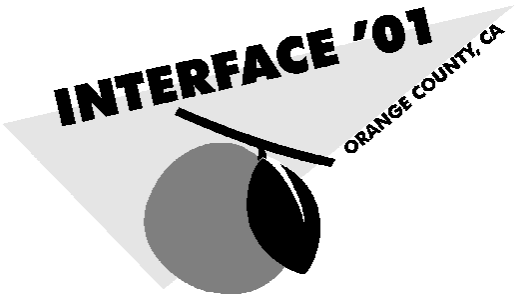
Computing Science and Statistics

Volume 33

Frontiers in Data Mining and Bioinformatics



Editors

Edward J. Wegman
George Mason University

Amy Braverman
NASA Jet Propulsion Laboratory

Arnold F. Goodman
University of California, Irvine

Padhraic Smyth
University of California, Irvine

Proceedings of the 33rd Symposium on the Interface

INTERFACE FOUNDATION OF NORTH AMERICA

About This Volume

The papers, discussions, and related materials in this volume are reproduced exactly as received from the authors. This issue contains the Proceedings of the 33rd Symposium on the Interface and the materials contained in this volume are presumed to be essentially as given at the 33rd Symposium on the Interface. The papers have been reviewed by the editors for appropriateness of content, but have not been formally refereed. The editors or publisher make no claims as to the originality or accuracy of their contents.

This volume is presented on a compact disc, but contains an electronic file (master.pdf) that is the image of the printed volume. For those interested in a hard copy, printing this file on a duplex color printer will result in a paper copy of this volume, which may be bound. Because of this, citations from this volume may be given in a traditional manner. Examples are given below.

> Solka, J. L. and Marchette, D. J. (2001), "Functional analysis of computer network data," *Computing Science and Statistics*, 33, 51-60

If more details are required, the editors and publisher may be added.

> Solka, J. L. and Marchette, D. J. (2001), "Functional analysis of computer network data," *Computing Science and Statistics*, 33, 51-60, (E. Wegman, A. Braverman, A. Goodman, P. Smyth, editors), Interface Foundation of North America, Inc., Fairfax Station, VA

Page numbers are the traditional means of uniquely locating an article within a paper volume. For an electronic volume, the directory structure uniquely locates an article within an electronic volume. Therefore, an alternative citation may be given by citing the appropriate directory location for the article. An example is given below.

> Solka, J. L. and Marchette, D. J. (2001), "Functional analysis of computer network data," *Computing Science and Statistics*, 33, /I2001Proceedings/JSolka/JSolka.pdf.

This volume is published by the Interface Foundation of North America, Inc., 9748 Thorn Bush Drive, Fairfax Station, VA 22039, USA

ISBN 1-886658-07-2

Produced in the U.S.A. (2001)

Multivariate Statistical Process Control and Signature Analysis Using Eigenfactor Detection Methods

Kuang-Han Chen, Duane S. Boning and Roy E. Welsch[1]

With the rapid growth of data-acquisition technology and computing resources, a plethora of process and product 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 paper proposes novel eigenfactor 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 it can reduce the data dimension, much like principal components analysis, yet retains the ability to detect and distinguish between subtle covariance structure changes.

1. Introduction and Motivation

In large and complex manufacturing systems, statistical methods are used to monitor whether the processes remain in control. This paper 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 of key parameters. Many businesses now use univariate statistical process control (USPC) (Montgomery (1996)) in both their manufacturing and service operations. Automated data collection, low-cost computation, products and processes designed to facilitate measurement, and demands for higher quality, lower cost, and increased reliability have accelerated the use of USPC.

However, in many situations the widespread use of USPC has caused a backlash as processes are frequently adjusted or shut down when nothing is really wrong because the probability of false positives (Type I error) is calculated based on USPC and takes little or no account of the multiple tests that are being performed or the correlation structure that may exist in the data. It is very likely that these variables will be correlated due to the large number of variables collected at a given time. 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. The approaches to deal with these issues focus on Bonferroni adjustments, Hotelling's T-squared statistics, and the generalized variance. Furthermore, data reduction strategies such as projection methods (principal component analysis or PCA) (Johnson and Wichern (1998)) are needed to address the high dimensionality problem in data rich environments. Often these methods indicate that some sort of change has taken place, but provide little information about the real

1. Kuang-Han Chen received his Ph. D. from Massachusetts Institute of Technology in 2001 and was associated with Microsystems Technology Laboratory. Duane S. Boning is Associate Professor at the Microsystems Technology Laboratory, Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139. Roy E. Welsch is Professor of Statistics and Management, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139 (E-mail: rwelsch@mit.edu)

nature of that change.

Some techniques described above require more than one new observation before making a decision about a process change. For example, one new observation allows a comparison with a previous mean (and associated control interval), but does not allow for the computation of a new variance to compare with an existing measure of variance. Many new observations will improve signal-to-noise ratios allowing increased detection sensitivity at the expense of delaying corrective measures. Trade-offs between updating and grouping of measurements become even more important in the multivariate setting.

1.1 Motivation

In this paper, we provide a multivariate detection method that is capable of detecting new events and subtle changes in the covariance structure of the process. Some of the changes discussed are natural behavior in the process and do not necessarily represent out-of-control behavior. As a result, most of the conventional multivariate methods are inadequate for such detection. Moreover, information regarding covariance structure in the process can be crucial for feedback, tuning, and control purposes.

This new multivariate detection approach allows us to solve some endpoint detection problems in the semiconductor industry. More specifically, the new detection method is used to solve the detection problem in the low open area situation where conventional multivariate techniques have not performed satisfactorily. Optical emission spectra have traditionally been used to detect endpoint in plasma etch (Wolf and Tauber (1986), Chang and Sze (1996)). Spectra are collected during the etch process and there are about one thousand spectral channels, which are sampled at high frequency.

The goal of this paper 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 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 Organization

We begin in Section 2 with a review of traditional multivariate detection strategies, including those based on χ^2, Hotelling T^2 statistics, principal components analysis, and generalized variance. A new eigenfactor detection approach and a corresponding eigenspace matrix $\boldsymbol{E}$ are proposed in Section 3, with an analysis of the properties and distribution of this matrix. In Section 4, simulations to verify the sampling distribution of $\boldsymbol{E}$ are presented. An application to a multivariate dataset, drawn from semiconductor manufacturing is presented in Section 5. Finally, conclusions are presented in Section 6.

