APPLICATION OF REDUCED-RANK MULTIVARIATE METHODS TO THE MONITORING OF SPATIAL UNIFORMITY OF WAFER ETCHING

Michael Nikolaou
Chemical Engineering Department
University of Houston
Houston, TX 77294-4004
e-mail: nikolaou@uh.edu

Andrew D. Bailey, III
Lam Research Corporation
4650 Cushing Parkway
Fremont, CA 94538-6470
e-mail: andrew.bailey@lamrc.com

KEYWORDS: Plasma etching, principal component analysis, singular value decomposition, image processing.

ABSTRACT: We provide a smooth introduction to reduced-rank analysis via singular-value decomposition, and show how it can be used to monitor images of etched silicon wafers. An industrial case study is discussed.

1 INTRODUCTION

Spatially uniformity is necessary for high yields in a number of crucial processes of the semiconductor manufacturing industry, such as etching or deposition of thin films and chemical-mechanical planarization (CMP). In plasma etching, good spatial uniformity is the result of both appropriate design of etching tools as well as development of successful recipes. For either of these tasks, the designer or operator must be able to assess spatial uniformity characteristics, understand similarities and differences between tools or recipes, and apply criteria for the monitoring of spatial uniformity from tool to tool or run to run. Because uniformity is usually expressed in terms of a single number (e.g., 3σ/[average etch depth]) very different spatial uniformity profiles may result in the same numerical value of uniformity (Figure 1), thus masking important information that could be useful in a number of ways related to tool or recipe performance.

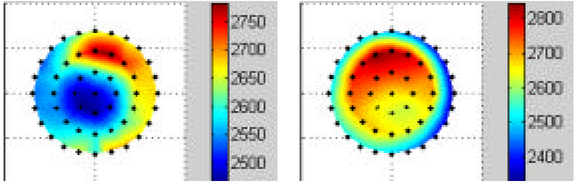


Figure 1 – Etch rate profiles on 300-mm wafer surface, interpolated over 49 measurement points (black dots). Both wafers correspond to virtually the same numerical uniformity value, but exhibit very different etch patterns.

In this presentation we provide a brief tutorial overview of the fundamentals of reduced-rank analysis, and show how it can be applied to the analysis, comparison, and monitoring of images corresponding to etch patterns of silicon wafers, as well as multivariate statistical process control.

2 COMPRESSION OF COLLINEAR DATA VIA SINGULAR VALUE DECOMPOSITION (SVD)

2.1 Basic case: Deterministic signals, no noise

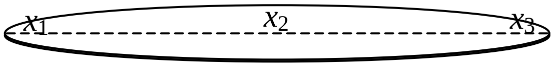


Figure 2 – Etch rate measurement points

❑ *An unrealistic but instructional example setting*

Suppose that etch rates, x_1, x_2, x_3 are exactly measured at three points (edge/center/edge) along the diameter of a wafer, as shown in Figure 2. As we collect data, wafer after wafer, we want to be able to use numbers that describe how similar the etch rate profiles are, and whether they are the result of a consistently performing process.

❑ *Noiseless data are collected*

Note that, for now, the data are assumed to be exact, i.e. *there is no measurement noise*. A set of data collected is shown in the matrix **X** below, and Figure 3.

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 2600 & 3348 & 3361 \\ 2700 & 3423 & 3311 \\ 2800 & 3392 & 2907 \\ 2900 & 3393 & 2609 \\ 3000 & 3527 & 2757 \\ 3100 & 3745 & 3182 \\ 3200 & 3900 & 3400 \\ 3300 & 3919 & 3163 \\ 3400 & 3882 & 2740 \\ 3500 & 3934 & 2614 \\ 3600 & 4118 & 2927 \\ 3700 & 4327 & 3324 \end{bmatrix} \end{array} \triangleq [\mathbf{x}_1 \; \mathbf{x}_2 \; \mathbf{x}_3] \qquad (1)$$

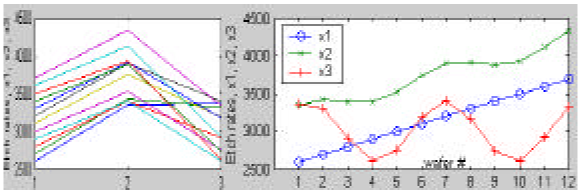


Figure 3 – Hypothetical etch rate profiles for 12 wafers (left) and Hypothetical local etch rates vs. wafer # (right).

