \). Our goal is to show that the sample eigenspace matrix \( E \) is an consistent estimator of \( F \). To do so we use the above theorem, and it is thus sufficient to show that the variances of the elements of \( E \) go to zero as \( n \) goes to infinity. Because \( E \) is a linear combination of the elements in the lower triangular matrix, it is sufficient to show that the variance of \( T \) goes to zero as the number of samples goes to infinity.

**Theorem 3.5.10.** Let \((n-1)S \sim Wp(n-1,1p)\) and \((n-1)S=(\sqrt{n-1}T)(\sqrt{n-1}T)^T\), where \(T=(t_{ij})\) is a lower triangular matrix with all its elements being independently distributed, \( t_{ii} > 0, \)
\[(n-1)t_{ii}^2 - \chi^2_{n-i+1}, 1 \leq i \leq p \quad \text{and} \quad \sqrt{n-1}t_{ij} - N(0, 1), 1 \leq j < i \leq p \]. Then, \(\text{var}(t_{ij})\) goes to zero as \(n\) goes to infinity for all \(i\) and \(j\).

Proof: For the off-diagonal elements \(t_{ij}\) of \(T\) when \(i \neq j\), we know \(\sqrt{n-1}t_{ij}\) has the standard normal distribution with variance 1. Therefore, the variance of \(t_{ij}\) can be computed:

\[
\text{var}(\sqrt{n-1}t_{ij}) = (n-1)\text{var}(t_{ij}) = 1 \Rightarrow \text{var}(t_{ij}) = \frac{1}{n-1}
\]

The limit of the variance tends to zero as \(n \to \infty\), i.e. \(\lim_{n \to \infty} \text{var}(t_{ij}) = \lim_{n \to \infty} \frac{1}{n-1} = 0\). As for the diagonal elements \(t_{ii}\) of \(T\), we do not have the distribution of \(t_{ii}\); however, we do know that 

\[(n-1)t_{ii}^2\]

has a chi-square distribution with degrees of freedom \(n - i + 1\). Consequently, we can derive the distribution of \(t_{ii}\) from \(t_{ii}^2\). We first show that the variance of \(t_{ii}^2\) goes to zero as \(n\) gets large. Again, using the fact that the variance of \(\chi^2_\nu\) is \(2\nu\), we then have

\[
\text{var}((n-1)t_{ii}^2) = (n-1)^2\text{var}(t_{ii}^2) = 2(n-i+1) \Rightarrow \text{var}(t_{ii}^2) = \frac{2(n-i+1)}{(n-1)^2}
\]

This limit also goes to zero as \(n\) gets large, \(\lim_{n \to \infty} \text{var}(t_{ii}^2) = \lim_{n \to \infty} \frac{2(n-i+1)}{(n-1)^2} = 0, 1 \leq i \leq p\). We are now ready to show that the variance of \(t_{ii}\) goes to zero as \(n\) gets large using an asymptotic formula for the gamma function; a formal proof follows later in this section.

We take the following approach to find the variance of \(t_{ii}\) from its density function, hence we need to find the density function of \(t_{ii}\) from the density function of \((n-1)t_{ii}^2\). Since \(t_{ii}\) is a function of \((n-1)t_{ii}^2\), whose distribution is known, we can use the transforma-
tion or change of variable technique (see [WM93]) to compute the distribution of $t_{ii}$. We break up the transformation into two steps. First, we compute the distribution of $t_{ii}^2$. From the transformation technique, let $x = (n - 1)t_{ii}^2$ and $y = \frac{t_{ii}^2}{n - 1} = f(x)$. Because this transformation is a one-to-one mapping, we can find a single value $y$ such that $x = f^{-1}(y)$. The Jacobian is then calculated and the following formula gives us the distribution of $y$.

$$
Prob(y) = Prob(f^{-1}(y))|J|
$$

In our case, $J = \frac{dy}{dx} = \frac{1}{dy} = n - 1$. So the density function of $y$ is

$$
Prob(y) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} (n - 1)^{\frac{\nu}{2} - \frac{2}{\nu}} e^{-\frac{(n - 1)y}{2}}, y > 0
$$

(Eq. 3-14)

We now make the second transformation to find the density function of $t_{ii}$. Let $z = t_{ii} = \sqrt{\frac{2}{t_{ii}}} = \sqrt{y}$, where $z > 0$. We again follow the above procedure using $y = z^2$ and compute the Jacobian $J = \frac{dy}{dz} = 2z$, to find the following:

$$
Prob(z) = \frac{2}{2^{\nu/2}\Gamma(\frac{\nu}{2})} (n - 1)^{\frac{\nu}{2} - 1} e^{-\frac{(n - 1)z^2}{2}}, z > 0
$$

(Eq. 3-15)

With this density function we can now compute the mean and variance of the random variable $z$ from the definition. Because of the density function, the first and second moments of $z$ depend on the following integral form

$$
\int_{0}^{\infty} z^k e^{-\frac{(n - 1)z^2}{2}} dz
$$

Such an integral can be transformed into a gamma function and there are several useful properties associated with the gamma function, in particular the recurrence relation, we
can take advantage of. If we let \( w = \frac{(n-1)z^2}{2} \), we then have \( z = \left( \frac{2w}{n-1} \right)^{\frac{1}{2}} \) and \( dw = (n-1)z\,dz \). Hence, the integral can be transformed into

\[
\int_0^\infty z^k e^{-\frac{(n-1)z^2}{2}} \, dz = \int_0^\infty \left( \frac{2w}{n-1} \right)^{\frac{k}{2}} e^{-w} \frac{dw}{(n-1)z} = \frac{2^{\frac{k-1}{2}}}{(n-1)^{(k+1)/2}} \int_0^\infty w^\frac{k-1}{2} e^{-w} \, dw
\]  

(Eq. 3-16)

Where \( \int_0^\infty w^\frac{k-1}{2} e^{-w} \, dw \) is the known gamma function with parameter \( \frac{k+1}{2} \) which is convergent for \( k + \frac{1}{2} > 0 \), i.e. \( \Gamma \left( \frac{k+1}{2} \right) = \int_0^\infty w^\frac{k-1}{2} e^{-w} \, dw \). Finally, the integral can be simplified to

\[
\int_0^\infty z^k e^{-\frac{(n-1)z^2}{2}} \, dz = \frac{2^{\frac{k-1}{2}}}{(n-1)^{(k+1)/2}} \int_0^\infty w^\frac{k-1}{2} e^{-w} \, dw = \frac{2^{\frac{k-1}{2}}}{(n-1)^{(k+1)/2}} \Gamma \left( \frac{k+1}{2} \right)
\]  

(Eq. 3-17)

With this information on hand, we can now calculate the first moment and second moment of random variable \( z \) and use those moments to find the mean and variance of \( z \).

The first moment or the mean of \( z \) is

\[
E\{z\} = \int_0^\infty z \cdot \text{Prob}(z) \, dz = \frac{2^{\frac{\nu-1}{2}}}{\sqrt{\nu} \Gamma \left( \frac{\nu}{2} \right)} \int_0^\infty z^\frac{(n-1)z^2}{2} \, dz
\]

From Eq. 3-17 set \( k = \nu \), we then find the mean of \( z \) to be

\[
E\{z\} = \frac{2^{\frac{\nu-1}{2}}}{\sqrt{\nu} \Gamma \left( \frac{\nu}{2} \right)} \left[ \frac{2^{\frac{\nu}{2}}}{(n-1)^{(\nu+1)/2}} \Gamma \left( \frac{\nu+1}{2} \right) \right]
\]

With some cancellation, the mean of \( z \) simplifies to

\[
E\{z\} = \sqrt{\frac{2}{n-1}} \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right)}
\]  

(Eq. 3-18)
We need to evaluate the limit of this expression as \( n \) goes to infinity. In order to evaluate this limit, we use an asymptotic formula for the gamma function 
\[
\Gamma(n + 1) = n! = \sqrt{2\pi n} n^{-n} e^{-n};
\]
this formula is also known as Stirling’s factorial approximation for \( n! \) when \( n \) is large and an integer. Because \( \frac{v}{2} \) or \( \frac{v + 1}{2} \) might not be integer, we only use this asymptotic formula for illustration, a formal proof will follow. Using this approximation, we can compute the ratio

\[
\lim_{v \to \infty} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} = \lim_{v \to \infty} \frac{\Gamma\left(\frac{v-1}{2} + 1\right)}{\Gamma\left(\frac{v-2}{2} + 1\right)} = \lim_{v \to \infty} \frac{n}{2^n} \frac{2\pi^{v-1/2}}{(v-1)^{(v-1)/2}} \left(\frac{v-2}{2}\right)^{(v-2)/2} e^{-\frac{v-2}{2}} \left(\frac{v}{2}\right)^{(v-1)/2}
\]

for large \( v \). Finally, since \( v = n - i + 1 \), we then have

\[
\lim_{n \to \infty} E\{z\} = \lim_{n \to \infty} \sqrt{\frac{2}{n-1}} \Gamma\left(\frac{v+1}{2}\right) = \lim_{n \to \infty} \sqrt{\frac{2}{n-1}} \Gamma\left(\frac{v}{2}\right) = \lim_{n \to \infty} \frac{n-i+1}{2} = 1 \quad (\text{Eq. 3-20})
\]

In order to find the variance of \( z \), we need to compute its second moment and use the relation that \( var(z) = E\{z^2\} - (E\{z\})^2 \). Again we compute the second moment of \( z \) from the definition,

\[
E\{z^2\} = \int_0^\infty z^2 \text{Prob}(z) dz = \frac{2}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} \int_0^\infty z^{v+1} e^{-\frac{(n-1)z^2}{2}} dz
\]
and from Eq. 3-17 the integral above has \( k = v + 1 \), so

\[
E\{z^2\} = \frac{2}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} (n-1)^{\frac{3}{2}} \int_0^\infty z^{v+1} e^{-\frac{(n-1)z^2}{2}} dz = \frac{2}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} (n-1)^{\frac{3}{2}} \frac{2^{v/2} \Gamma\left(\frac{v}{2} + 1\right)}{(n-1)^{(v+2)/2} \Gamma\left(\frac{v}{2}\right)}
\]

With some cancellation, the above equation can be simplified to

\[
E\{z^2\} = \frac{2}{2^{v/2} \Gamma\left(\frac{v}{2}\right)} (n-1)^{\frac{3}{2}} \frac{2^{v/2} \Gamma\left(\frac{v}{2} + 1\right)}{(n-1)^{(v+2)/2} \Gamma\left(\frac{v}{2}\right)} = \frac{2}{n-1} \frac{\Gamma\left(\frac{v}{2} + 1\right)}{\Gamma\left(\frac{v}{2}\right)}
\]

Now using the recurrence formula for the gamma function \( \Gamma(n+1) = n\Gamma(n) \) to further simplify the above expression, we have

\[
E\{z^2\} = \frac{2}{n-1} \frac{\Gamma\left(\frac{v}{2} + 1\right)}{\Gamma\left(\frac{v}{2}\right)} = \frac{2}{n-1} \frac{\frac{v}{2} \Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} = \frac{v}{n-1} = \frac{n-i+1}{n-1}
\]

(Eq. 3-21)

The limit of the second moment goes

\[
\lim_{n \to \infty} E\{z^2\} = \lim_{n \to \infty} \frac{n-i+1}{n-1} = 1
\]

Finally, we show that the variance of \( z \) goes to zero as \( n \) gets large since

\[
\lim_{n \to \infty} var(z) = \lim_{n \to \infty} E\{z^2\} - \left(\lim_{n \to \infty} E\{z\}\right)^2 = 1 - 1 = 0
\]

(Eq. 3-22)

With all the variances of elements in the matrix \( T \) going to zero, we have shown that the variances of the elements in the eigenspace matrix \( E \) go to zero, because every element in \( E \) is just a linear combination of the elements in matrix \( T \). Let \( (e_{ij}) \) be elements of \( E \);
then the mean and variance of \((e_{ij})\) can be expressed as the following

\[
E(e_{ij}) = E\left(\sum_{k=1}^{p} b_{ik} t_{kj}\right) = \sum_{k=1}^{p} b_{ik} E(t_{kj})
\]

\[
\text{var}(e_{ij}) = \sum_{k=1}^{p} b_{ik}^2 \text{var}(t_{kj})
\]

where \((b_{ij})\) are the linear coefficients. Hence \(\lim_{n \to \infty} \text{var}(e_{ij}) = \sum_{k=1}^{p} b_{ik}^2 \lim_{n \to \infty} \text{var}(t_{kj}) = 0\).