2. Traditional Multivariate Detection Strategies

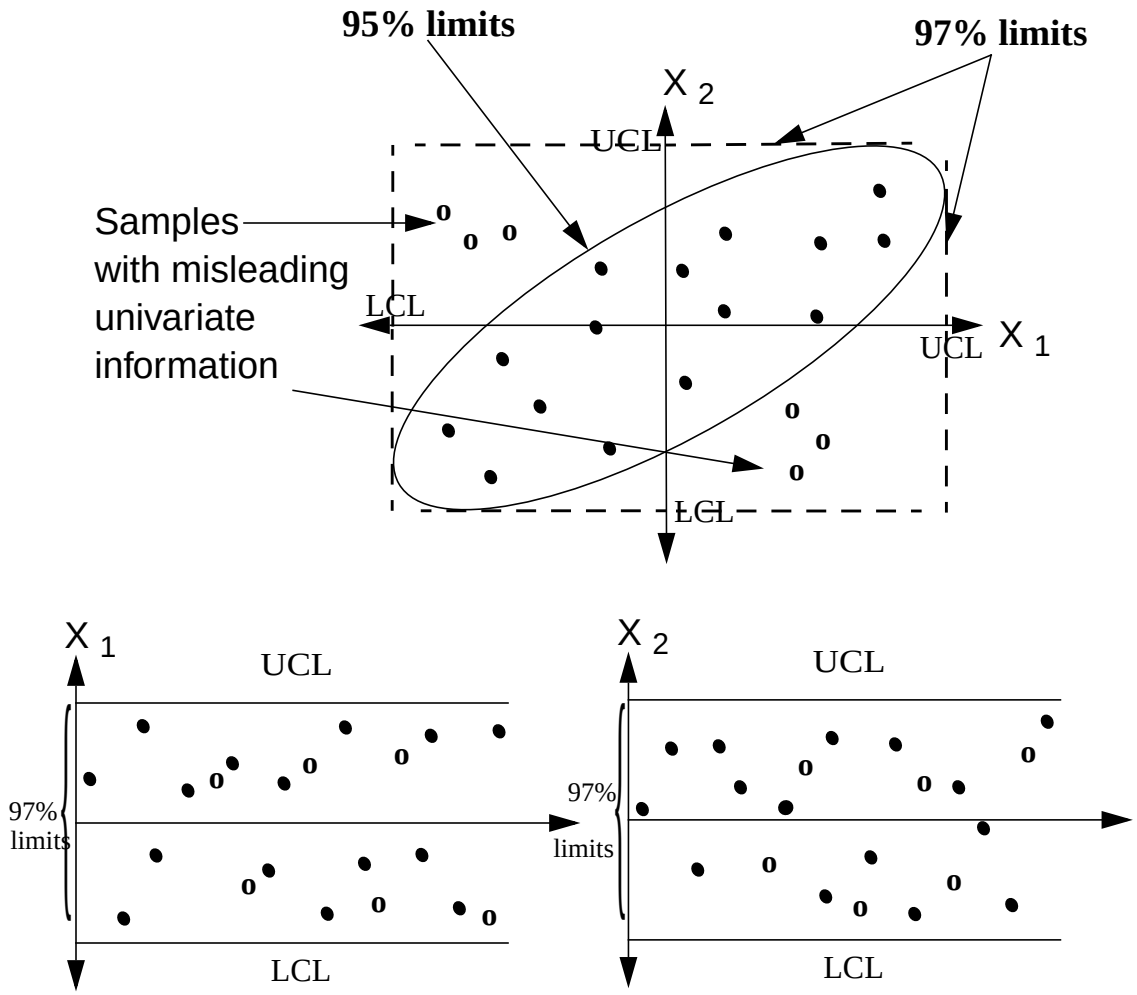
2.1 Multivariate Quality Control: χ^2 and Hotelling's T^2 statistic

Because of rapid sensor advancement and modern manufacturing system 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: distortion of the body results in correlated deviations in these measurements.

By dealing with all of the variables simultaneously, multivariate quality control methods not only can extract information on individual characteristics, but also can identify and monitor the 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 2-1. Here we have two quality variables (x_1 and x_2). Suppose that, when the process 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 2-1. The ellipse represents a contour for the in-control process with 95% confidence limits; both (•) and (**o**) represent observations from the process. The same observations are also plotted in Figure 2-1 as individual Shewhart charts on x_1 and x_2 with their corresponding upper (UCL) and lower (LCL) control limits (roughly 97% 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 (**o**) 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 (**o**) observations are outside the joint confidence region (with the corresponding covariance structure) and are thus different from the normal in-control population of products.

Figure 2-1: Multivariate statistical analysis vs. univariate statistical analysis.



A natural multivariate extension to the univariate Shewhart chart is the Hotelling multivariate control chart. This procedure assumes that *p* quality characteristics are jointly distributed as *p*-variate normal and that random samples of size *n* are collected across time from the process. The Hotelling multivariate control chart signals that a statistically significant shift in the mean has occurred as soon as χ^2 is larger than a threshold limit and χ^2 is defined to be

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

If Σ and μ are unknown, then a T^2 statistic is the appropriate statistic for the Hotelling multivariate control chart. In this case, the sample covariance matrix, ***S***, and sample mean vector,

$\bar{x}$, are used to estimate Σ and μ, respectively.

2.2 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 Johnson and Wichern (1998) and Sharma (1996). In many instances, it is found that the data can be adequately explained using just 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 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. For example, imagine that PCA is performed on some stock market data, one might identify the first principal component as the general market index (average of all companies) and the second principal component might be the market segment component that shows the contrast among different industries. Algebraically, PCA relies on eigenvector decomposition of the covariance or correlation matrix from the variables of interest. An alternate approach to obtain principal components is to use a singular value decomposition on the given data matrix. The mathematical details of PCA can be found in Sharma (1996).

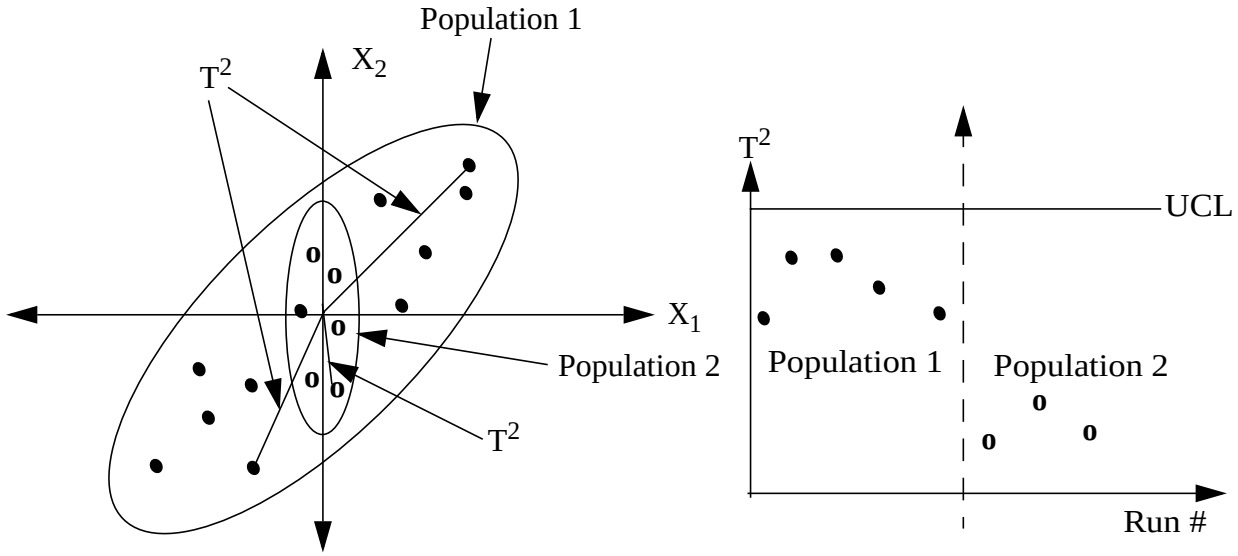
2.3 PCA and T^2 methods

PCA provides great advantages for data compression: instead of dealing with hundreds of variables, we are now dealing with a few principal components. However, the T^2 statistic (MacGregor and Kourti (1995)) only tracks the data in the projection hyperplane; one must also track the Q statistic (Wise, *et. al.*, (1990)) in order to detect if the PCA model no longer describes the process. Note that T^2 is a statistical distance measure, so it cannot resolve the differences in directionality. Moreover, the T^2 statistic cannot be computed when process variables are highly correlated, because the sample covariance matrix $\boldsymbol{S}$ is almost non-invertible.