❑ *Data collinearity and computation of matrix rank*

Are the variables x_1, x_2, x_3 linearly dependent? I.e. is there a nonzero vector $\mathbf{a} \triangleq [a_1 \; a_2 \; a_3]^T$ such that

$$a_1 x_1 + a_2 x_2 + a_3 x_3 = 0 \Leftrightarrow \mathbf{x}^T \mathbf{a} = 0 \qquad (2)$$

If so, then the data satisfy the relationship (**model equation**)

$$a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + a_3 \mathbf{x}_3 = 0 \Leftrightarrow \mathbf{X}\mathbf{a} = \mathbf{0} \quad \text{for } \mathbf{a} \neq \mathbf{0} \qquad (3)$$

E.g., if two of the column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ (in $\Re^{12}$) are linearly independent, the rank of the full data set matrix **X** will be 2, and we will be able to express each of the column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ (in $\Re^{12}$) as a linear combination of two basis vectors $\mathbf{y}_1$, $\mathbf{y}_2$ (in $\Re^{12}$).

A numerically robust method to check whether eqn. (3) is valid is the singular value decomposition (SVD).

Sidebar – Singular value decomposition (SVD)
Theorem: A matrix **X** of dimensions $m \times n$ can be factorized as

$$\overset{n}{\boxed{\mathbf{X}}}_{m} = \overset{m}{\boxed{\mathbf{U}}}_{m} \overset{n}{\boxed{\begin{matrix} s_1 & & \\ & \ddots & \\ & & s_r \end{matrix}}}_{m} \overset{n}{\boxed{\mathbf{V}^T}}_{n} \qquad (4)$$

where $\mathbf{U} \triangleq [\mathbf{u}_1 \vdots \cdots \vdots \mathbf{u}_m] \in \Re^{m \times m}$ and $\mathbf{V} \triangleq [\mathbf{v}_1 \vdots \cdots \vdots \mathbf{v}_n] \in \Re^{n \times n}$ are orthonormal matrices[1] (i.e. $\mathbf{UU}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$) whose columns are the (normalized) eigenvectors of the matrices $\mathbf{XX}^T$ and $\mathbf{X}^T\mathbf{X}$, respectively; $s_1 \geq \cdots \geq s_r$ are the square roots of the nonzero eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$ (or $\mathbf{XX}^T$) and $r \hat{=} rank(\mathbf{X}) = rank(\mathbf{XX}^T) = rank(\mathbf{X}^T\mathbf{X})$.

Remark: Eqn. (4) can also be written as

$$\overset{n}{\boxed{\mathbf{X}}}_{m} = s_1 \underbrace{\overset{1}{\boxed{\mathbf{u}_1}}_{m}}_{\text{"}score\text{"}1} \underbrace{\overset{n}{\boxed{\mathbf{v}_1^T}}_{1}}_{\text{"}loading\text{"}1} + \cdots + s_r \underbrace{\overset{1}{\boxed{\mathbf{u}_r}}_{m}}_{\text{"}score\text{"}r} \underbrace{\overset{n}{\boxed{\mathbf{v}_r^T}}_{1}}_{\text{"}loading\text{"}r} \tag{5}$$

$$\equiv \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \equiv \sum_{i=1}^{r} \mathbf{y}_i \mathbf{v}_i^T$$