We now conclude this chapter by providing the formal proof of \(\lim_{n \to \infty} \text{var}(t_{ij}) = 0\).

**Theorem 3.5.11.** Given that \(\lim_{n \to \infty} \text{var}(x^2) = 0\). Then for \(y = \sqrt[n]{x^2} = x \geq 0\), we have \(\lim_{n \to \infty} \text{var}(y) = 0\).

**Proof:** Because the variance of \(x^2\) goes to zero, its distribution tends to a delta function at a given point \(a\), in other words with probability one \(x^2 = a\). We prove this statement by contradiction. Assume that as \(n\) gets large \(x^2\) has some non-zero probability in more than one place, then its variance can not be zero from the definition, i.e.

\[
\text{var}(x^2) = \int_{0}^{\infty} (x^2 - E(x^2))^2 f(x^2) dx = (x_1^2 - E(x^2))^2 f(x^2 = x_1^2) + (x_2^2 - E(x^2))^2 f(x^2 = x_2^2) \neq 0
\]

Therefore, the distribution of \(x^2\) tends to a delta function as \(n\) gets large. Now we use the fact that the positive square root \(y = \sqrt[n]{x^2}\) is a continuous, monotone function with one-to-one mapping; thus the distribution of \(y\) must also be a delta function at \(y = \sqrt[n]{a}\), i.e.
\[ \lim_{n \to \infty} \text{Prob}(y = \sqrt{a}) = 1 \]. Hence \[ \lim_{n \to \infty} \text{var}(y) = 0. \]

As a result, knowing that \[ \lim_{n \to \infty} \text{var}(t_{ii}^2) = 0 \], we have

\[ \lim_{n \to \infty} \text{var}(t_{ii}) = 0 \]

Theorem 3.5.11 concludes that the each element of the eigenspace matrix \( E \) converges to something since its variance goes to zero as \( n \) gets large. However, we want to find out exactly what \( E \) converges to; in particular, we wish to determine if \( E \) converges to \( F \) (the population eigenspace matrix). We conclude this chapter by proving that the sample eigenspace matrix \( E \) is a consistent estimator of the population eigenspace matrix \( F \).

**Theorem 3.5.12.** Suppose \( E \) is the sample eigenspace matrix of a sample covariance matrix \( S \), which converges to a population covariance matrix \( \Sigma \). Let \( F \) be the eigenspace matrix of \( \Sigma \). Then \( E \) converges to \( F \).

**Proof:** As \( n \) gets large, we know that \( E \) converges to a matrix \( \hat{E} \) (from Theorem 3.5.11). Let us assume that \( \hat{E} \neq F \) and prove the theorem by contradiction. As \( n \) gets large, we know

\[ S = \hat{E}\hat{E}^T \to \Sigma = FF^T \]

But the eigenspace matrix is unique, as a result \( \hat{E} = F \).
Chapter 4

Monte Carlo Simulation of Eigenspace and Cholesky Detection Strategy

4.1 Introduction

In this chapter, we focus on simulation of the sample eigenspace and Cholesky matrices. The sampling distribution of the Cholesky decomposition of the sample covariance matrix is verified with its precise distribution. We also look at the sampling distribution of the eigenspace matrix and compare that with the distribution $E = FT$, where $F$ is the eigenspace matrix of the population covariance matrix $\Sigma$ and $T$ is the lower triangular matrix from Bartlett’s decomposition theorem (Theorem 3.5.3). Finally, detection strategies based on deviation from the expected distribution of the eigenspace or Cholesky matrices are applied to a few simulated datasets with known changes induced in the datasets, and the results are compared with those of the $T^2$ and the generalized variance multivariate detection strategy.

Both eigenfactors and columns of the Cholesky matrix can be thought of as a set of factors as in factor analysis [JW98]; these factors attempt to explain the covariance relationship among the original variables using only a few factors. The factors in the eigenspace matrix have a geometrical interpretation similar to that in the principal components analysis described in Section 2.4 and Figure 2-4. However, the factors in the Cholesky matrix are oblique and their interpretation remains unclear. The main purpose of factor analysis is to approximate the covariance matrix $\Sigma$ using only a few underlying (but unob-
servable) random quantities called factors. Moreover, in factor analysis most of the focus is on orthogonal factors and it is essential that those orthogonal factors have physical meaning or are rotated until some interpretation can be imposed on the factors. In contrast to factor analysis, our approach seeks to capture the directionality of the covariance matrix; in doing so the covariance matrix is naturally decomposed into the eigenspace and the Cholesky matrix. Moreover, such decompositions maintain full description of the covariance matrix, i.e. no approximation is required. As a result, the distribution of the decompositions can be derived from the distribution of the sample covariance matrix $S$. Finally, because the eigenspace and the Cholesky matrices have such close ties with factor analysis, approximation using only a few columns of the matrices falls out naturally from the theory of factor analysis. Therefore, the eigenspace detection strategy works efficiently when dominant eigenfactors exist in the data, and in some examples provided in this chapter detection using only the first eigenfactor is often sufficient. Figure 4-1 shows the usage of a dominant eigenfactor to detect both variance volume change and eigenvector angle change. The original population is represented by the solid line ellipse, the small dotted ellipse represents eigenvalue changes induced in the process and the larger dotted ellipse represents the eigenvector changes induced in the process.
Figure 4-1: Eigenfactor used to detect variance volume and eigenvector angle change

In Section 4.2, we provides Monte Carlo simulations of the sampling distributions of eigenspace and Cholesky decomposition matrices and compare these to predicted distributions. The goal of this section is to investigate the sampling distributions from oracle data (by oracle data we mean mathematically generated data with specified properties) and verify the sampling distributions from the predicted distributions. Moreover, in this section we want to verify that the sampling distribution gets tighter around the true value when the sample size increases.

The distribution of matrix $T$ has some helpful properties; in particular the elements are independent and the off-diagonal elements are normally distributed. The independence is a very useful property that allows us to find the overall variance in an element $(e_{ij})$ of $E$ from its linear combination of $t_{ij}$. However, there is a drawback in $T$, namely the distribu-
tion of the diagonal elements $t_{ii}$ is complicated. In section 4.3, we investigate the possibility of approximating the distribution of $t_{ii}$ as $n$ gets large. Such an approximation allows us to simplify the distribution of $E$ and $M$, and eventually we want to use the distribution approximation to compute control limits.

Section 4.4 discusses sensitivity issues of the eigenspace detection method, the discussion is especially focused on dominant eigenfactors estimation. Covariance matrix estimates become singular as the number of variables exceeds the sample size. Therefore, only a handful of non-zero eigenvalues can be found from the samples. In this section, we investigate the sampling distribution of eigenspaces as $p$ gets large. More specifically, we study the behavior of the sampling distribution when there are a few dominant eigenfactors. Such situations arise when data collected exhibits high correlation among variables and a few principal components are adequate to describe the process.

Simulated or oracle data is used to test eigenspace detection methods against other multivariate methods, in particular $T^2$ and generalized variance detection methods in section 4.5. Different fault scenarios are created in the oracle data, and emphasis is placed on faults associated with small mean-shift and covariance structure changes.

4.2 Simulated Distributions of Eigenspace Matrix and Cholesky Decomposition Matrix

In this section, we use the distribution of eigenspace and Cholesky decomposition matrices derived from Theorem 3.5.5 and Theorem 3.5.6 and compare those distributions with the sampling distributions obtained from multi-nomial random samples of size $n$ from a finite population size $N$. Note that Theorem 3.5.5 provides a possible distribution for the eigenspace matrix $E$, while Theorem 3.5.6 provides the exact distribution for the
The Cholesky matrix $M$.

We include the Cholesky decomposition matrix in addition to the eigenspace matrix mainly because of some useful mathematical properties associated with the Cholesky matrix. First, for any symmetric and positive definite matrix $S$, $S$ can be uniquely factored into $S = MM^T$, where $M$ is a lower triangular matrix with diagonal elements greater than zero. There is an advantage working with the Cholesky decomposition matrix $M$: since its diagonal elements are always positive, we do not need to check whether or not the eigenvectors computed from the eigenvalue decomposition or singular value decomposition are all on one side of the hyperplane. Hence, we do not have to worry about flipping all the eigenvectors to one side of the hyperplane to achieve uniqueness as we need to do in the eigenspace matrix $E$. The second property is the availability of an exact distribution for $M$ derived in Theorem 3.5.6.

We now describe how the simulation is carried out. Random data of $T$ is generated for the predicted distribution of $E$ ($E=FT$) and $M$ ($M=LT$) based on a sample size $n$, i.e. the distribution of $E$ and $M$ depends on $T$ and the sample size. This is done repeatedly $k$ times, and the predicted distributions of $E$ and $M$ are obtained. Then a population of size $N$ with multivariate normal distribution is generated with mean $\mu$ and covariance matrix $\Sigma$. Now sampling with replacement is done repeatedly in the population $k$ times and each time with $n$ samples; hence the sampling distributions of $E$ and $M$ are obtained. Note that $n$ is much less than $N$, so we do not need to apply a finite population correction factor. We then compare the predicted distribution with the sampling distribution using quantile-to-quantile plots; such plots allow us to compare the key features of the two distributions graphically and are useful for determining whether two samples come from the same distribution. In
other words, we want to confirm that the predicted distributions are an adequate descrip-
tion of the sampling distributions. Therefore, the control limits can be established from the
predicted distributions rather than bootstrap simulations.

This is done for several sample sizes \( n \). Also normal data with the sample size \( k \) and its
quantile-to-quantile plots are provided for benchmark purposes. Note that for a \( p \)-variate
normal distribution, there are \( p \) eigenvectors and each eigenvector has \( p \)-coordinates.

In the following three examples, we are using an bivariate population with \( p=2 \). In all
cases, we choose \( N=20,000 \) because we want to have a large population size compared to
any given sample size provided in the examples.