As a result, a T^2 statistic based on PCA can be used to eliminate the invertability issue of the sample covariance matrix $\boldsymbol{S}$. However, such a strategy is not capable of detecting certain covariance structure changes. An example of this scenario is shown in Figure 2-2. Here we have two populations; one of the populations has more variation in all directions and 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 shifts 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 cannot detect the change depicted in Figure 2-2, 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 paper is to provide single-statistic detection methods that enable both covariance structure change detection and classification.

Figure 2-2: Drawback of T^2 and T^2 with PCA: Reduction in variance not detected



2.4 Generalized Covariance

Generalized covariance methods collapse all the information in a data matrix into a single generalized measure of the covariance of that data (e.g., by taking the determinant of the covariance matrix); it is then easy to monitor this single number. However, generalized covariance is invariant under rotation, i.e., multiplication of the covariance matrix by a rotational matrix $\boldsymbol{U}$ whose determinant is 1. We then have

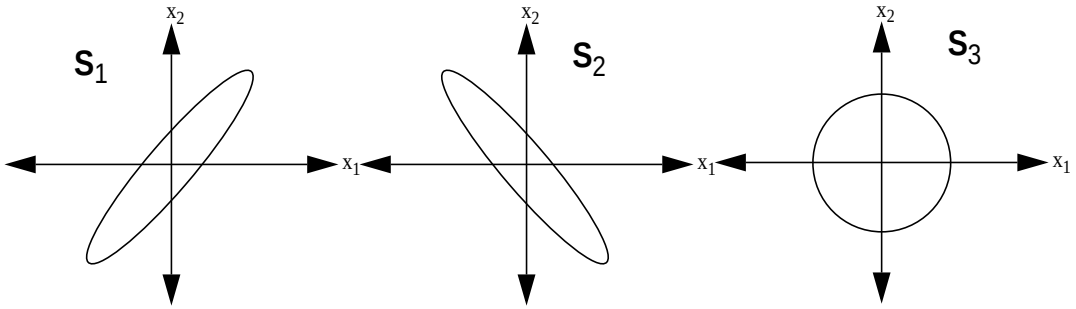
$$det(\Sigma) = det(\boldsymbol{U}\Sigma) = det(\boldsymbol{U})det(\Sigma) = det(\Sigma). \qquad \textit{(Eq. 2-1)}$$

Figure 2-3 shows problems of this nature in two dimensions. Variables in the first populations are positively correlated with sample covariance matrix $\boldsymbol{S}_1$, while the other populations show negative correlation ($\boldsymbol{S}_2$) or no correlation ($\boldsymbol{S}_3$). All populations have the same volume. In this case, the determinants of $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, and $\boldsymbol{S}_3$ are all identical. Note that although the generalized covariance cannot detect a rotational change in the covariance, the T^2 method would detect such a

rotational change as an out of control event.

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

$$\boldsymbol{S}_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \qquad \boldsymbol{S}_2 = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix} \qquad \boldsymbol{S}_3 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$



3. Eigenfactor Detection

3.1 Eigenspace Detection Method

In this section we present a new detection method which we term “eigenspace detection” 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 the eigen-decomposition of the covariance matrix. Because every sample covariance matrix $\boldsymbol{S}$ is real and symmetric, there is a real orthogonal matrix $\boldsymbol{V}$ and a real diagonal matrix Λ, such that $\boldsymbol{S} = \boldsymbol{V}\Lambda\boldsymbol{V}^T$. Furthermore, $\boldsymbol{S}$ has a spectral decomposition, and one can write $\boldsymbol{S}$ as

$$\boldsymbol{S} = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \sum_{i=1}^{p} (\sqrt{\lambda_i}\boldsymbol{v}_i)(\sqrt{\lambda_i}\boldsymbol{v}_i)^T, \qquad \textit{(Eq. 3-1)}$$

where p is the number of variables, λ_i is an eigenvalue and a diagonal element of Λ, and $\boldsymbol{v}_i$ is an eigenvector in $\boldsymbol{V}$. The above equation resembles very much how the sample covariance matrix is

computed. Let each column vector $\boldsymbol{X}_i = \left[x_{i1}\ x_{i2}\ \ldots\ x_{ip}\right]^T$ represent a *p*-variate random vector with density function $f(\boldsymbol{X}_i)=f(x_1,x_2,...,x_p)$; note the subscript *i* is omitted in the distribution because all $\boldsymbol{X}_i$ have identical distributions. If all column vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, ..., \boldsymbol{X}_n$ form *n* independent identically distributed observations, and we write $\boldsymbol{X} = \left[\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_i\ \ldots\ \boldsymbol{X}_n\right]^T$, then its mean corrected matrix is $\boldsymbol{X}_m = \left[\boldsymbol{X}_1 - \bar{\boldsymbol{X}}\ \boldsymbol{X}_2 - \bar{\boldsymbol{X}}\ \ldots\ \boldsymbol{X}_n - \bar{\boldsymbol{X}}\right]^T$. We then can express the sample variance-covariance matrix as the following

$$\boldsymbol{S} = \frac{\boldsymbol{X}_m^T \boldsymbol{X}_m}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}}). \qquad \textit{(Eq. 3-2)}$$

The matrix summation from the above equation is very similar to that in Eq. 3-1.

In short, the eigenspace matrix will be a decomposition of the sample covariance matrix obtained from multiple samples. Detection strategies using the eigenspace matrix compute second order statistics and use this information to detect subtle changes in the process. Probability distributions of the matrix are discussed in Section 3.2. 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. Moreover, 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. Application of eigenfactor analysis in semiconductor manufacturing is demonstrated in Section 5.