Application of SVD (e.g. in Matlab ®) to the data matrix **X**, eqn. (1), yields that the rank of **X** is 2, and the matrix **X** can be decomposed as

$$\mathbf{X} = 19973 \underbrace{\overset{1}{\boxed{\begin{matrix}-0.26882\\-0.2727\\-0.26381\\-0.25876\\-0.26976\\-0.29076\\-0.30431\\-0.30144\\-0.2919\\-0.29302\\-0.30999\\-0.32996\end{matrix}}}_{12}}_{\text{"}score\text{"}1,\,\mathbf{y}_1} \underbrace{\overset{3}{\boxed{\begin{matrix}-0.54865 & -0.65112 & -0.52443\end{matrix}}}_{1}}_{\text{"}loading\text{"}1,\,\mathbf{v}_1^T}$$

$$+1233.7 \underbrace{\overset{1}{\boxed{\begin{matrix}0.53687\\0.44752\\0.14112\\-0.10007\\-0.068095\\0.1339\\0.20911\\0.005124\\-0.31245\\-0.44865\\-0.31508\\-0.13062\end{matrix}}}_{12}}_{\text{"}score\text{"}2,\,\mathbf{y}_2} \underbrace{\overset{3}{\boxed{\begin{matrix}-0.52217 & -0.22301 & 0.82317\end{matrix}}}_{1}}_{\text{"}loading\text{"}2,\,\mathbf{v}_2^T}$$

$$+ \underbrace{\overset{1}{\boxed{\begin{matrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{matrix}}}_{12}}_{\text{"}score\text{"}3,\,\mathbf{y}_3} \underbrace{\overset{3}{\boxed{\begin{matrix}-0.65293 & 0.72548 & -0.21764\end{matrix}}}_{1}}_{\text{"}loading\text{"}3,\,\mathbf{v}_3^T} \tag{6}$$

The above eqn. (6) implies that each row of the matrix **X** can be written as a linear combination of the row vectors *loading*1 and *loading*2, i.e.

$$\underbrace{[x_1 \;\; x_2 \;\; x_3]}_{\mathbf{x}^T} = \underbrace{y_1}_{\text{"}score\text{"}1} \underbrace{\overset{3}{\boxed{\mathbf{v}_1^T}}}_{\text{"}loading\text{"}1} + \underbrace{y_2}_{\text{"}score\text{"}2} \underbrace{\overset{3}{\boxed{\mathbf{v}_2^T}}}_{\text{"}loading\text{"}2} \tag{7}$$

Because **V** is orthonormal, eqn. (7) yields the sought eqn. (2), i.e.

$$\mathbf{x}^T \mathbf{v}_3 = 0 . \tag{8}$$

❑ *Loadings can be thought of as basic shapes that can be used to represent the raw data*

Note that the row vectors *loading*1 and *loading*2 in eqn. (7) are **the same** for all rows of data triplets x_1, x_2, x_3; they appear to be related to the system and not to any individual wafer. Therefore, *loading*1 and *loading*2 can be thought of as **two basic shapes** (Figure 4)**, whose linear combination (sum weighted by score entries) can produce any of the 12 measured shapes**.

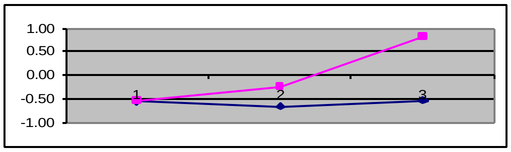


Figure 4 – Loadings, eqn. (6). Because loadings are orthonormal, the order of magnitude of their entries is 1. The sign is unimportant. The two shapes attempt to capture the curvature in the etch rate profile.

❑ *Monitoring scores gives a complete picture of the data*

It follows from the preceding discussion that **one can simply observe the scores (compressed data, values of *principal components* – hence PCA), to capture all information about the original data**. In other words, instead of looking at Figure 3, one can look at Figure 5.