### 4.2.1 Example 1: \( N=20,000, n=50 \) and \( k=100 \) times

The covariance matrix is

\[
\Sigma = \begin{bmatrix}
21.81 & -12.05 \\
-12.05 & 27.89
\end{bmatrix}
\]

and it can be decomposed to

\[
\Sigma = \begin{bmatrix}
3.75 & 2.78 \\
-4.82 & 2.17
\end{bmatrix} \begin{bmatrix}
3.75 & -4.82 \\
2.78 & 2.17
\end{bmatrix} = FF^T = \begin{bmatrix}
4.67 & 0 \\
-2.58 & 4.61
\end{bmatrix} \begin{bmatrix}
4.67 & -2.58 \\
0 & 4.61
\end{bmatrix} = LL^T \quad \text{(Eq. 4-1)}
\]

In this example, we have two eigenvectors and each eigenvector has two coordinates.
In other words, there are four elements in matrices \( E \) and \( M \) and we will compare the dis-
tribution for each element. Note that because the Cholesky decomposition matrix \( M \) is a
lower triangular matrix, it only has three non-zero elements.

Figure 4-2 shows samples from the predicted distributions of \( E=FT \) having a linear
relationship with samples from the sampling distribution of \( E \); thus, these plots provide
plausibility of using the predicted distribution \( E=FT \) to establish the control limits.
Figure 4-3 also shows a linear relationship between the samples of the predicted distribu-
tion of \( M=LT \) and the samples of the sampling distribution of \( M \). This is expected since
the predicted distribution of \( M \) is the precise distribution of \( M \). The top plot in Figure 4-2
is a normal quantile-quantile plot with the sample size $n=100$ samples to provide sensitivity of the tails as a function of the sample size, i.e. fewer samples tend to have more tail action. Note that in the normal quantile-quantile (Q-Q) plot, even when the data is generated by a Gaussian distribution, we can never see a perfect straight line (see top plot in Figure 4-2).

**Figure 4-2: Normal quantile plot for 100 samples and Q-Q plot for eigenfactor elements**
Moreover, because the slopes of all plots in Figure 4-2 and Figure 4-3 are close to 1, the two distributions (predicted vs. sampling) are almost identical. We also plot the joint distributions of eigenfactor 1 in \( \mathbf{E} \) and the joint distribution of \( \mathbf{E}_1 \) is centered around the first column of \( \mathbf{F} \) (see Eq. 4-1), as shown in Figure 4-4. In this case, the joint sampling distribution of \( \mathbf{E} \) is more spread out than that of the predicted; this can be attributed to small sample size and number of repetitions. The joint distributions of column one of \( \mathbf{M} \) is plotted in Figure 4-5 and the joint distribution is centered around the first column of \( \mathbf{L} \) (see Eq. 4-1). It can be seen in Figure 4-5 that both samples of the predicted and sampling distributions exhibit certain orientation, i.e. both distributions have negative correlation between the elements.
Figure 4-4: Joint Distributions of eigenfactor 1 in E

Figure 4-5: Joint Distributions of column one of M
4.2.2 Example 2: $N=20,000$, $n=500$ and $k=500$ times

The covariance matrix is the same as in example 1, $\Sigma = \begin{bmatrix} 21.81 & -12.05 \\ -12.05 & 27.89 \end{bmatrix}$ and it can be decomposed to

$$\Sigma = \begin{bmatrix} 3.75 & 2.78 \\ -4.82 & 2.17 \end{bmatrix} \begin{bmatrix} 3.75 & -4.82 \\ 2.78 & 2.17 \end{bmatrix} = \mathbf{F} \mathbf{F}^T = \begin{bmatrix} 4.67 & 0 \\ -2.58 & 4.61 \end{bmatrix} \begin{bmatrix} 4.67 & -2.58 \\ 0 & 4.61 \end{bmatrix} = \mathbf{L} \mathbf{L}^T$$ (Eq. 4-2)

In this example, the covariance matrix remains the same, but both the sample size $n$ and the number of sampling times $k$ are increased. In particular, the sample size has increased by a factor of 10 from 50 samples to 500 samples. Thus we expect the sampling distributions to be more tightly centered around the true value (first column of $\mathbf{F}$ and $\mathbf{L}$ Eq. 4-2) and we also expect to see a straighter line in the quantile-quantile plots between the sampling and predicted distributions.
In Figure 4-6, the top plot is the normal Q-Q plot; the line is expected to be a straight line since the data is generated from a Gaussian distribution and the tail action is less noticeable (c.f. example 1) due to the larger sample size. Moreover, as expected for every element of $\mathbf{E}$ (there are four elements) the samples from the sampling distribution align...
very well with the samples from the predicted distribution (see Figure 4-6). In all four plots in Figure 4-6, the range of the sampling distribution (x axis) is very close to that of the predicted distribution (y axis) and the slope is very close to 1, meaning the two distributions are almost identical. The following figure shows the Q-Q plot for the Cholesky matrix $M$.

**Figure 4-7: Q-Q plot for Cholesky Decomposition Matrix $M$**

![Q-Q plot of Cholesky matrix](image)

Each plot in Figure 4-7 represents an element of $M$, and all the plots in Figure 4-7 follow the straight line very well because the number of samples have increased. Therefore, from these plots the samples from the sampling distributions match pretty well with the samples from the actual distribution. We also plot the joint distributions of eigenfactor 1 of the matrix $E$ in Figure 4-8, and because $n=500$ samples, we can see the joint distributions
for the predicted and actual distributions look much more alike than that of example 1 (c.f. Figure 4-4). Moreover, as expected the samples of the joint distributions are more closely clustered around the true value $\begin{bmatrix} 3.75 \\ -4.82 \end{bmatrix}$ (compare with Figure 4-4 of example 1).

**Figure 4-8: Joint distribution of eigenfactor 1**

Joint sampling distribution of eigenspace column 1

Joint predicted distribution of eigenspace column 1
The joint distributions of column 1 in matrix $\mathbf{M}$ are shown in Figure 4-9. The plots in Figure 4-9 show that both samples of the sampling and predicted distributions can be encompassed by the same ellipse oriented in a certain direction. This is because the predicted distribution is the true distribution. Following are the two correlation matrices among the elements of the Cholesky decomposition matrices computed from the samples in the sampling and predicted distributions. For $p=2$, we only have three elements in the Cholesky matrix. The two correlation matrices are very close to each other, especially for

$$
\begin{bmatrix}
1 & -0.377 & -0.014 \\
-0.377 & 1 & 0.059 \\
-0.014 & 0.059 & 1 \\
\end{bmatrix}
$$

and

$$
\begin{bmatrix}
1 & -0.375 & 0.027 \\
-0.375 & 1 & -0.019 \\
0.027 & 0.019 & 1 \\
\end{bmatrix}
$$

the one significant correlation coefficient $-0.37$ (the one coefficient not nearly zero). In
order to justify close, we compute the root mean square residual (RMSR) which is defined as the following

\[ RMSR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=1}^{p} res_{ij}^2}{p(p-1)/2}} \]

where \( res_{ij} \) is the difference between the elements of the two correlation matrices. The RMSR is computed to be \( 6.0934 \times 10^{-2} \), which is small.

The purpose of these two examples is to illustrate that the eigenfactors or Cholesky columns can be bounded by a confidence region which gets tighter as the number of samples increases. The confidence region can be established using the sampling distribution, but it is desirable to establish the confidence region using the predicted distribution. If all the elements of \( T \) have normal distribution, then from \( M = [M_1, M_2, ... M_p] = [LT_1, LT_2, ... LT_p] = LT \) (see Eq. 3-13) we see that each column of \( M \) has a multivariate normal distribution and the \( T^2 \) value of each column has \( \chi^2 \) distribution. It is not easy to establish the control limits for the \( T^2 \) value of each column when the elements of \( T \) are not all normally distributed. Therefore, in Section 4.3 we show that the diagonal elements of \( T \) can be approximated by normal distribution when \( n \) is large. Moreover, we want to point out that there is no correlation between any two columns of \( M \), because \( T_i \) and \( T_j \) are independent of each other when \( i \neq j \).

Our next example focuses on higher dimensions of data, in particular in this example \( p=8 \).
4.2.3 Example 3: N=20,000, n=500 and k=500 times (8-variate)

The covariance matrix is

\[
C = \begin{bmatrix}
-11.11 & -31.90 & 110.08 & -87.07 & 8.03 & 21.60 & 4.00 & 121.83 \\
16.21 & 66.33 & -87.07 & 135.94 & -13.63 & 35.55 & 14.46 & -83.79 \\
0.20 & -4.48 & 8.03 & -13.63 & 13.15 & 0.35 & -11.40 & -10.71 \\
24.05 & 18.55 & 21.60 & 35.55 & 0.35 & 153.07 & 18.65 & 19.81 \\
\end{bmatrix}
\]

This example has eight eigenvectors and eigenvalues and the Cholesky decomposition matrix is also 8 by 8. Therefore, we only select a few elements from the matrices to demonstrate the closeness between the distributions. Figure 4-10 shows the Q-Q plots between the samples of the predicted and sampling distributions of elements of the eigenspace matrix \( E \). Like the previous examples, the normal quantile plot for the same number of samples is provided to serve as a benchmark for the tail.
The straight line relationship between the samples of the sampling and predicted distributions of $E$ shown in Figure 4-10 provides qualitative justification of using $E=FT$ as the distribution of the eigenspace matrix $E$. With $E=FT$, we can write the columns of $E$ as linear combinations of columns of $T$, i.e.

$$E = [E_1, E_2 \ldots E_p] = FT = [FT_1, FT_2 \ldots FT_p]$$  \hspace{1cm} (Eq. 4-3)

Such prediction allows us to find the mean and covariance matrix of each column of $E$ conveniently as the following

$$E\{E_i\} = FE\{T_i\} \quad \text{and} \quad var(E_i) = Fvar(T_i)F^T$$  \hspace{1cm} (Eq. 4-4)

Furthermore, control limits can be established easily for each eigenfactor through the
mean vector and covariance matrix in Eq. 4-4 when $n$ is large.

Figure 4-11: Q-Q plot for Cholesky Decomposition Matrix $M$

Figure 4-11 shows Q-Q plots of the sampling and predicted distributions of some elements in the Cholesky matrix $M$. These Q-Q plots show strong linear relationship between the two samples of the sampling and predicted distribution, just as we have seen in the previous examples.