A corresponding second order detection method is also proposed that not only provides information on volume change but also identifies when there is an orientation shift in the covariance structure. Let λ_i be arranged in descending order, just like the ordering of singular values, and let $\boldsymbol{v}_i$ be *almost* the eigenvector associated with λ_i except that $\boldsymbol{v}_i$ is selected uniquely. We will come back to the selection of $\boldsymbol{v}_i$ 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

$$\sqrt{\lambda_i}\boldsymbol{v}_i,\ i = 1, 2, \ldots, p \qquad \textit{(Eq. 3-3)}$$

is called the *eigenfactor* and the matrix containing all the eigenfactors is the *eigenspace matrix* $\boldsymbol{E}$. 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 rewrite the spectral decomposition as

$$\boldsymbol{S} = \sum_{i=1}^{p}\lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \sum_{i=1}^{p}(\sqrt{\lambda_i}\boldsymbol{v}_i)(\sqrt{\lambda_i}\boldsymbol{v}_i)^T = (\boldsymbol{V}\Lambda^{1/2})(\boldsymbol{V}\Lambda^{1/2})^T = \boldsymbol{E}\boldsymbol{E}^T. \qquad \textit{(Eq. 3-4)}$$

So instead of tracking all eigenfactors $\sqrt{\lambda_i}\boldsymbol{v}_i,\ i = 1, 2, \ldots, p$ individually, we can monitor the full eigenspace matrix $\boldsymbol{E}$.

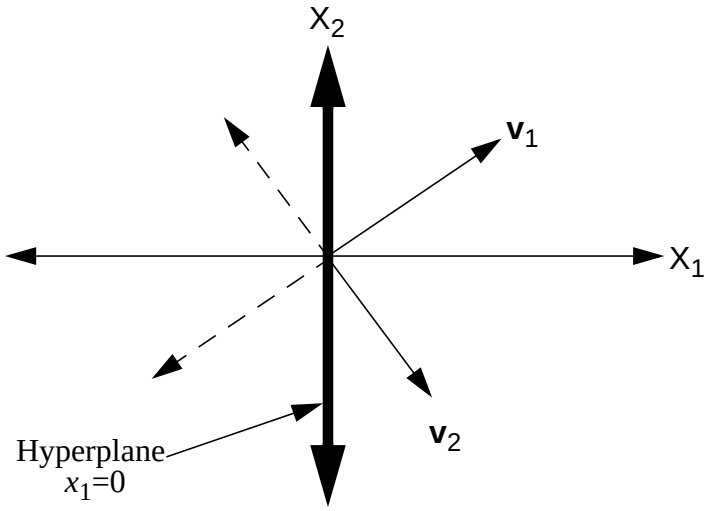
We now provide a selection procedure for eigenvectors such that $\boldsymbol{E}$ becomes a unique decomposition matrix from the sample covariance matrix $\boldsymbol{S}$. From its previous definition $\boldsymbol{E} = \boldsymbol{V}\Lambda^{1/2}$, there are still two possibilities that we can pick when selecting an eigenvector, i.e., if $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{S}$ then $-\boldsymbol{v}$ is also an eigenvector of $\boldsymbol{S}$. Note that for any square matrix any scalar multiple of an eigenvector is also an eigenvector. Because of this ambiguity associated with the eigenvectors, we desire a procedure to find a unique eigenvector for each positive eigenvalue. One way to do so is to find a hyperplane in $\Re^p$; 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*

$$\boldsymbol{n}_{\perp} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

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 $90°$ (this corresponds to the case $x_1 \geq 0$). Figure 3-1 presents graphically the selection of two unique eigenvectors in the two dimensional case; both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are on the right side of the hyperplane (x_1=0).

Figure 3-1: Selection procedure of a unique eigenvector: choose eigenvectors such that all lie on the same side of a given hyperplane.



Such a selection of eigenvector enables us to obtain a unique eigenvector. This eigenvector is called the unique eigenvector and $\boldsymbol{E}$ 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.

3.2 Distribution of Eigenspace Matrix *E*

We now provide some fundamental properties of the distribution of the sample eigenspace matrix $\boldsymbol{E}$. In order to establish the distribution for matrix $\boldsymbol{E}$, we must first study properties of the distribution of the sample covariance matrix $\boldsymbol{S}$. The distribution of $\boldsymbol{S}$ is called the Wishart distribution after its discoverer (Gupta and Nagar (2000), Arnold (1981)); it is defined as the sum of independent products of multivariate normal random vectors, as in Eq. 3-2. We shall think of the Wishart distribution as a generalized extension of the χ^2-distribution into a multivariate scenario.

Definition: *Let* $\boldsymbol{Z}=[\boldsymbol{Z}_1\ \boldsymbol{Z}_2\ ...\ \boldsymbol{Z}_n]^T$ *such that the* $\boldsymbol{Z}_i$ *are independently distributed as the p-variate normal distribution* $N_p(0,\Sigma)$*. Let* $\boldsymbol{W}=\boldsymbol{Z}^T\boldsymbol{Z}$*. Then* $\boldsymbol{W}$ *is said to have a Wishart distribution with n degrees of freedom (d.o.f.). Note that* $\boldsymbol{W}$ *is a p by p matrix and positive definite.*

From the above definition, we then summarize the sampling distribution results below. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be a random sample of size *n* from a *p*-variate normal distribution with mean μ and covariance matrix Σ. Then the following statement is true:

- $(n\text{-}1)\boldsymbol{S}$ where $\boldsymbol{S}$ is defined in Eq. 3-2, is distributed as a Wishart random matrix with *n*-1 d.o.f., i.e. $(n-1)\boldsymbol{S} \sim W_p(n-1, \Sigma)$.

More properties associated with the Wishart distribution are now stated. The proof of these properties is either provided or a reference is given. 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 $\boldsymbol{E}$. In the following theorems and corollaries, the (*n*-1) term from $(n\text{-}1)\boldsymbol{S}$ has been suppressed for simplicity.

Theorem 3.1 *Let* $\boldsymbol{S}\sim W_p(n,\Sigma)$ *and let* $\boldsymbol{S} = (\boldsymbol{V}\Lambda^{1/2})(\boldsymbol{V}\Lambda^{1/2})^T = \boldsymbol{E}\boldsymbol{E}^T$*. Then the distribution of* $\boldsymbol{E}$ *has a functional form of* $\boldsymbol{E} = \boldsymbol{A}^{-1}\boldsymbol{T}\boldsymbol{U}$*, where* $\boldsymbol{A}$ *is a square root matrix of the inverse of the population covariance matrix* $\Sigma^{-1} = \boldsymbol{A}^T\boldsymbol{A}$*,* $\boldsymbol{U}$ *is a given unitary matrix and* $\boldsymbol{T}$ *is associated with Bartlett's decomposition of a Wishart distribution matrix* (Gupta and Nagar (2000)).

Proof: In order to show that the distribution of matrix $\boldsymbol{E}$ depends on the distribution of $\boldsymbol{T}$, we need to use a transformation theorem found in Gupta and Nagar (2000) which allows us to transform any Wishart distribution with parameters (n,Σ) to a special form of Wishart distribution with the parameters $(n,\boldsymbol{I})$. Together with Bartlett's decomposition result, we have the following equality

$$\boldsymbol{S}_{new} = \boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{E}\boldsymbol{E}^T\boldsymbol{A}^T = (\boldsymbol{A}\boldsymbol{E})(\boldsymbol{A}\boldsymbol{E})^T, \qquad \textit{(Eq. 3-5)}$$

where $\boldsymbol{S}_{new}\sim W_p(n,\boldsymbol{I})$. As a result, we have

$$\boldsymbol{A}\boldsymbol{E} = \boldsymbol{T}\boldsymbol{U} \Rightarrow \boldsymbol{E} = \boldsymbol{A}^{-1}\boldsymbol{T}\boldsymbol{U}, \qquad \textit{(Eq. 3-6)}$$

where $\boldsymbol{U}$ is a unitary matrix and $\boldsymbol{T}$ is a lower triangular matrix with $t_{ii}>0$. Then $t_{ij},\ 1 \le j \le i \le p$ are independently distributed, with $t_{ii}^2 \sim \chi_{n-i+1}^2,\ 1 \le i \le p$ and $t_{ij} \sim N(0,1),\ 1 \le j < i \le p$.