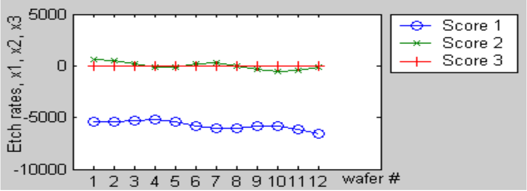


Figure 5 – Scores for the data in Figure 2, according to eqn. (6). Note that Score 3 is identically 0, which is precisely the equation sought in eqn. (2).

❑ *Eqn. $rank(\mathbf{X}) = 2$ implies data points fall on a plane*

Figure 6 shows 3-D plots of the data from two different viewpoints. The second viewpoint clearly shows that data fall on a plane. The new axes (not shown) in the figure) are produced by multiplying **V**, eqn.(4), times each of the original axes.

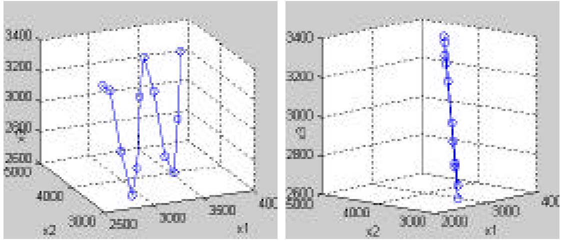


Figure 6 – 2-D world in 3-D data ("collinearity").

❑ *Loadings can also be thought of as weights used to relate original data to scores (compressed data)*

An immediate corollary of eqn. (7) is that each of the data column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ can be written as a linear combination of the column vectors ("scores") $\mathbf{y}_1$ and $\mathbf{y}_2$. If the score vectors $\mathbf{y}_1$, $\mathbf{y}_2$ are thought of as corresponding to two new variables, y_1, y_2, then y_1, y_2 are related to x_1, x_2, x_3 as follows: Because the loadings are orthonormal, we can post-multiply eqn. (5) by $\mathbf{v}_j$ to get

$$\overset{n}{\boxed{\mathbf{X}}}_{m} \; \overset{1}{\boxed{\mathbf{v}_j}}_{n} = s_j \underbrace{\overset{1}{\boxed{\mathbf{u}_j}}_{m}}_{\text{"}score\text{"}j} \hat{=} \mathbf{y}_j \tag{9}$$

[1] The analysis is valid for complex-valued matrices as well, with *Hermitian* in place of *transpose.*

or, row by row,

$$y_j = [x_1 \cdots x_n]\mathbf{v}_j \equiv \mathbf{x}^T\mathbf{v}_j = \mathbf{v}_j^T\mathbf{x} \tag{10}$$

or, in vector/matrix form,

$$\boxed{\mathbf{y}} = \boxed{\quad \mathbf{V}^T \quad}\boxed{\mathbf{x}} \Leftrightarrow \boxed{\mathbf{x}} = \boxed{\quad \mathbf{V} \quad}\boxed{\mathbf{y}} \tag{11}$$

(The new variables **y** are also called ***principal components***, and the reason why will become clear in section 2.3.)

Thus, for this particular example we get, using eqn.(10), that the two nonzero score variables are

$$y_1 = [x_1 \ \ x_2 \ \ x_3]\begin{bmatrix} -0.54865 \\ -0.65112 \\ -0.52443 \end{bmatrix}, \quad y_2 = [x_1 \ \ x_2 \ \ x_3]\begin{bmatrix} -0.52217 \\ -0.22301 \\ 0.82317 \end{bmatrix} \tag{12}$$

and that the last score variable should be trivially equal to zero, i.e.

$$y_3 = [x_1 \ \ x_2 \ \ x_3]\begin{bmatrix} -0.65293 \\ 0.72548 \\ -0.21764 \end{bmatrix} = 0 \tag{13}$$

which is the same as eqn. (8).