Now that we have verified the validity of using the predicted distribution of $E$, we can establish the level of the significance or control limits for use in statistical process control. To get ahead of ourselves, we want to find out if there is a way to approximate all these predicted distributions. In particular we want to determine if the Gaussian distribution is appropriate for this approximation. Therefore, we compare the predicted and sampling distribution in this example with a Gaussian distribution. The top four plots in Figure 4-12
provides quantile-to-quantile plots of the samples from the predicted distributions of $E$ and $M$ against a Gaussian distribution. The results show that these distributions are not too far off from a normal distribution. Such a result is useful when approximation is used to establish the control limits; the details are discussed in the next section. Similar results are found for the sampling distributions of $E$ and $M$. The bottom four plots in Figure 4-12 show Q-Q plots of the samples between the samples of the sampling distributions of $E$ and $M$ and a Gaussian distribution. Again, the sampling distributions of $E$ and $M$ are shown to be qualitatively close to a Gaussian distribution.
Figure 4-12: Normal Q-Q plot of predicted and sampling distribution of E and M
4.3 Approximation of Distribution of $E$ and $M$ for Large $n$

The distributions of $E$ and $M$ depend totally on the distribution of $T$, i.e. every element of $E$ and $M$ is a linear combination of the elements in $T$. There are two different types of distributions in the elements of matrix $T$. The off-diagonal elements ($t_{ij}$) ($i \neq j$) are independent random variables having a normal distribution from which we can easily derive any confidence intervals or control limits. However, the distributions of the diagonal elements of $T$ ($t_{ii}$) are not standard known distributions, and there is not a statistical software package providing simulation for such a distribution. Because the distributions of the diagonal elements ($t_{ii}$) are functions of the sample size $n$, we study the plausibility of finding an approximation for the distributions of $t_{ii}$ as $n$ gets large. In other words, we want to approximate the asymptotic distributions of $t_{ii}$ by a normal distribution, and a formal normality test such as the Shapiro-Wilk test (see [SW65] and [SF72]) is used to justify quantitatively this approximation. Such approximation allows us to construct confidence limits on individual eigenfactors or columns of the Cholesky matrix, i.e. the $T^2$ value $T^2 = (E_i - F_i) \Sigma_i^{-1} (E_i - F_i)$ where $F_i$ is the true $i$-th eigenfactor) of the eigenfactor then has a chi-square distribution and the confidence limits of the $T^2$ value can be easily constructed from the chi-square distribution.

In this section, we run several simulations of different sample size $n$ on the distributions of $t_{ii}$. The properties of the simulations are observed and discussed. Moreover, we compare the properties observed from the simulation to those we derived in the previous chapter; such properties include the fact that the mean of $t_{ii}$ tends to 1 and the variance tends to zero as $n$ gets large (see Eq. 3-20 and Eq. 3-22).
4.3.1 Example 1: $p=2, n=3, k=100$ times

In this case, we have a small sample size simulated 100 times. Since the sample size is small, the quantile to quantile plot of $t_{ii}$ against a Gaussian distribution is expected to deviate from a straight line. This is expected since $t_{ii}^2$ has a chi-square distribution with $\nu = n - i + 1$ (d.o.f.) and a chi-square distribution has a heavy tail for small $\nu$; the heavy tail can be seen in Figure 4-13 for both $t_{11}$ and $t_{22}$. The W-test from the Shapiro-Wilk test for normality (see [SW65] and [SF72]) has small p-value (less than 0.01) for both $t_{11}$ and $t_{22}$, and the small p-value suggests that we reject the null hypothesis which assumes $t_{ii}$ has a normal distribution. As a result for small $n$, we can not approximate the distribution of $t_{ii}$ using a normal distribution.

**Figure 4-13: Normal Quantile plot for $t_{ii}$**

![Normal Quantile plot of $t_{11}$](image1)

![Normal Quantile plot of $t_{22}$](image2)
4.3.2 Example 2: \( p=2, n=100, k=100 \) times

In this example, the sample size is increased to 100. The Q-Q plots (Figure 4-14) between the samples of \( t_{ii} \) and a Gaussian distribution show close to a straight line. Such behavior is directly the result of a chi-square distribution as \( \nu \) gets large, i.e. the heavy tail becomes less noticeable when \( \nu \) (d.o.f.) is large. The argument follows from the Central Limit Theorem; a chi-square distribution \( t_{ii}^2 \) of d.o.f \( \nu \) can be thought as a summation of \( \nu \) independent random variables and when \( \nu \) is large the Central Limit Theorem can be invoked and \( t_{ii}^2 \) approaches a normal distribution. Moreover, the W-test from the Shapiro-Wilk test for normality has p-value greater than 0.1 for both \( t_{11} \) and \( t_{22} \), and the large p-value suggests that we do not reject the null hypothesis which assumes \( t_{ii} \) has a normal distribution. Therefore, we can approximate the distribution of \( t_{ii} \) using a normal distribution for large \( n \).
4.3.3 Example 3: p=10, n=500, k=500 times

In this example, we have a 10-variate example and the sample size is now 500, with k=500 sample sets drawn. Moreover, because $p=10$, there are ten diagonal elements in $T$. We observe that almost all the graphs ($t_{1,1}$ through $t_{10,10}$) in Figure 4-15 show straight line relationship against the normal distribution, and this is a result of the chi-square distribution for large $\nu$ as discussed in example 2. We have also performed the Shapiro-Wilk normality test on $t_{ii}$, and found the p-value for all $t_{ii}$ to be larger than 0.1 meaning that the null hypothesis in not rejected (the null hypothesis states that the test distribution comes from a normal distribution). Thus, this suggests that the distribution of $t_{ii}$ could be approximated by a normal distribution when $n$ is large.
Figure 4-15: Normal Quantile Plot of $t_{ii}$ (high dimension)
We also present here some observed properties of the \( t_{ii} \) distribution from the simulations. Table 4-1 and Table 4-2 summarize findings from the simulations of all three examples. As expected, the means of all \( t_{ii} \) tend to 1 as \( n \) gets large and the variances tend to zero for large \( n \). These trends match the properties of \( t_{ii} \) we derived in Chapter 3. Thus, \( T \) becomes the identity matrix \( I \) as \( n \) gets large. Moreover, the expected value of \( T_i \) in Eq. 4-4 becomes

\[
E\{T_i\} = \begin{bmatrix}
0 \\
\vdots \\
0 \\
1 \\
0 \\
\vdots \\
0
\end{bmatrix}\rightarrow \text{ith element}
\]

and the expected value of the eigenfactor becomes

\[
E\{E_i\} = FE\{T_i\} = F_i \quad \text{as} \quad n \to \infty \quad \text{(Eq. 4-5)}
\]

**Table 4-1: Properties of \( t_{ii} \) from simulated distribution**

<table>
<thead>
<tr>
<th>Examples</th>
<th>Mean ( t_{1,1} )</th>
<th>Mean ( t_{2,2} )</th>
<th>Variance ( t_{1,1} )</th>
<th>Variance ( t_{2,2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1 (n=3)</td>
<td>0.8762</td>
<td>0.5735</td>
<td>0.3135</td>
<td>0.1519</td>
</tr>
<tr>
<td>Example 2 (n=100)</td>
<td>0.9971</td>
<td>0.9980</td>
<td>0.0043</td>
<td>0.0053</td>
</tr>
</tbody>
</table>

**Table 4-2: Mean and variance from example 3 (p=10 and n=500)**

<table>
<thead>
<tr>
<th></th>
<th>( t_{1,1} )</th>
<th>( t_{2,2} )</th>
<th>( t_{3,3} )</th>
<th>( t_{4,4} )</th>
<th>( t_{5,5} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.0009</td>
<td>0.9977</td>
<td>0.9985</td>
<td>0.9958</td>
<td>0.9972</td>
</tr>
<tr>
<td>Variance</td>
<td>0.0010</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( t_{6,6} )</th>
<th>( t_{7,7} )</th>
<th>( t_{8,8} )</th>
<th>( t_{9,9} )</th>
<th>( t_{10,10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.9944</td>
<td>0.9925</td>
<td>0.9923</td>
<td>0.9919</td>
<td>0.9980</td>
</tr>
<tr>
<td>Variance</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

In this section we have shown that the distribution of \( t_{ii} \) can indeed be approximated
using normal distribution as \( n \) gets large. Hence, the distribution of all elements of \( \mathbf{E} \) and \( \mathbf{M} \) can be approximated by a normal distribution for large \( n \). Confidence intervals and control limits can then be established using a multivariate normal distribution.

### 4.4 Sensitivity Issues: Estimation of Dominant Eigenfactors for Large \( p \)

In this section, we provide simulation results on sensitivity issues related to estimation of the eigenspace matrix \( \mathbf{E} \) for large \( p \). In particular, the examples are set up in such a way that there are a few (one or two) dominant eigenfactors. For a fixed window size, we want to investigate the sampling distribution of the dominant eigenfactors as a function of \( p \) (number of variables). This topic is especially important in data rich environment when we are dealing with thousands of variables, but the window size might only be a fraction of that in order to achieve early detection of abnormality in the process. However, with a small window size, the samples within the window do not contain enough information to compute all of the eigenfactors (note \( p \) is larger than \( n \)). Therefore, it is desirable to study the effect of \( p \) on key dominant eigenfactors.

Before we discuss the simulation aspect of the issues, we want to give some qualitative discussion of the distribution of \( \mathbf{E} \). The distribution of \( \mathbf{E} \) is closely related to that of \( \mathbf{T} \). Moreover, because only the diagonal elements \( (t_{ii}) \) are dependent on \( p \), we study what happens to the variance of \( t_{ii} \) or \( t_{ii}^2 \) as function of \( p \). Since the variance of \( t_{ii} \) depends on Sterling’s factorial formula, we focus our attention on the variance of \( t_{ii}^2 \). Section 3.5.2 shows that the variance of \( t_{ii}^2 \) is exactly \( \text{var}(t_{ii}^2) = \frac{2(n-i+1)}{(n-1)^2} \) for \( 1 \leq i \leq p \); thus for \( 1 \leq i \leq n+1 \) (assuming large \( p>n \)) the variance of \( t_{ii}^2 \) is close to zero for a given large \( n \) (i.e. the \( \text{var}(t_{ii}^2) = \frac{2(n-i+1)}{(n-1)^2} \) becomes negative and ill defined as \( i>n+1 \)). As a result, if the dominant eigenfactor only depends on the distribution of \( t_{ii}^2 \) for \( 1 \leq i \leq n+1 \), then that distribu-
tion is well defined and the variance still goes to zero even when \( n < p \). This information is particularly useful when we use \( \text{FT} \) to predict the distribution of \( \mathbf{E} \), because the first \( k \) dominant eigenfactors of \( \mathbf{E} \) only depend on the first \( k \) columns of \( \mathbf{T} \) (see Eq. 4-3 in Section 4.2) and the first \( k \) columns of \( \mathbf{T} \) only have the diagonal elements \( t_{ii} \) for \( 1 \leq i \leq k \) (i.e. up to \( t_{kk} \)). As a result, it is conceivable that the distribution of the dominant eigenfactors remain well defined.

In order to see the effect of \( p \) on the distribution of the dominant eigenfactor, we observe a distance measure between the sample mean eigenfactor of the sampling distribution and the true eigenfactor (i.e. \( d = \mathbf{E}_i - \mathbf{F}_i \)). We also observe the variance or standard deviation of the sampling distribution of the eigenfactor as functions of \( p \). We now summarize the results in the following tables.

### 4.4.1 Example 1:

First PC captures around 66% and second PC captures about 22% of the total variation and the window size is 300 samples. The sampling distribution is generated 500 times with replacement (total population 20,000 samples).