Eq. 3-6 shows that the distribution of $\boldsymbol{E}$ is a function of the distribution of $\boldsymbol{T}$. Therefore, the asymptotic properties regarding the distribution of $\boldsymbol{E}$ depends strongly on the asymptotic properties of $\boldsymbol{T}$. We then study the asymptotic distribution of $\boldsymbol{T}$. The following theorem shows that the variance of each element in $\boldsymbol{T}$ goes to zero when the number of samples approaches infinity.

Theorem 3.2 *Let* $(n-1)\boldsymbol{S} \sim Wp(n-1, \boldsymbol{I}_p)$ *and* $(n-1)\boldsymbol{S} = \sqrt{n-1}\boldsymbol{T})(\sqrt{n-1}\boldsymbol{T})^T$, *where* $\boldsymbol{T}=(t_{ij})$ *is a lower triangular matrix with all its elements being independently distributed,* $t_{ii}>0$, $(n-1)t_{ii}^2 \sim \chi_{n-i+1}^2,\ 1 \le i \le p$ *and* $\sqrt{n-1}\,t_{ij} \sim N(0,1),\ 1 \le j < i \le p$. *Then, 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 $\boldsymbol{T}$ when $i \ne j$, we know $\sqrt{n-1}\,t_{ij}$ has the standard normal distribution with variance 1. Therefore, the variance of t_{ij} can be computed:

$$var(\sqrt{n-1}\,t_{ij}) = (n-1)var(t_{ij}) = 1 \Rightarrow var(t_{ij}) = \frac{1}{n-1}.$$

The limit of the variance tends to zero as $n \to \infty$, i.e. $\lim_{n\to\infty} var(t_{ij}) = \lim_{n\to\infty} \frac{1}{n-1} = 0$. As for the diagonal elements t_{ii} of $\boldsymbol{T}$, 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 χ_v^2 is 2υ, we then have

$$var((n-1)t_{ii}^2) = (n-1)^2 var(t_{ii}^2) = 2(n-i+1) \Rightarrow var(t_{ii}^2) = \frac{2(n-i+1)}{(n-1)^2}.$$

This limit goes to zero as n gets large, $\lim_{n\to\infty} var(t_{ii}^2) = \lim_{n\to\infty} \frac{2(n-i+1)}{(n-1)^2} = 0,\ 1 \le i \le p.$ We are now ready to show that the variance of t_{ii} goes to zero as n gets large.

Theorem 3.3 *Assume* $\lim_{n\to\infty} var(x^2) = 0$. *Then for* $y = \sqrt{x^2} = x \ge 0$, *we have* $\lim_{n\to\infty} var(y) = 0$.

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 cannot be zero from the definition, i.e.

$$var(x^2) = \int_0^\infty (x^2 - E(x^2)) f(x^2) dx = (x_1^2 - E(x^2))^2 f(x^2 = x_1^2) + (x_2^2 - E(x^2))^2 f(x^2 = x_2^2) \ne 0.$$

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 $y = \sqrt{x^2}$ is a continuous, monotone function with one-to-one mapping; thus the distribution of y must also be a delta function at $y=\sqrt{a}$, i.e. $\lim_{n\to\infty} Prob(y= \sqrt{a}) = 1$. Hence $\lim_{n\to\infty} var(y) = 0$.

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

Theorem 3.3 concludes that the each element of the eigenspace matrix $\boldsymbol{E}$ converges to something since its variance goes to zero as n gets large. However, we want to find out exactly what $\boldsymbol{E}$ converges to; in particular, we wish to determine if the sample eigenspace matrix $\boldsymbol{E}$ (where $\boldsymbol{E}$ is based on the sample covariance matrix $\boldsymbol{S}$) converges to $\boldsymbol{F}$ (the population eigenspace matrix based on Σ). We conclude this chapter by proving that the sample eigenspace matrix $\boldsymbol{E}$ is a consistent estimator of the population eigenspace matrix $\boldsymbol{F}$.

Theorem 3.4 *Suppose* $\boldsymbol{E}$ *is the sample eigenspace matrix of a sample covariance matrix* $\boldsymbol{S}$*, which converges to a population covariance matrix* Σ*. Let* $\boldsymbol{F}$ *be the eigenspace matrix of* Σ*. Then* $\boldsymbol{E}$ *converges to* $\boldsymbol{F}$**.**

Proof: As n gets large, we know that $\boldsymbol{E}$ converges to a matrix $\hat{\boldsymbol{E}}$ (from Theorem 3.3). Let us assume that $\hat{\boldsymbol{E}} \neq \boldsymbol{F}$ and prove the theorem by contradiction. As n gets large, we know

$$\boldsymbol{S} = \hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^T \to \Sigma = \boldsymbol{F}\boldsymbol{F}^T.$$

But the eigenspace matrix is unique, as a result $\hat{\boldsymbol{E}} = \boldsymbol{F}$.

4. Simulation Results

In this section, we study the sampling distribution of the sampling eigenfactors as a function of the sample size n from a finite population of size N. The sampling distribution is obtained by sampling with replacement k times. The simulation results show that the sampling eigenfactors indeed converge to the true eigenfactors. In both examples the population covariance matrix used to generate N random samples is

$$\Sigma = \begin{bmatrix} 21.81 & -12.05 \\ -12.05 & 27.89 \end{bmatrix} = \begin{bmatrix} 3.75 & 2.78 \\ -4.82 & 2.17 \end{bmatrix} \begin{bmatrix} 3.75 & -4.82 \\ 2.78 & 2.17 \end{bmatrix} = \boldsymbol{F}\boldsymbol{F}^T,$$

and the first population eigenfactor $\boldsymbol{F}_1$ is $\begin{bmatrix} 3.75 & -4.82 \end{bmatrix}^T$. In example 1, N=20,000, n=50 and k=100 times. In example 2, N=20,000, n=500 and k=500 times.

Figure 4-1: The sampling distribution of the first eigenfactor. Dots indicate sample eigenfactors, where F_1 is the known "true" first population eigenfactor.