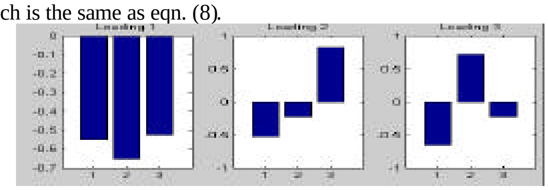


Figure 7 – Loadings as weights of original variables used to construct new variables by linear combination. The first two bar charts present the exact same numbers as in Figure 4.

This gives us the second interpretation of loadings: They are the **vectors of coefficients by which we weight the original variables in linear combinations that produce a new set of variables (the "scores") or the linear relationships among the original variables**. To emphasize this interpretation we are showing the corresponding loadings in Figure 7.

❑ *The preceding findings about X can be used to monitor the system*

If the system etches subsequent wafers in the same way, **it is reasonable to expect that data points** (x_1, x_2, x_3) **will be produced that are related as before**, i.e. by eqn. (2). That means, equivalently, that if one first constructs 2 new variables y_1, y_2 in terms of eqn. (10) (or, equivalently, (11)), then the value of the *residual error* (cf. eqn. (7))

$$\mathbf{e}^T \triangleq \underbrace{[x_1 \ \ x_2 \ \ x_3]}_{\mathbf{x}^T} - \left(\underbrace{y_1}_{\text{"}score\text{"}1} \underbrace{\overset{3}{\boxed{\quad \mathbf{v}_1^T \quad}}}_{\text{"}loading\text{"}1} + \underbrace{y_2}_{\text{"}score\text{"}2} \underbrace{\overset{3}{\boxed{\quad \mathbf{v}_2^T \quad}}}_{\text{"}loading\text{"}2} \right) \tag{14}$$

$$= (\mathbf{x} - \mathbf{P}\mathbf{P}^T\mathbf{x})^T$$

for each new data triplet should be equal to zero, or, equivalently,

$$\|\mathbf{e}\|^2 \triangleq \mathbf{e}^T\mathbf{e} = 0 \Leftrightarrow \mathbf{x}^T(\mathbf{I} - \mathbf{P}\mathbf{P}^T)\mathbf{x} = 0 \tag{15}$$

where the matrix **P** consists of the first *r* columns of **V**. (The reason for using eqn. (15), instead of simply $\mathbf{e} = \mathbf{0}$, is that it can easily be extended to handle noisy data, as will be shown below).

Consider now the new data shown in Figure 8 below.

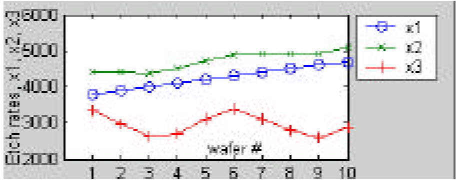


Figure 8 – Data set from 10 new wafers.

Applying the test of eqn. (15) to the new data shown above yields the results of Figure 9. It is clear that two data points (#7 and #8) do not fall on the zero line as they should. These points indicate that the behavior of the system that etched these wafers is different from before.

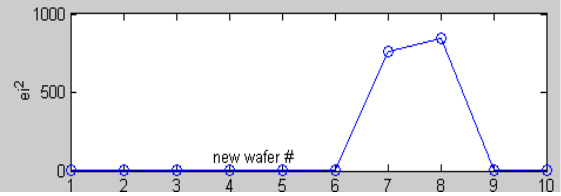


Figure 9 – Square errors for the 10 new data sets, Figure 8.

2.2 Noisy signals

❑ *SVD on the noisy counterpart of* **X** *reveals similar relationship among* x_1, x_2, x_3.