<table>
<thead>
<tr>
<th>number of variables (eigenvalue)</th>
<th>2 norm of ( d = \mathbf{E}_i - \mathbf{F}_i )</th>
<th>Normalized 2 norm</th>
<th>Normalized Total Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (662.3296)</td>
<td>0.3524</td>
<td>0.0137</td>
<td>7.1553e-2</td>
</tr>
<tr>
<td>100 (2.1959e4)</td>
<td>1.5233</td>
<td>0.0103</td>
<td>6.9661e-2</td>
</tr>
<tr>
<td>300 (1.2456e5)</td>
<td>2.0465</td>
<td>0.0058</td>
<td>6.8048e-2</td>
</tr>
<tr>
<td>600 (3.2093e5)</td>
<td>6.4869</td>
<td>0.0115</td>
<td>6.7711e-2</td>
</tr>
<tr>
<td>1000 (7.2541e5)</td>
<td>8.1687</td>
<td>0.0096</td>
<td>7.0431e-2</td>
</tr>
</tbody>
</table>
The summary from Table 4-3 and Table 4-4 shows that although the sample mean eigenfactor seems to get further from the actual eigenfactor in Euclidean norm, the normalized difference remains more or less the same. That is, as $p$ increases the eigenvalue increases (see first column of Table 4-3 and Table 4-4), thus we normalize the Euclidean norm by the eigenvalue ($\frac{d}{\|F\|}$). Such normalization shows the 2-norm of the difference between the sample mean eigenfactor and true eigenfactor remains unchanged as a function of $p$. In other words, the overall offset between the sampling mean and true eigenfactor seems to get larger, but that is because the length of the eigenfactor is getting larger too; so when the offset is normalized against the length of the eigenfactor, the change remains close to constant (see third column of Table 4-3 and Table 4-4). Furthermore, the results are consistent for the second eigenfactor. To summarize, the increase in the 2-norm can also be attributed to the increase in the eigenvalues. The geometrical interpretation of the normalization is in Figure 4-16.
Figure 4-16: Normalization of $d$ and total standard deviation

In a statistics context, one can not just talk about the mean of a random variable without mentioning its standard deviation. The summary regarding the standard deviation in the fourth column of Table 4-3 and Table 4-4 suggests that the normalized of the total standard deviations remains fairly constant (fluctuate around a mean) as a function of $p$. We use the total variance by adding all variances from each coordinate in the sampling eigenfactor; this quantity provides some notion about the size of the uncertainty in the sampling mean eigenfactor $\mathbf{E}_i$ (see the dotted box in Figure 4-16). We then take the square root of the total variance to get the total standard deviation and normalize that quantity against the length of the true eigenfactor. Moreover, when $p$ gets large, there is no significant change in the normalized total standard deviation. Similar results are also observed in the second eigenfactor (summary in Table 4-4). Figure 4-17 provides visual enhancement of the results just discussed; the values are taken from Table 4-3 of the first eigenfactor.
The top plot in Figure 4-17 is the normalized 2-norm as a function of $p$, which shows that the value fluctuates around a mean value represented by the solid line. The bottom plot in Figure 4-17 shows a similar result for the normalized total standard deviation, i.e. it also fluctuates around the solid line.

**Figure 4-17: Normalized 2-norm and total standard deviation as function of $p$**

Therefore, the variance of the dominant eigenfactors does not seem to grow without bound even when $p$ becomes larger than $n$. These simulation results grant us confidence in our ability to estimate the few dominant eigenfactors in a data rich environment when $p$ exceeds the sample size $n$. However, the above example still uses a large sample size $n=300$, The following example investigates the normalized offset in the means and the total standard deviation with a smaller sample size $n=100$ from a smaller population (5,000). Also we reduce the dominance of the first eigenfactor from 66% to 33% of the
total variance.

4.4.2 Example 2:

First PC captures around 33% and second PC captures about 22% of the total variation and the window size is 100 samples. The sampling distribution is generated 100 times with replacement (total population 5,000 samples).

Table 4-5: Summary of statistics as function of \( p \) for eigenfactor 1

<table>
<thead>
<tr>
<th>number of variables (eigenvalue)</th>
<th>2 norm of ( d = \mathbf{E}_i - \mathbf{F}_i )</th>
<th>Normalized 2 norm</th>
<th>Normalized Total Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (774.4038)</td>
<td>2.6394</td>
<td>9.4847e-2</td>
<td>0.3523</td>
</tr>
<tr>
<td>100 (2.4228e4)</td>
<td>8.6871</td>
<td>5.5811e-2</td>
<td>0.3519</td>
</tr>
<tr>
<td>300 (1.1335e5)</td>
<td>22.4359</td>
<td>6.6640e-2</td>
<td>0.3576</td>
</tr>
<tr>
<td>600 (3.2871e5)</td>
<td>24.4444</td>
<td>4.2636e-2</td>
<td>0.3296</td>
</tr>
<tr>
<td>1000 (7.1841e5)</td>
<td>54.6202</td>
<td>6.4442e-2</td>
<td>0.4029</td>
</tr>
</tbody>
</table>

Table 4-6: Summary of statistics as function of \( p \) for eigenfactor 2

<table>
<thead>
<tr>
<th>number of variables (eigenvalue)</th>
<th>2 norm of ( d = \mathbf{E}_i - \mathbf{F}_i )</th>
<th>Normalized 2 norm</th>
<th>Normalized Total Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (516.2692)</td>
<td>3.2275</td>
<td>0.1420</td>
<td>0.4503</td>
</tr>
<tr>
<td>100 (1.6152e4)</td>
<td>17.3731</td>
<td>0.1367</td>
<td>0.4356</td>
</tr>
<tr>
<td>300 (7.5564e4)</td>
<td>38.0880</td>
<td>0.1386</td>
<td>0.4663</td>
</tr>
<tr>
<td>600 (2.1914e5)</td>
<td>69.4581</td>
<td>0.1484</td>
<td>0.4809</td>
</tr>
<tr>
<td>1000 (4.7894e5)</td>
<td>85.8821</td>
<td>0.1241</td>
<td>0.4880</td>
</tr>
</tbody>
</table>

In this example, the first principal component is only about half as dominant (33%) as that in the first example (66%), and the second PC captures about the same percentage variation out of the total variation as in the first example. We expect the normalized mean offset \( d \) and total standard deviation in this example to be larger than in the first example.
since the eigenfactor 1 is not as dominant as in example 1 and the sample size is also smaller in this example. Such results are clearly demonstrated in Table 4-5 and Table 4-6. For every $p$, the normalized total standard deviation (column 4) has increased compared to that of the previous example. For example, when $p=10$, the normalized total deviation (column 4 in Table 4-5) is 0.3523 as compared to 0.0716 (column 4 in Table 4-3) for eigenfactor 1; when $p=300$, the normalized mean offset has increased from 0.0058 (column 3 in Table 4-3) to 0.0666 (column 3 in Table 4-5). Other than the inherent increase in the mean shift and total variance, we still observe similar results as in example 1. That is, the normalized mean shift does not seem to increase with $p$ and the normalized total standard deviation does not grow out of bound for large $p$. This example reflects what happens in data rich environments where the largest $p$ is often ten times the sample size. These examples provide encouraging results for the applications from data rich environments which and suggest that we can apply eigenspace detection methods since the estimation of a few dominant eigenfactors are justifiable.

The examples in this section discuss the validity of estimating only a few dominant eigenfactors from a dataset. It must be noted that the dataset must be highly correlated in order for dominant eigenfactors to emerge as a result of the high correlation. This is often the case in data rich scenarios. Moreover, the discussion of dominant eigenfactors estimation takes place under the set up that the sample size $n$ is fixed and $p$ is varied, i.e. we investigate the sensitivity of the estimation of the eigenfactors as a function of $p$. However, we want to point out for any $p$ (large or small), we can have a sample size $n$ large enough such that the estimates of the eigenfactors converge to the true eigenfactors (i.e. the variance of the estimates goes to zero). This is discussed in Table 4-1 and Table 4-2 where $p$ is
fixed and \( n \) is varied.

### 4.5 Oracle Data Simulation

Several *oracle* datasets are provided in this section. By *oracle* datasets, we mean synthetic or computer generated datasets. Traditional multivariate statistical methods described in Chapter 2 such as \( T^2 \) and generalized variance are applied to these datasets as well as eigenspace and Choleskey matrix detection methods. We compare the results using these different methods. Furthermore, observations are discussed with regard to the motivation described in Chapter 2.

The first example is a low dimensional dataset with \( p=2 \), for which graphical interpretation can be provided. In this example, data reduction is not applied. Several out of control scenarios are considered; in particular these scenarios are all related to covariance structure change with small mean shift since large mean shift detection is well studied in the literature [Alt84]. Results for higher dimensional datasets can be extended from this example, since data reduction can be applied to the high dimensional dataset and multivariate analysis is then carried out only on the reduction part of the dataset. An example of high dimensional data is discussed later in this section.

The eigenfactor analysis allows us to work on the eigenfactor space instead of the original data space, and individual eigenfactors can be bounded by a confidence region centered around the expected eigenfactor computed from the training data. Therefore, a \( T^2 \) value is computed for the eigenfactor

\[
T^2 = (E_i - E_{i,\text{EXP}})^T S_{E_{i,\text{EXP}}}^{-1} (E_i - E_{i,\text{EXP}}) \tag{Eq. 4-6}
\]

where \( E_{i,\text{EXP}} \) is the expected \( i \)-th eigenfactor and \( S_{E_{i,\text{EXP}}} \) is the covariance matrix of the \( i \)-th
eigenfactor established from the training data, and \( E_i \) is the eigenfactor calculated from the test samples. The geometrical interpretation of the above discussion is shown in Figure 4-18, where the confidence region of the eigenfactor gets smaller as the sample size \( n \) increases (each \( x \) denotes an estimate of the eigenfactor based on \( n \) samples).

**Figure 4-18: Confidence interval for individual eigenfactor**

4.5.1 Example 1:

In this example, the first 1000 samples are generated using \( C_1 = \begin{bmatrix} 21.81 & -12.05 \\ -12.05 & 27.89 \end{bmatrix} \). The next 1000 samples are generated with the same eigenstructure of the previous covariance matrix with the exception that the second singular value is only half of its original value, i.e. the covariance matrix becomes \( C_2 = \begin{bmatrix} 17.94 & -15.06 \\ -15.06 & 25.54 \end{bmatrix} \). The last 1000 samples are generated by swapping the order of the eigenvectors, i.e. the covariance matrix is now \( C_3 = \begin{bmatrix} 21.81 & 12.05 \\ 12.05 & 27.89 \end{bmatrix} \). The following figures present the results using the different detection methods. Note with the exception of the \( T^2 \) method, all second order statistics methods are computed using a non-overlapping successive window of size 100 samples and the second order statistics are computed based on the data within the window. The control limit for
each multivariate method is also provided with $\alpha = 0.01$. Note that the mean is always zero for the three blocks of the data.

**Figure 4-19: The $T^2$ statistic detection**
Both the $T^2$ and generalized variance detection methods build their probability models of the data using the first 1000 samples. In Figure 4-19, the control limit for the $T^2$ detection is established using 99% confidence intervals, or 1% false alarm rate. The $T^2$ statistic detects abnormality in the last 1000 samples, but nothing seems to be out of control in the second block of 1000 samples (one can see that in Figure 4-19, the $T^2$ values are smaller in those samples due to the decrease in the second singular value). Figure 4-20 shows detection results using generalized variance strategy. The control limit for the generalized variance is established using the first two moments of the distribution together with the property that most of the probability distribution of $\text{det}(S)$ is contained in the interval within three standard deviations from its mean (see [Alt84]), that corresponds to $\alpha = 0.0013$. Although the $T^2$ statistic can not detect changes in the second block of 1000
samples, generalized variance finds covariance structure change in those samples. Overall, combining both methods, we are able to detect the two changes induced in the data.