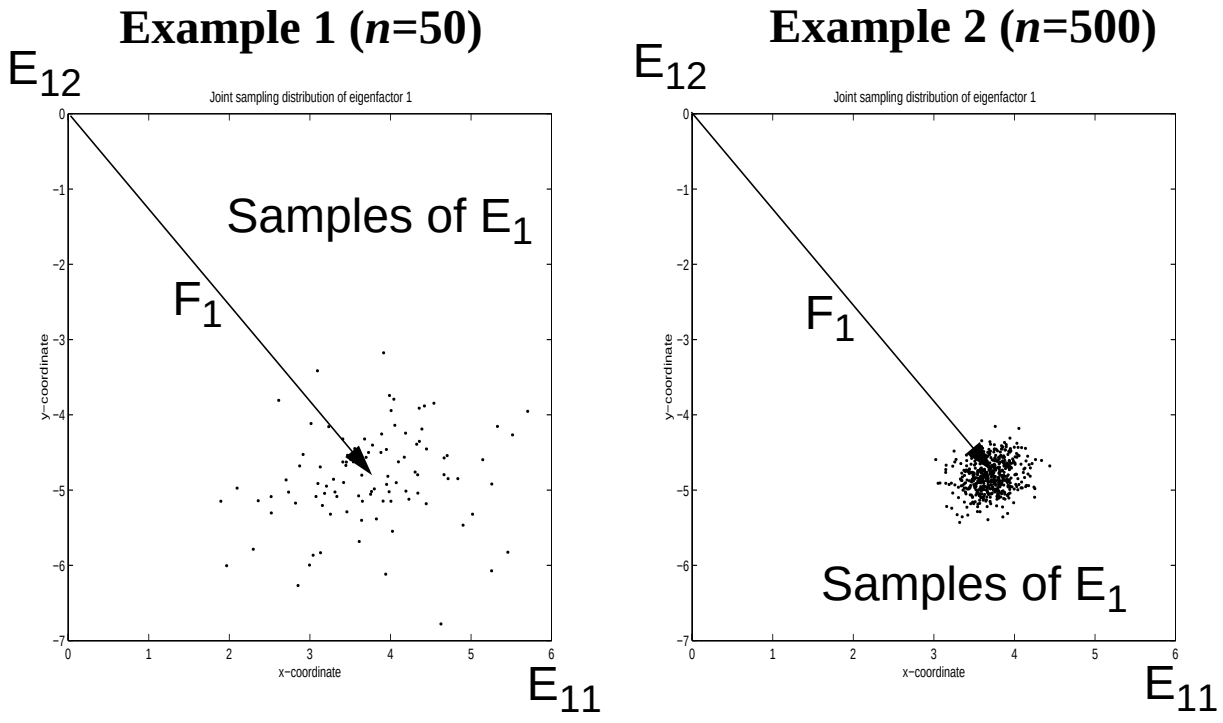


Figure 4-1 illustrates that the sampling distribution of the first eigenfactor becomes more tightly centered around the true value $\boldsymbol{F}_1 = \begin{bmatrix} 3.75 \\ -4.82 \end{bmatrix}$ as the number of samples increases. As predicted, the variance of the sampling eigenfactor gets smaller as the number of samples increases. More simulation results can be found in Chen (2001).

We now have the apparatus needed to monitor the covariance structure of a manufacturing process. We select one or more eigenfactors to monitor. After gathering each new set of samples (within some given window size), we compute the sample eigenfactor, and compare to the known distribution we expect of such vectors based on prior process characterization. If either the directionality or magnitude of this vector deviates significantly from the distribution we signal an out of control condition. Furthermore, we can decide if the new eigenfactor belongs to one of several alternative distributions indicative of specific error conditions.

5. Application to Optical Emission Spectra (OES)

5.1 Eigenspace Analysis on Optical Emission Spectra (OES)

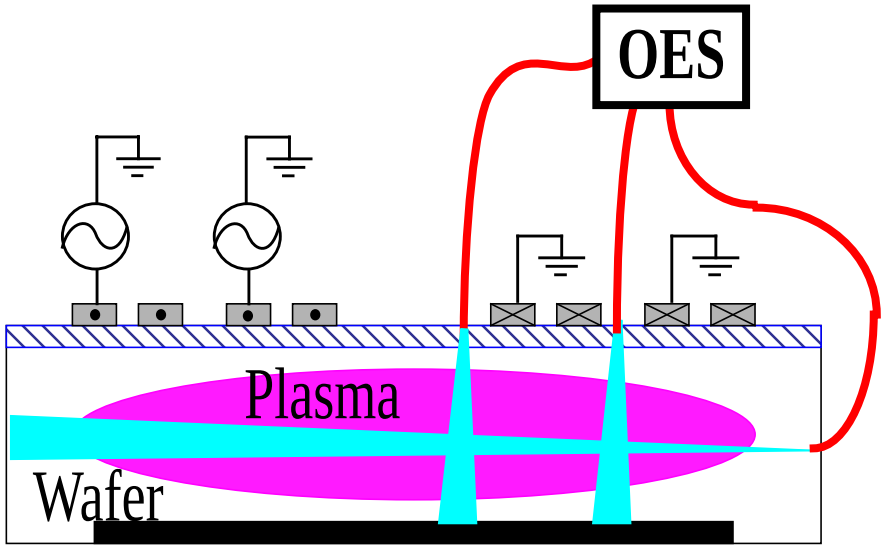
We are now ready to demonstrate this technique on an application that requires multivariate analysis. Optical emission spectra (OES) have traditionally been used to detect endpoints in semiconductor plasma etch processes (Wolf and Tauber (1986), Chang and Sze (1996)). We first describe the experimental setup and data collection.

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 Lee, *et. al.*, (1995), Litvak (1996), Lee and Spanos (1995), and Spanos, *et. al.*, (1992)). Real-time equipment data together with multivariate analysis is used to detect possible faults or process transitions.

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 the endpoint, with only one or two spectral wavelengths used for endpoint detection.

Figure 5-1: Optical emission spectroscopy experiment setup



Spectra from a side view optical port were collected during an etch process consisting of approximately one thousand spectral channels each sampled every 600 milliseconds (Le (1997)). 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 (Le (1997)).