Table 1 – Noisy data

#	x_1	x_2	x_3
1	2585	3373	3353
2	2874	3586	3374
3	2809	3311	2861
4	2759	3355	2562
5	3175	3602	2763
6	3071	3753	3258
7	3424	3933	3486
8	3368	3974	3263
9	3526	3887	2709
10	3523	4034	2735
11	3546	4209	2910

Consider that measurements of x_1, x_2, x_3 are obtained with measurement noise. The data of Table 1 are obtained (cf. eqn. (1)). SVD on the data of Table 1 yields singular values equal to 20219, 1206.5, 226.15 (cf. eqn.(6)). Corresponding eigenvalues (singular values squared) are shown in Figure 10. The smallest singular value is two orders of magnitude smaller than the largest one, indicating that it is probably equal to zero. But the second singular value is also an order of magnitude smaller than the largest singular value. Is it really non-zero or zero? How many singular values should be retained? What is the underlying rank of the data? (Figure 11-b indicates that, viewed from a certain viewpoint, the data appear to fall on a plane, verifying that the smallest singular value is most probably zero.)

❑ *How many singular values of* **X** *are really nonzero?*

We need to understand how the singular values of the noisy data of Table 1 are related to the singular values of the noiseless data in Figure 3. Let us call the noiseless data matrix Ξ. Then

$$\mathbf{X} = \Xi + \mathbf{E} \tag{16}$$

where **E** is a matrix that contains measurements errors. Note that for the data in Table 1

$$rank(\mathbf{X}) = 3 > rank(\Xi) = 2 \tag{17}$$

There is no closed-form expression that relates the singular values of **X**, $\sigma_{\mathbf{X}}$, to those of Ξ, σ_{Ξ}, but there are asymptotic, Taylor-series-type results that are valid for "small" **E**, and bounds such as ([1], p. 419):

$$\left|s_i(\mathbf{X}) - s_i(\Xi)\right| \le \|\mathbf{E}\|_{i2} = s_{\max}(\mathbf{E}) \tag{18}$$

Note that no assumptions about the statistics of **E** need to be made for eqn. (18) to be valid. Eqn. (18) indicates that zero singular values of Ξ will appear as "small" singular values of **X**.

How many of the "large" singular values of **X** are really nonzero (cf. Figure 10)? The answer should be such that $\mathbf{X} - \Xi$ should be a realization of the noise model assumed for the noise matrix **E**, eqn. (16). Two simple criteria for detecting the number of essentially nonzero singular values of **X** are

(a) visual inspection of the singular value plot such as in Figure 10, and

(b) fidelity of reconstruction of the original data in **X** by use of a reduced number of basis vectors (*principal components*, cf. eqn. (7)).

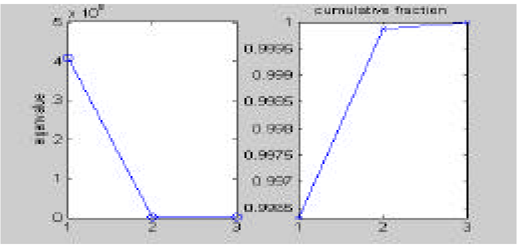


Figure 10 – Squared singular values (eigenvalues) for data in Table 1. (a) individual, (b) cumulative.

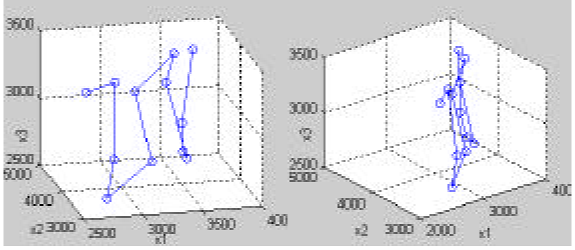


Figure 11 – 2-D world in noisy 3-D data.(cf. Figure 6).

❑ *Singular values quantify the goodness of data fit by a matrix of reduced rank*

If the underlying structure of the data in Ξ is such that only a "small" number of principal components is important, what is the best estimate of Ξ (with rank $r < n$) given the data in **X**?

Answering this question will allow us to construct scores and loadings, and to monitor the system as subsequently etched wafers arrive, in the same way as we did in the noiseless case. The difference is that what should have been ideally zero errors, eqn. (14) should now be "small" (more in the sequel).