Figure 4-21 and Figure 4-22 show detection results using eigenspace and Cholesky matrix methods respectively. The probability models of the two eigenfactors are characterized using the first 1000 samples when the process operates normally. Then the sample eigenfactors are computed using the successive non-overlapping window of size 100 samples. The sample eigenfactors are compared with characterized eigenfactors through the $T^2$ statistic

$$T^2 = (E_i - E_{ic})^T S_{E_i}^{-1} (E_i - E_{ic})$$  \hspace{1cm} (Eq. 4-7)

where $E_{ic}$ is the characterized $i$-th eigenfactor and $S_{E_i}$ is the covariance matrix of the $i$-th eigenfactor.

**Figure 4-21: Eigenspace detection**
A multivariate normal distribution assumption is applied in Eq. 4-7 because of the large window sample size \((n=100)\). Therefore, the mean and covariance matrix are estimated from the first 1000 samples, and the estimates are used to compute the \(T^2\) statistic across windows of samples. The control limits for both methods are established using a \(\chi^2\) distribution approximation and 99% confidence intervals are selected.

In the top plot of Figure 4-21, eigenfactor 1 detects the large change which occurs in the last block of 1000 samples, this is because the eigenvectors were swapped in that block. The change in the second block of 1000 samples is not detected using eigenfactor 1 because the change is induced on the second singular value (reduced to half of its original value); as a result eigenfactor 1 is unable to see this change. However, if the change is to magnify the magnitude of the second singular value such that it becomes larger than the
first singular value, then eigenfactor 1 would detect such a change. Note that such change is analogous to swapping the eigenvectors. Although eigenfactor 1 can not detect the change which occurs in the second block of the data, eigenfactor 2 detects the change; this is shown in the bottom plot of Figure 4-21. Note that in Figure 4-21 the y-axis scale of the $T^2$ plot of the second eigenfactor has been expanded to emphasize the out of control points in the second block of the 1000 samples. More interestingly, column 1 of the Cholesky matrix alone is enough to detect both changes induced in the data (see Figure 4-22), although such finding could be coincidental.

4.5.2 Example 2:

This example sheds some light on advantages of the new detection methods applied to higher dimensional dataset. We also investigate the need of monitoring the residual space (counterpart of the Q statistic) in the eigenspace detection in high dimensional dataset. The dataset has 50 variables and PCA is first applied to the dataset; only two principal components are kept for analysis. In this example, the first 20,000 samples are generated by a given covariance matrix $C$. We then induce an out of control scenario by swapping the second principal component with a less significant (small eigenvalue) principal component from the original covariance matrix $C$.

Before showing the results from the control charts, we provide some qualitative discussion about what to expect from each detection method. The PCA with $T^2$ detection method can not detect these changes since the projection of the test data remains well within the eigenspace computed from the first 20,000 historical samples. The Q statistic can detect these changes in since one of the eigenvalues in the residual space trained from the historical data is magnified. The generalized variance detection is not capable of
detecting these changes due to the fact that the first two eigenvalues remain unchanged through the whole dataset. The eigenspace detection should detect this change in the second eigenfactor, i.e. the second eigenfactor from the first 20,000 samples is very different from that of the last 10,000 samples. Non-overlapping successive windows of sample size 100 are employed for all second order statistics detection methods examined here.

**Figure 4-23: PCA with $T^2$ and Q plots**
As expected, the \( T^2 \) value shown in Figure 4-23 and generalized variance shown in Figure 4-24 do not detect the changes in the dataset, though the \( T^2 \) values are well within the limits in the last 10,000 samples. The Q statistic plot (see bottom plot in Figure 4-23) detects a change in the dataset by finding many data points out of control in the last 10,000 samples.
Figure 4-25: Eigenspace detection using two columns with 100-sample window

![Eigenspace #1 T2 control chart with 100-sample window](image)

![Eigenspace #2 T2 control chart with 100-sample window](image)

Figure 4-26: Cholesky detection using the first column with 100-sample window

![Cholesky column #1 T2 control chart with 100-sample window](image)
Using the first two principal components, the eigenspace detection detects the changes induced in the dataset. The first eigenfactor $T^2$ plot (top plot in Figure 4-25) shows a few points out of control in the last 100 samples. The $T^2$ plot (bottom plot in Figure 4-25) of the second eigenfactor shows all the points are out of control in the last 100 samples; the $T^2$ plots in Figure 4-25 show that a 100-sample window provides accurate estimation for the eigenfactors. Similarly, the first column of the Cholesky matrix also detects the changes in the data and this is seen from the $T^2$ plot in Figure 4-26, i.e. many samples after the 200th sample have $T^2$ value larger than the control limits.

This example seems to suggest that there is no need to monitor the residual space for eigenspace detection. This is only partially true. We next consider all possible out of control scenarios regarding high dimensional datasets when PCA is used to pre process the data, i.e. $p$ variables are compressed down to a small number of principal components. In order to do so, we first argue that a dataset with high correlation can be broken up into two parts consisting of a signal space and a noise space [Lee00]. The separation between signal and noise space can be done through the eigenvalues; as stated before small eigenvalues (little variance) suggest no information and hence are classified as noise. Three scenarios are considered.

- When an out of control event or change causes the noise components to become the signal. Such a scenario is provided in example 2. PCA with $T^2$ alone is not capable of detecting such a change, the Q statistic must be used in order to detect this change. However, the eigenspace detection approach can detect such a change since the signal space has changed.
- When the direction of a noise component is swapped with that of another noise component. This scenario does not necessarily correspond to out of control behavior. This scenario can only be detected if we keep track of the residual eigenspace.
However, since the noise components easily fluctuate up and down, swapping among the noise components is considered natural behavior in the noise. Such behavior need not to be detected.

- When the variance of some noise components gets *significantly* larger than its original value, but not large enough to make the noise become the signal. Such a scenario can be detected by the Q statistic. If we keep track of the residual eigenspace this change can also be detected using the eigenspace detection approach. However, this scenario perhaps suggests that the model should be updated. Thus, one could just track the largest eigenvalue from the least significant set (the eigenvalues that were left out from the PCA model), and if that eigenvalue gets significantly larger then one should update the model (i.e. keep more PCs in the model).

We conclude this section by pointing out the key difference between the eigenspace detection method and traditional multivariate methods. Almost all the conventional multivariate fault detection techniques rely on a model and subspace computed from the historical data. With the computed model and subspace in hand, test data can then be compared to the historical data and a conclusion reached to whether or not the test data is produced by such model. However, the eigenspace and Cholesky matrix detection techniques compute the model and subspace of the test data (using more than one sample) and compare that with the model and subspace obtained from the historical data. That is, because the eigenspace matrix detection method compares an entire subspace with another subspace, the residual space is not as critical as the Q statistic is to the $T^2$ statistic. There is a trade-off between early detection and robust detection; namely collecting one sample at a time allows quick detection but detection based on several samples should provide more information on process changes.
Chapter 5

Application of Eigenspace Analysis

This chapter focuses on applications of eigenspace detection. In this chapter, we demonstrate the use of the eigenspace analysis methods discussed earlier on two different manufacturing environments: the semiconductor and automotive industries. In both applications, the experiment setups and physical meaning of the datasets collected are described, and the key objectives of the analysis are presented. The motivation for using the eigenspace detection method is discussed. Section 5.1 investigates the application of eigenspace detection to end point detection in a plasma etch process. Section 5.2 describes application of eigenspace analysis for identification of the root cause variables when data is discovered to be out of control.

5.1 Eigenspace Analysis on Optical Emission Spectra (OES)

From the theoretical work in Chapter 3 to Monte Carlo simulation results on eigenspace analysis, we are now ready to demonstrate this technique on an application that requires multivariate analysis. Optical emission spectra (OES) have traditionally been used to detect endpoint in semiconductor plasma etch processes [WT86], [CS96]. We first describe the experiment setup and data collection. Then the motivation for applying eigenspace analysis to OES data is discussed; in particular we discuss why single point detection methods are not adequate in this application. Results from eigenspace analysis follow to conclude this section.
5.1.1 Optical Emission Spectra Experiment Setup

In recent years, multivariate analysis techniques such as PCA, PLS and $T^2$ have been applied in the semiconductor industry (see [LBL95], [Lit96], [LS95] and [SGM92]). Real-time equipment data together with multivariate analysis is used to detect possible faults.

The Ocean Optics SQ2000 optical emission spectrometer uses optical fibers placed on the side-port of a plasma etch reactor with a clear view of the chamber to look across or down onto the wafer. The optical sensor is capable of multiple fibers, shown in Figure 5-1, so spatial resolution can be achieved. However, in the experiment described here, only the horizontal fiber is used for simplicity. Conventionally, narrow bandpass filters have been used to detect endpoint, where only one or two spectral wavelengths are used for endpoint detection.

**Figure 5-1: Optical emission spectroscopy experiment setup**

Spectra from a side view optical port were collected during the etch process consisting
of approximately one thousand spectral channels each sampled every 600 milliseconds [Le97]. An example of the time evolution of the spectral lines is shown in Figure 5-2. The spectra measure the emission intensity of excited atoms and molecules which in turn provide information on relative concentrations of chemical species. The relative concentrations of chemical species is a useful measure of the plasma state since as different layers of materials are etched the chemistry of the plasma changes. For example, as the oxide layer is etched away, less and less oxide remains until the oxide layer is totally etched away and the silicon layer starts to etch; the chemistry of the plasma thus changes when oxide is replaced by silicon as a surface reactant. Two spectral lines exhibiting the above behavior are presented in Figure 5-3. The data can be divided into two regions: main etch and clear or end point region (see dotted lines in Figure 5-3); also there are two sharp drop-offs known as the plasma turn-on and turn-off states. This OES data is shown to have no time serial correlation during the main etch stage [Le97].

**Figure 5-2: Time evolution of spectral lines in an oxide plasma etch process.**
5.1.2 Endpoint Detection

In semiconductor fabrication, plasma etching is used to produce patterns on the silicon wafer by the selective removal of particular regions of thin film layers on the wafer surface. A photoresist mask is typically used to protect desired surface regions from the etchant and this mask is then stripped after the etching has been completed [Sze88]. The goal of the analysis is to find out when to stop etching so that the erosion or over etch of the underlying layer is minimized. Such detection is critical to proper functionality of a device since both underetch and overetch could render the device inoperative (see Figure 5-4).

Multivariate techniques such as $T^2$ statistic and PCA with $T^2$ have been demonstrated to work well with OES data in certain case ([WGG00] and [Le97]). Both of these methods use all spectral channels to improve the signal to noise ratio in the system, where PCA
provides data reduction through correlation among spectral channels. However, the signal to noise ratio decreases significantly when the area of the etched layer is relatively small compared to the protected area. Such a situation is referred as the low open area problem and endpoint detection becomes very challenging under this circumstance. Both the $T^2$ statistic and PCA with $T^2$ are shown to detect endpoint for large open area wafers with success, but these techniques have not performed satisfactorily in the low open area situation.