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

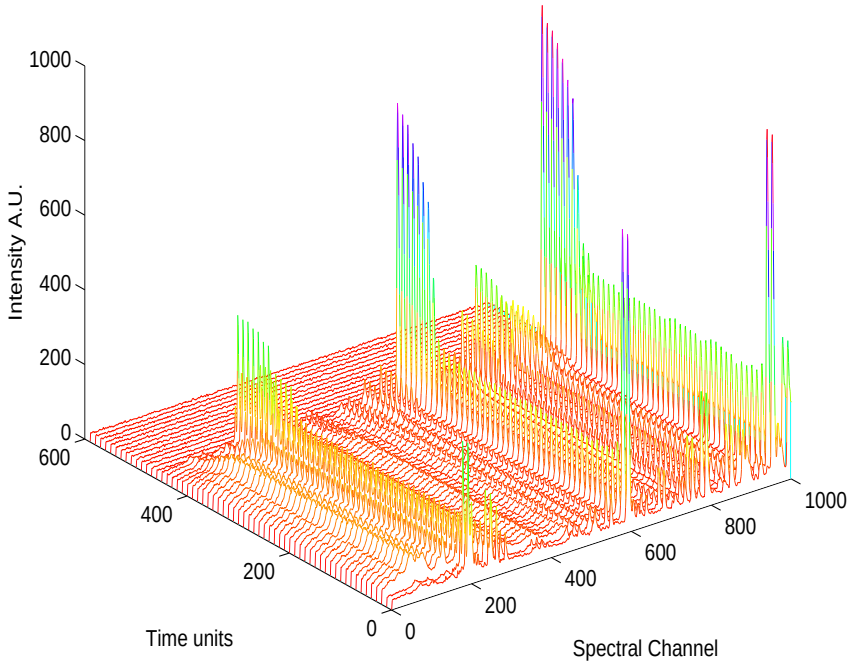
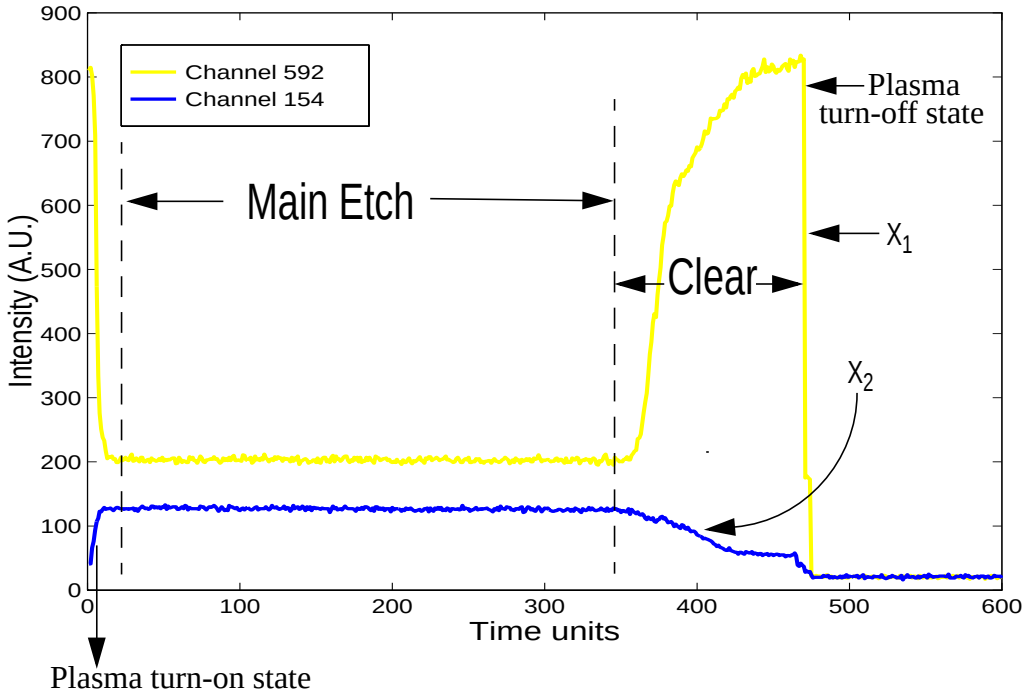


Figure 5-3: Two spectral lines showing different behavior as endpoint is reached.

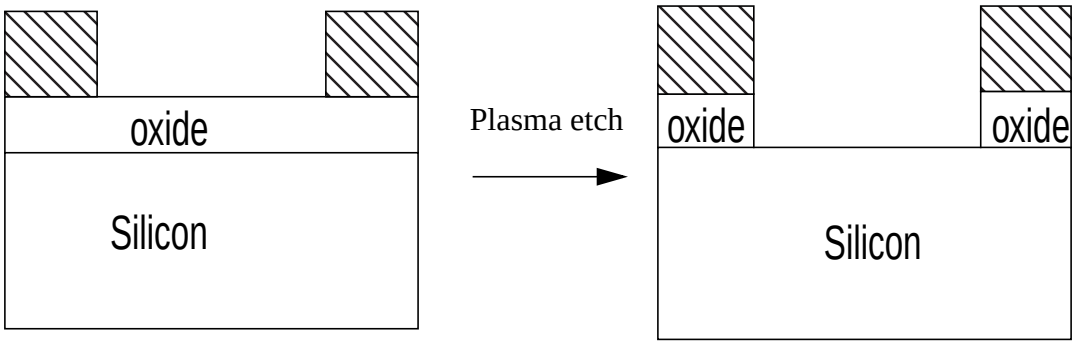


5.1.2 Endpoint Detection

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 (Sze (1988)). 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 the T^2 statistic and PCA with T^2 have been demonstrated to work well with OES data in certain cases (White, *et. al*., (2000) and Le (1997)). 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 these circumstances. 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: Endpoint is reached when the intended etch layer (oxide) is completely 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 on 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 the main etch region (see Figure 5-3) are calculated from the data. The sample mean of the endpoint is then computed from the data. With the assumption that the sample mean and covariance matrix are fairly good estimates of the population mean and covariance, 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

$$T^2 = (\boldsymbol{x} - \bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq \chi_p^2(\alpha).$$

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.

5.2 Low Open Area OES Endpoint Detection Results

For the low open area OES data, spectra were collected at 5 Hertz with an integration time of 15 milliseconds (Le (1997)). 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. In other words, we seek to characterize an eigenfactor $\boldsymbol{E}_{\text{ep1}}$ that identifies the "endpoint" condition.

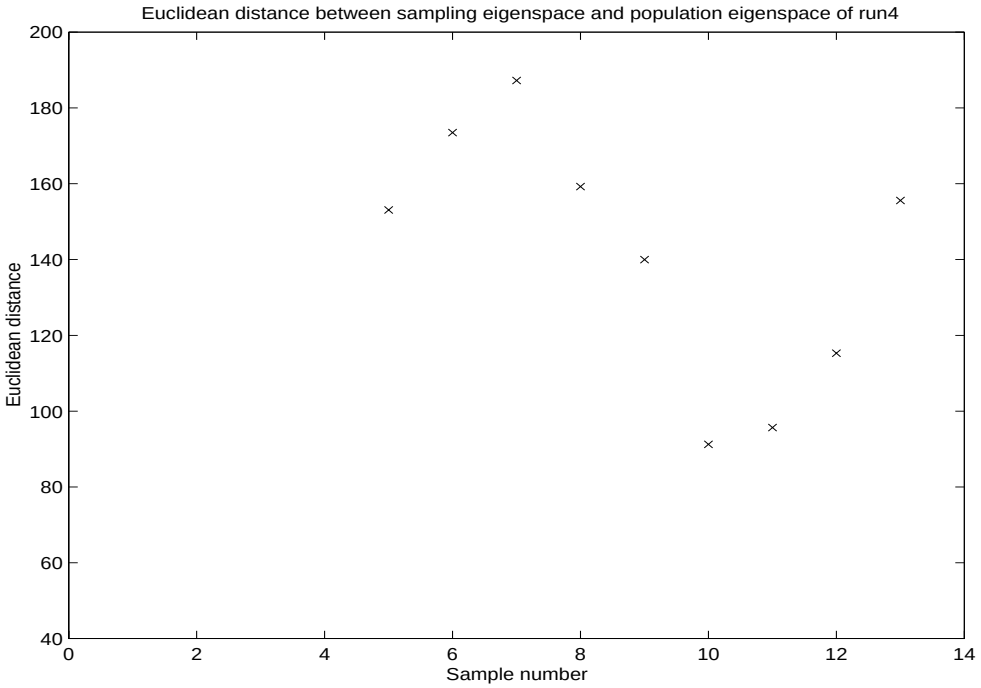
The endpoint is characterized using 100 samples prior to the plasma's turn-off. Furthermore, we use principal components analysis for data reduction, and the first principal component alone captures about 80% of the total variation out of 1087 spectra. We then only monitor the first eigenfactor through 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 1 and the endpoint eigenfactor 1 is defined to be

$$\left\| \boldsymbol{E}_1 - \boldsymbol{E}_{ep1} \right\| = \sqrt{\sum_{i=1}^{1087} (\boldsymbol{E}_{1i} - \boldsymbol{E}_{ep1i})^2}, \qquad \textit{(Eq. 5-1)}$$