To find the best estimate $\hat{\Xi}$ of Ξ given $\mathbf{X}$ we can minimize the distance between Ξ and **X**, i.e. find

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\| \qquad (19)$$

It can be shown [2] that if the matrix norm that appears in eqn. (19) is either the *induced 2-norm* ($\|\mathbf{A}\|_{i2} \hat{=} \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|\mathbf{Av}\|_2}{\|\mathbf{v}\|_2} = s_{\max}$, where $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$ is the Euclidean vector norm) or the *Frobenius norm* ($\|\mathbf{A}\|_F \hat{=} \sqrt{\sum_i \sum_j a_{ij}^2} = \sqrt{\sum_i s_i^2}$)[2] the minimization in eqn. (19) produces the same $\hat{\Xi}$ in terms of the SVD of **X** as

$$\hat{\Xi} = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \qquad (20)$$

(cf. eqn. (5)). Moreover, the optimal difference can be shown to be

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\|_{i2} = \|\mathbf{X}-\hat{\Xi}\|_{i2} = s_{r+1} \qquad (21)$$

and

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\|_{F} = \|\mathbf{X}-\hat{\Xi}\|_{F} = \sqrt{\sum_{i=r+1}^{n} s_{r+i}^2} \qquad (22)$$

Note that the singular vectors (loadings) of **X** could be very different from the singular vectors (loadings) of Ξ [4]. Figure 12, shows loadings for **X**. Comparison with Figure 4 and Figure 7 shows little difference.

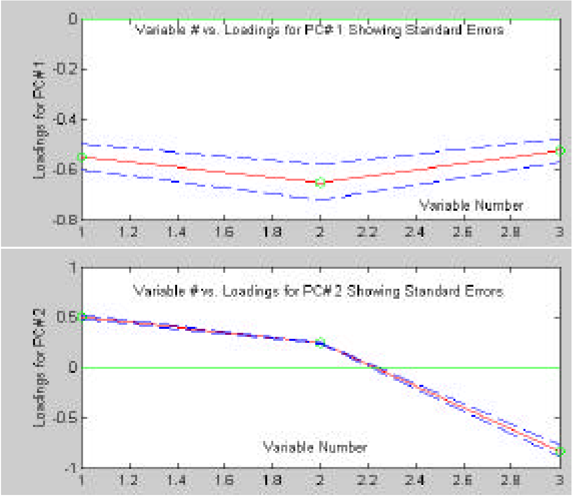


Figure 12 – Loadings(with error bounds) for noisy data of Table 1 (by PLS-toolbox® [3]*) (cf. Figure 4 and Figure 7).*

To test whether the fit is generalizable, tests must be done, such as a test on the PRESS (Prediction Error Sum of Squares) statistic: One datum is left out, a model that fits the remaining data is computed, and the square error between the model prediction and the datum left out is computed. The process is repeated by considering all data points, one by one, and finally summing the square errors and comparing them to the total square error.

❑ *Process monitoring by looking at residual errors (cf. p. 3)*

Once the relationship among x_1, x_2, x_3 has been identified by the counterpart of eqn. (8) with noisy loading $\mathbf{v}_3$, the value of the *residual error* (i.e. counterpart of eqn. (14) for noisy loadings) for each new data point (x_1, x_2, x_3) arriving in the future can be checked. If the relationship among x_1, x_2, x_3 remains the same, then the residual error should be "small". This leads to the counterpart of eqn. (15) for noisy data. Specifically, if the residual error is normally distributed (very often a reasonable assumption) then $\|\mathbf{e}\|^2 = \mathbf{e}^T\mathbf{e}$ follows a chi-square distribution, from which one can construct Q-confidence bounds [3] as (cf. eqn. (15))

$$\mathbf{e}^T\mathbf{e} = \mathbf{x}^T(\mathbf{I}-\mathbf{PP}^T)\mathbf{x} < d^2 \qquad (23)$$

2.3 Stochastic signals

❑ *For multiple random variables* principal components *are uncorrelated new variables, a few of which capture most variance*

SVD can provide additional insight if the vector variable $\mathbf{x}$ is stochastic, i.e. it takes values according to a certain probability distribution (the particular distribution is not important). The analysis is known as *principal component analysis* (PCA) [5].