Figure 5-4: Plasma etch endpoint is reached when the intended etch layer (oxide) is completed removed

5.1.3 Motivation for Application of Eigenspace Analysis to Low Open Area OES

Test spectra were collected using the experimental setup described in the previous section an oxide etch process at Digital Semiconductor. The wafers processed were patterned for contact etch with about 1% open area. Part of the motivation has been stated: none of the previous multivariate analysis techniques have been shown to work well with low open area. Moreover, in this application we are trying to detect a particular event, rather than any out of control or fault data points. An event such as endpoint can exhibit a subtle correlation change rather than a large mean shift; thus $T^2$ techniques might not be appropriate for such an application.
We now provide a quantitative view of why single sample detection methods do not work with low open area data. In order for a single sample detection approach to work effectively, we need the two populations to be far away enough from each other such that the separation is within the power of resolution of the single sample detection method. In other words, a single sample detection method is not capable of separating two close populations. A sample mean of sample size $n$ is sometimes used to resolve two close populations, because the variances of the sample mean decrease as $\frac{1}{\sqrt{n}}$. This behavior makes a shift in mean between two close populations discernible if enough samples are used.

The sample mean and covariance matrix during main etch region (see Figure 5-3) are calculated from the data. The sample mean of endpoint is then computed from the data. With the assumption that the sample mean and covariance matrix are fairly good estimates of the mean and covariance matrices, we then ask the question “How far is the mean of endpoint from the population of main etch?” We can compute such a statistical squared distance using

$$T^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \leq \chi^2_p(\alpha)$$

We find the squared distance to be 673.70 for our test data. This squared distance is less than the 95% confidence interval with degree of freedom equal to 1087, i.e. $\chi^2_{1087}(0.05) = 1198$. Therefore, the two populations cannot be resolved using a single sample detection approach because their means are too close in a statistical sense relative to the underlying variation in the data. The information above together with the need for event detection makes a single sample detection approach inadequate for the low open
area etch application. If a multiple sample detection method is used, we can explore covariance structure change as well as mean shift.

### 5.1.4 Low Open Area OES Endpoint Detection Results

We want to point out that for the low open area OES data, spectra were collected at 5 Hertz with an integration time of 15 milliseconds [Le97]. Since the purpose is to identify endpoint, we want to characterize the endpoint population. This characterization enables us to verify whether the etching process has reached the endpoint. Characterizations of the main etch alone can only provide information about whether the process is still in the main etch state. In addition, if the process is determined not to be in main etch, no additional information can be drawn about whether or not the process is in endpoint or some other fault condition.

The endpoint is characterized using 100 samples prior to the plasma’s turn-off. Furthermore, we use principal components analysis for data reduction, and the first principal component alone captures about 80% of the total variation out of 1087 spectra. We then only monitor the first eigenfactor through a non-overlapping successive window of size 50 samples. Note that this is a strongly dominant eigenfactor because the second eigenfactor only captures about 0.66% of the total variation.

Before discussing control limits on the eigenfactor control chart, we want to provide some qualitative analysis through analytic geometry. From each sample window, we get the first eigenfactor of that window. The Euclidean distance between this eigenfactor $\mathbf{E}_i$ and the endpoint eigenfactor $\mathbf{E}_{ep}$ is defined to be

$$
\|\mathbf{E}_i - \mathbf{E}_{ep}\| = \sqrt{\sum_{i=1}^{1087} (E_{i1} - E_{ep1})^2}
$$

(Eq. 5-1)
where $E_{ep1}$ is the eigenfactor 1 of the endpoint. This distance is computed to provide a measure of closeness. Note that the Euclidean distance does not include any variance or standard deviation term; the variance is later discussed and included in the control limits.

Figure 5-5 and Figure 5-6 represent two typical wafer runs found in the OES data, showing the distance statistic for successive non-overlapping windows.

**Figure 5-5: Euclidean distance of ($E_1 - E_{ep1}$) in run 4**

Because the sensors/fibers start collecting data prior to plasma turn-on (see Figure 5-3), the data points show a sharp drop near the start of the process when the plasma is just turned on. As a result, those points are not included in the analysis. Both Figure 5-5 and Figure 5-6 are scaled in such a way that the data points prior to the plasma turn-on state are eliminated. Both Figure 5-5 and Figure 5-6 show that the Euclidean distance is large at the beginning of the etch, and when the window approaches the end point the Euclidean
norm becomes small. The Euclidean norm diverges when the sampling window leaves the endpoint population. Also note that the sampling size (50) of the window is smaller than the sample size used to characterize the endpoint population (100). That is why we see two small Euclidean norm points in both figures since there are some overlapping samples in these windows with the samples used to characterize the endpoint.

Figure 5-6: Euclidean distance of \((E_1-E_{ep1})\) in run 5

In the above examples, the endpoint population is characterized for each run, hence the ability to detect that endpoint within the same run is not surprising. We next want to investigate the robustness of the eigenfactor method. In this case, the eigenfactor 1 of the sampling window of any run is compared with the eigenfactor 1 characterized by a previous run. In particular, we characterize the endpoint population using the second run from our data and then compute the Euclidean norm of the difference between the sampling eigen-
factor and endpoint eigenfactor on subsequent runs.

In order to establish a benchmark, we first show the Euclidean norm when the characterization is done using data within the run (see Figure 5-7). We then compute the Euclidean norm again, but this time the characterization is done based on the previous run (Figure 5-8).

**Figure 5-7: Euclidean norm using characterization data from within the run (run 3)**
Figure 5-8: Euclidean norm of run 3 using characterization data from run 2

The key features of the graphs (Figure 5-7 and Figure 5-8) remain unchanged; this suggests that the method can be robust across different runs. Specifically, the minimum remains identifiable in Figure 5-8, where the characterization is done based on the previous run (run 2). Further subsequent runs are used to show the robustness of the eigenspace detection method. Figure 5-9 shows the Euclidean norm of the difference of run 4 when the characterization is again done using the second run, and Figure 5-10 plots the Euclidean norm of run 5 using characterization data from run 2. Both Figure 5-9 and Figure 5-10 show similar results as in Figure 5-8, i.e. from run 3 to run 5 the minimum is detectable even when the characterization is done on run 2.
Figure 5-9: Euclidean norm of run 4 using characterization data from run 2

Figure 5-10: Euclidean norm of run 5 using characterization data from run 2
In order to get the most out of the eigenspace method, it is desirable to reduce the window size. Therefore, instead of 50 sample windows, we want to try using 30 sample windows and check if endpoint detection is still plausible. Indeed, this is the case; Figure 5-11 and Figure 5-12 show the Euclidean distance plot for runs 2 and 4. In the figures, the first few samples are ignored due to plasma turn-on state, and the minimum near the endpoint remains visible in both Figure 5-11 and Figure 5-12.

**Figure 5-11: Euclidean distance for n=30 samples run 2**
5.1.5 Control Limit Establishment for the Eigenspace Detection Method

The normal approximation of the distribution of $t_{ii}$ discussed in Section 4.3 can be used to establish control limits for detection based on a selected eigenfactor. When the sample size is large, the control limit can be established based on percentage points of a multivariate normal distribution, because each eigenfactor can be approximated using a multivariate normal distribution. However, due to the fact that the number of variables $p$ is much larger than the number of samples $n$, there is not enough information to estimate the covariance matrix of each eigenfactor. More precisely, a nonsingular sample covariance matrix cannot be found given the number of samples, thus the inverse of the sample covariance matrix is not defined. In order to work around this problem, for a given eigenfactor we only estimate the individual sample variance for each coordinate in the eigenfactor and assume there is no correlation among the coordinates. This assumption is equivalent to
using a cube instead of an ellipse as control limits on our eigenfactor distribution; it is analogous to imposing univariate Bonferroni control limits on multivariate data. Figure 5-13 shows such a scenario in a 2-D environment, where $\sigma_1$ and $\sigma_2$ are standard deviations corresponding to $E_{11}$ and $E_{21}$, and $E_{11}$ and $E_{22}$ are the first and second coordinate of the eigenfactor 1. This approximation is adequate especially when there is no correlation among the variables, i.e. the power of detection using Bonferroni limits is not affected much when such assumptions hold, and the use of Bonferroni confidence intervals allows us to minimize the type I error (see Section 2.2).

**Figure 5-13: Univariate control limits on multivariate normal data, applied to a selected eigenfactor**

Therefore, we compute the confidence interval for each coordinate using

$$Pr(|x_i - \mu_i| < k\sigma_i) = 1 - \alpha$$  \hspace{1cm} (Eq. 5-2)

In order to get the Bonferroni limits, the inverse of the student $t$ distribution with $n-1$ d.o.f.
must be computed (i.e. we are estimating the variance). With $n=50$ samples, $p=1087$ and with $\alpha = 0.05$, we then have $t_{50-1} \left( \frac{0.05}{2 \times 1087} \right) = 4.4721$, i.e. an overall 5% false alarm rate.

These control limits can be applied directly to the Euclidean norm of the differences between the sampling eigenfactor (c.f. $x_i$ in Eq. 5-2) and characterized eigenfactor (c.f. $\mu_i$ in Eq. 5-2). Therefore, for our 1087-coordinate data we have the control limit associated with the Euclidean norm to be

$$\sum_{i=1}^{1087} \left( 4.4721 \sigma_i \right)^2$$

We further simplify the above equation with the average variance obtained similar to the results in Table 4-3 and Table 4-4

$$(1087 \times (4.4721 \bar{\sigma})^2)^{1/2} = 4.4721 \bar{\sigma} \sqrt{1087} \quad \text{(Eq. 5-3)}$$

In the OES data, taking into consideration the size of the first eigenvalue and the percentage of total variation captured by that eigenvalue, we find the average standard deviation to be 0.4862 for $n=50$ samples. Thus the control limit for the Euclidean norm when $n=50$ samples is 100.89. Such control limits could be further expanded if the numbers used to determine the control limit are only estimates. The control chart with the control limits for the OES data with $n=50$ samples is shown in Figure 5-14 for run 5. From the figure we see that there are only two samples within the control limits, thus the rest of the samples can be thought as out of control or lack of similarity points indicating that the process is in a different state than endpoint.
To conclude this section, the OES data is analyzed using an overlapping moving window. The overlapping moving window discards one test sample from the previous window and introduces a single new test sample point into the window. This moving window allows us to compute the test statistic immediately after each new sample is collected rather than waiting for $n$ new test samples as used in the non-overlapping successive window. This strategy enables us to detect abnormality in the process early and quickly. An overlapping sliding window of sample size $n=50$ samples used in run 5 is shown in Figure 5-15 with the control limits. The key features of the graph in Figure 5-15 are very similar to those of Figure 5-14, and the Euclidean norm is a continuous curve as the overlapping window slides from the beginning of the main etch to the endpoint. Also note that the first minimum of the curve in Figure 5-15 is induced by the plasma turn-on state and is
disregarded even though it is within the control limits.

**Figure 5-15: Euclidean norm of run 5 using an overlapping sliding window of 50 samples**

![Euclidean norm of the differences of run 5 using a sliding window of 50 samples](image)

---

**5.2 Eigenspace Analysis on Automotive Body-In-White Data**

The objective in this section is to apply multivariate detection methods to automotive vehicle body data in order to detect any unusual or out of control behavior in the process. It is desired that the multivariate methods be capable of detecting small mean shifts/drifts while exercising greater control over the risk of false alarms. The data analyzed in this section is generated from vision-system measurements from a vehicle left door. The results from the traditional multivariate methods (PCA with $T^2$) are compared to those of the eigenspace analysis.