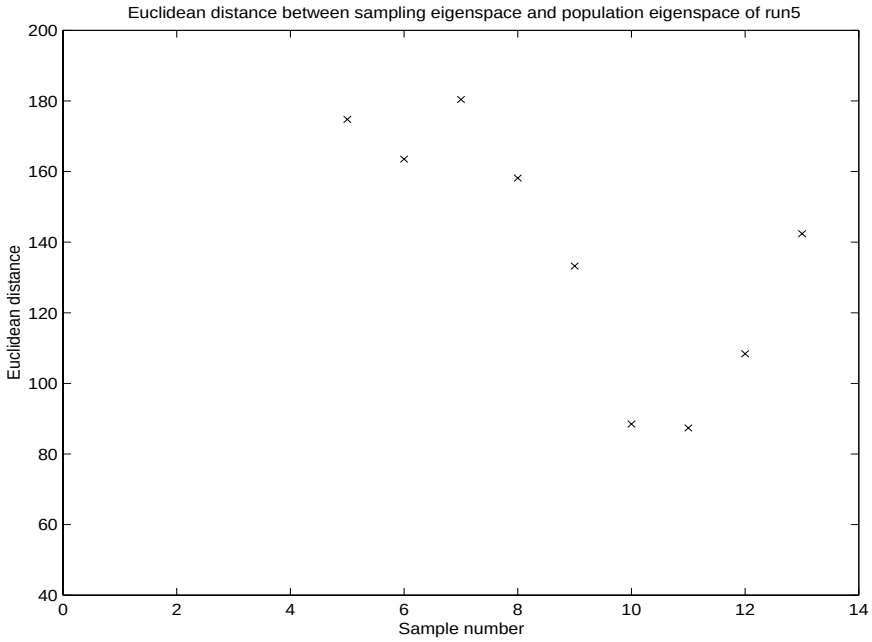
where $\boldsymbol{E}_{\text{ep1}}$ is eigenfactor 1 of the endpoint. This distance is computed to provide a measure of closeness. Note that the Euclidean distance does not include any variance or standard deviation term; the variance is later discussed and included in the control limits. Figure 5-5 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 ($\boldsymbol{E}_1$-$\boldsymbol{E}_{\text{ep1}}$) 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 indicating a "match" with the endpoint condition near these samples. The Euclidean norm diverges when the sampling window leaves the endpoint population indicating decreasing similarity between the plasma state and the endpoint condition.

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



In Chen (2001), additional examples and discussion of eigenspace detection applied to OES are presented. In particular, robustness issues are considered when the eigenfactor of the sampling window of any run is compared with the eigenfactor characterized in a previous run.

6. Conclusions

In this paper, we introduce a new eigenspace detection strategy to detect subtle covariance structure change. The uniqueness in the eigenspace enables us to address the consistency issues associated with the estimator. Key theorems related to eigenspace detection strategy and probability density distribution of the eigenspace matrix are derived. With the known probability density function, control limits of certain confidence interval can then be established for the eigenspace matrix or dominant eigenfactors. Simulation results on sampling distribution as a function of the sample size demonstrate this consistency.

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 semiconductor manufacturing has proven useful for low open area endpoint detection.

7. Acknowledgment

The authors are grateful for the financial support of the M. I. T. Leaders for Manufacturing Variation Reduction Research Program and the M. I. T. Leaders for Manufacturing GM Dimen-

sional Analysis Project.

8. References

S. F. Arnold, *The Theory of Linear Models and Multivariate Analysis*, John Wiley & Sons, 1981.

K. H. Chen, *Data-Rich Multivariate Detection and Diagnosis Using Eigenspace Analysis*, Ph. D. Thesis, MIT Aeronautics/Astronautics, 2001.

C. Y. Chang and S. M. Sze, *ULSI Technology,* McGraw-Hill, 1996.

A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*, Chapman & Hall/CRC, 2000.

R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, *Fourth Edition*, Prentice-Hall, 1998.

S. F. Lee, E. D. Boskin, H. C. Liu, E. H. Wen and C. J. Spanos, "RTSPC: A Software Utility for Real-Time SPC and Tool Data Analysis," *IEEE Trans. Semicond. Manufact.*, vol. 8, no. 1, pp. 17-25, Feb. 1995.

M. S. Le, *Variation Reduction in Plasma Etching via Run-to-Run Control and Endpoint Detection*, Masters Thesis, MIT Electrical Engineering and Computer Science, 1997.

H. Litvak, "End point control via optical emission spectroscopy," *J. Vac. Sci. Technol.*, vol. 14, no. 1, pp. 516-520, Jan./Feb. 1996.

S. F. Lee and C. J. Spanos, "Prediction of Wafer State After Plasma Processing Using Real-Time Tool Data," *IEEE Trans. Semicond. Manufact*., vol. 8, no. 3, pp. 252-261, Nov. 1995.

J. F. MacGregor and T. Kourti, "Statistical Process Control of Multivariate Processes," *Control Eng. Practice*. Vol. 3, pp. 403-414, 1995.

D. C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 1996.

C. J. Spanos, H. F. Guo, A. Miller and J. Levine-Parrill, "Real-Time Statistical Process Control Using Tool Data," *IEEE Trans. Semicond. Manufact.*, vol. 5, no. 4, pp. 308-318, Nov. 1992.

S. Sharma, *Applied Multivariate Techniques*, Wiley, 1996.

S. M. Sze, *VLSI Technology, Second Edition*, McGraw-Hill, 1988.

D. A. White, B. E. Goodlin, A. E. Gower, D. S. Boning, K. Chen, H. H. Sawin and T. J. Dalton, "Low Open-Area Endpoint Detection Using a PCA-Based T^2 Statistic and Q Statistic on Optical Emission Spectroscopy Measurements," *IEEE Trans. Semicond. Manufact.*, vol. 13, no. 2, pp. 193-207, May 2000.

B. M. Wise, N. L. Ricker, D. J. Veltkamp and B. R. Kowalski, “A Theoretical Basis for the Use of Principapal Components Models for Monitoring Multivariate Processes,” *Process Control and Quality*, vol. 1, no. 1, pp. 41-51, 1990.

S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era (Volume 1 - Process Technology*), Lattice Press, 1986.