All principal components analysis in the following subsections is performed on the covariance matrix because all the vision-system measurements are in the same units and
variations across variables are expected to be more or less the same. As for how many principal components one should keep, there are several different criteria in the literature. In this work, we choose to use the scree graph to determine the number of principal components to keep (see [ShS96], [Lee00]). We also point out that the principal components analysis could be performed on the correlation matrix rather than the covariance matrix, and the number of principal components kept in this case might be those whose eigenvalues are larger than 1. This is because when the correlation matrix is used, all variables are normalized to have variance 1; therefore eigenvalues substantially smaller than one are considered less significant.

In multivariate analysis, when data compression schemes such as PCA or PLS are used, methods for determining the cause of out of control points are often quite elaborate. In essence, the out of control samples are detected using the reduced $T^2$ values; we then must find out which projection components (known as scores) contribute the most to the reduced $T^2$. Finally, the dominant variables in those projection components can then be identified using the loading/eigenvector [KM96]. Later in this section, we provide an alternative way using the eigenspace analysis to identify the root cause.

5.2.1 Vision-system Measurements

The in-line data is collected from vision stations measuring several variables. These measurement points are generated by laser cameras on the vision-system frames. The laser camera measures the distance of a point on the rigid frame to the camera, but only the offset from a predefined value is recorded. The data measured from a rigid body is expected to have substantial correlation among variables; this is because the structure of the rigid body forces the variables to be interdependent on each other. As a result, PCA is used to
take advantage of this correlation and reduce the dimensionality. Moreover, because of such a condition, it is reasonable to consider eliminating some of the laser cameras on the vision-system frame. However, the topic of measurement points elimination is not discussed in this work, such discussion can be found in [CWB00]. Figure 5-16 shows an example of possible vision-system measurement points on a vehicle door.

Our goal in this section is to develop different ways (c.f. [MK96]) to map out-of-control signals back into the underlying process variables in order to help diagnose problems. We explore the use of eigenspace analysis to map the reduced dimensional measurement into dimensions directly related to process variables when a production disruption is detected.

Figure 5-16: Possible vision-system measurements on a door
5.2.2 Out of Control Detection Using PCA and $T^2$ Technique on the Left Door Data

The dataset analyzed in this section is a collection of vision-system measurements from the left doors of a certain type of automotive. There are 62 measurement points collected from each left door. In this dataset variables are highly correlated, since four principal components (PCs) capture 94% of the total variation. Therefore, the PCA model is built based on the first four PCs, and the $T^2$ statistic is computed based on these four PCs. From the scree plot shown in Figure 5-17, we see that the first principal component alone captures a large amount of information (about 73% of the total variation). We can examine the loading of the first principal component to see which variables contribute highly to the first principal component (see bottom plot of Figure 5-19).

**Figure 5-17: Scree Plot for the left door data**
The $T^2$ plot in Figure 5-18 shows that there are many out of control points concentrated in the first 400 records. We investigate the 50th record which is an out of control sample by plotting the scores contribution for that records. As seen in Figure 5-19 (top plot), the first score (projection on PC #1) contributes the most to the $T^2$ value of the 50th record. Furthermore, similar results are found in the other out of control records. Knowing that the first score is responsible for the out of control points, we then plot the first loading/eigenvector to see which variables have large coefficients in magnitude. Those variables then become the potential root cause of the out of control behavior. The following figure shows that variable 20 is the source of the problem.
Plotting the standardized variable 20 across records (see Figure 5-20), we find that there are several samples within the first 400 records with the standardized variable 20 valued near seven. Those samples have variable 20 values which are seven standard deviations away from the mean. Moreover, those out of control samples shown in variable 20 match those we find from the T² plot. In Figure 5-20 control limits of ±3 are provided as a visual aide (where σ has been normalized to 1).
5.2.3 Eigenspace Analysis

In this section, the eigenspace analysis is performed on the same dataset as in the previous section. Because the dataset exhibits high correlation among the vision-system measurements (i.e. the first PC alone captures roughly 73% of the total variance), it is sufficient to monitor the first dominant eigenfactor. The first eigenfactor of the population is characterized using samples that are within control. A non-overlapping successive window of size 100 samples is used in the analysis; the first sampling eigenfactor is then computed using the samples within the window. The difference between the first sampling and population eigenfactor is calculated and the Euclidean norm of the difference is determined.

Figure 5-21 plots the Euclidean norm of the difference vector, where it can be seen that three out of the first four samples are quite different from the rest of the samples. The
control limits are established using the Bonferroni approach [JW98]. In this application the size of the window is 100 and $p=62$, hence $t_{100-1}(\frac{0.05}{2 \times 62}) = 3.4559$ with $\alpha = 0.05$. The average variance in each coordinate (see Section 4.4) is found to be 0.0343. Therefore, the control limits are $3.4559 \times \sqrt{0.0343} \times \sqrt{62} = 5.04$.

Figure 5-21: The 2-norm of the difference between the sampling and population eigenfactor 1

From the out of control samples, we investigate the contribution of individual coordinates leading to the large magnitude in the Euclidean norm. By plotting the difference vector across the variables, we find that variable 20 is the main cause of out of control behavior in the three samples. Figure 5-22 shows the individual contribution to the Euclidean norm of the difference for the first sample. Other samples also demonstrate similar results. The conclusion from the eigenspace analysis verifies the results found in PCA.
Moreover, similar results are obtained when a smaller window size is used. Figure 5-23 shows the results for a window size of 50 samples. The top plot in Figure 5-23 shows six out of the first eight samples are above the control limit, and the bottom plot in Figure 5-23 shows the individual contribution to the Euclidean norm of the difference for the first sample.
One advantage of using the eigenspace analysis is its simplicity. Instead of a three step procedure to identify root causes using PCA, one can find the root cause just in two steps. Moreover, using the PCA approach it is conceivable that an out of control sample must be detected in the residual space through the Q statistic. If that is the case, one needs to check both $T^2$ and $Q$ plots for out of control points, and use the two step procedure described above to determine the root cause. However, if the eigenspace analysis is used, one only needs to check the $T^2$ plot of the dominant eigenfactor (or the Euclidean norm of the difference) for out of control points, and an additional plot of contribution of the underlying variables takes us to the root cause.
Chapter 6

Conclusions and Future Work

6.1 Summary

The problem of detecting subtle covariance structure change using information about the eigenvector direction and magnitude is addressed in this thesis. In particular, the focus of this thesis is to develop a second order statistical detection method not only sensitive to rotational changes in the covariance structure, but also capable of detecting any directional change in the covariance structure. Moreover, the detection method provides additional power to resolve two close populations, i.e. when multiple test samples are collected, a second order statistic formed from the test samples can be used to differentiate the two populations. Conventional multivariate methods such as the $T^2$ statistic work well for detecting many out of control conditions, but such a strategy might not be adequate for process control integration purposes. Event or process drift information can be crucial for equipment shut down or fault prevention. In addition, such information supports planning of maintenance schedules that can boost overall throughput of the equipment.

The eigenspace and Cholesky matrix second order statistic detection methods are introduced in the thesis. The eigenspace is defined such that the no ambiguity exists in the eigenvectors of the correlation or covariance matrix, and a selection procedure to choose a unique eigenvector is discussed. The uniqueness in the eigenspace enables us to address the consistency issues associated with the estimator. The approach used here to decompose the sample covariance matrix into the eigenspace and Cholesky matrix allows us to
investigate process change in a smaller set of factors or components. This strategy works well when a few dominant eigenfactors emerge from the data and is much desired in data rich environments.

Monte Carlo simulations from oracle/synthetic data provide better understanding of the capability of the detection methods studied in this thesis. Although the $T^2$ statistic combined with the $Q$ statistic and the generalized variance provide a powerful detection scheme, individually these methods are quite vulnerable against particular induced changes in the process. Consequently the eigenspace analysis becomes an attractive stand alone or complementary tool to those multivariate detection methods.

The eigenspace analysis has a wide range of applications. First, it can be used for conventional detection of out of control conditions. Second, it can be extended to subtle event detection. Finally, it supports identification of the root cause of an out of control point or other event. Application of eigenspace analysis to the semiconductor manufacturing and automotive industry is demonstrated. In semiconductor fabrication, end point detection is desired in the plasma etch process using optical emission spectra; this problem is very challenging when the area exposed for etch is less than about 1.5%. The eigenspace analysis method is demonstrated to achieve this task. In the automotive industry, mapping out of control signals back into the underlying process variables using vision-system measurements is crucial for determination of appropriate corrective or control action. Once more, the eigenspace analysis provides an efficient way to identify the root cause of the out of control samples.

6.2 Directions for Future Research

A second order statistics detection strategy increases detection sensitivity by collecting
many new observations to improve signal-to-noise ratios. However, such increase in sensi-
tivity is achieved at the expense of delaying corrective action. As trade-offs between early
detection and robust detection become more important in the multivariate setting, future
extensions of this thesis may consider this issue.

In the development of the eigenspace and Cholesky matrix variate distribution, the
data is assumed to be Gaussian distributed and to have no time serial correlation. The
Gaussian assumption is upheld when large number of samples are collected so that the
central limit theorem can be invoked. However, when the observations across time are no
longer independent of each other (i.e. the data exhibits time series behavior), multivariate
time series analysis becomes necessary. Although time series analysis is rather well devel-
oped and understood for the univariate case, the situation is not so complete for the multi-
variate case.

The decomposition of the data in this thesis is focused on correlation among variables,
and different decomposition schemes should be explored. In particular, multiway data
analysis discussed in [Bre92] provides several decompositions including multiway PCA,
parallel factor analysis (PARAFAC), Tucker models and the generalized rank annihilation
method (GRAM). These decompositions allow simultaneous modeling of time and vari-
ables, and investigate the data from a three-way array perspective (rather than a matrix)
with the following axes: objects, variables and time. Decompositions of a three-way array
might be possible, nevertheless interpretations in each dimension remain unclear. For
example, multiway PCA adds the temporal information by rearranging a 3-D array into a
2-D matrix (see Figure 6-1). With standard PCA performed on the 2-D matrix, the time
serial correlation is captured as correlation between variables. Moreover, time correlation
captured using multiway PCA is different from the autocorrelation defined in the time series literature [BJR94], since the autocorrelation is discussed with a constant separation of time interval or known as lag \( k \). As a result, interpretation of the time decompositions remains to be investigated.

Finally, covariance structure information can be useful for feedback tuning or control purposes. A single point often does not provide enough information regarding the control strategy; in many instances a control strategy based on a single point could increase the variation of the process. Instead, the covariance information calculated in the eigenspace analysis using the new observations can be used to drive the current covariance structure to a target covariance matrix. Depending on the input controllability of the system, such a control problem can be formulated and optimization can be carried out. The control strat-
egy is not discussed in this thesis, but it could have significant impact on processes. Moreover, the control action becomes especially important for equipment shut down maintenance/prevention, as a non-scheduled shut down can be quite costly. Therefore, research in this area could further be explored.
References