Consider the random variable vector $\mathbf{x} \hat{=} [x_1 \cdots x_n]^T$, and assume, without loss of generality, that $E[\mathbf{x}] = \mathbf{0}$ [3] where E denotes

[2] Both the induced 2-norm and the Frobenius norm are frequently called Euclidean norm in literature! The reason is that the induced-2 norm is induced by a Euclidean vector norm, and the Frobenius norm would be a Euclidean vector norm if the matrix **A** were re-organized as a vector.

[3] If the average of $\mathbf{x}$ is not zero, a new deviation variable can trivially be defined as $\mathbf{x}-E[\mathbf{x}]$. There is much higher chance that deviation variables (as opposed to original variables) are linearly dependent. Indeed, if the variables $\mathbf{x}$ satisfy the relationship $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, Taylor series expansion around $E[\mathbf{x}]$ yields

$$\mathbf{0} = \mathbf{f}(\mathbf{x}) \approx \underbrace{\mathbf{f}(E[\mathbf{x}])}_{\approx\mathbf{0}} + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}=E[\mathbf{x}]} (\mathbf{x}-E[\mathbf{x}]) \hat{=} \mathbf{B}\cdot\Delta\mathbf{x}$$

which implies linearly dependent $\Delta\mathbf{x}$.

expected value. Denote the covariance matrix of $\mathbf{x}$ by

$$\mathbf{C} = E[\mathbf{x}\mathbf{x}^T] \in \Re^{n\times n} \qquad (24)$$

It can be shown [5] that we can use the modal matrix $\mathbf{A} \hat{=} [\mathbf{a}_1 \vdots \cdots \vdots \mathbf{a}_n]$ of $\mathbf{C}$ (i.e. the matrix whose columns are the orthonormal eigenvectors of $\mathbf{C}$) to construct a new, zero-mean, vector random variable $\mathbf{y}$ as

$$\overset{1}{\boxed{\mathbf{y}}}_n = \overset{n}{\boxed{\mathbf{A}^T}}_n \overset{1}{\boxed{\mathbf{x}}}_n \Leftrightarrow \overset{1}{\boxed{\mathbf{x}}}_n = \overset{n}{\boxed{\mathbf{A}}}_n \overset{1}{\boxed{\mathbf{y}}}_n \qquad (25)$$

(**principal components**) that has the following important property

$$\operatorname{var}(y_i) = \max_{\substack{\|a_i\|_2 = 1 \\ E[y_i y_{j<i}]=0}} \operatorname{var}\left(a_i^T \mathbf{x}\right) = l_i , \qquad (26)$$

That is, each principal component, y_i is a weighted sum of the original variables $x_1, \ldots, x_n$, (eqn. (25)) such that

(a) its variance is maximal and equal to the i-th eigenvalue of the original covariance matrix $\mathbf{C}$ (eqn. (26)), and

(b) y_i is orthogonal to all previous principal components y_{i-j}, $i \geq 2$, $j = 1,\ldots, i-1$ (eqn. (26)).

There are various criteria for selecting the number of principal components, as discussed above and in [5].

❑ *SVD on covariance estimate produces values of principal components*

Because the matrix $\mathbf{C}$ is unknown, it has to be estimated from data. The best estimate of $\mathbf{C}$ is

$$\overset{n}{\boxed{\mathbf{C}}}_n \approx \frac{1}{m-1} \overset{m}{\boxed{\mathbf{X}^T}}_n \overset{n}{\boxed{\mathbf{X}}}_m \qquad (27)$$

where $\mathbf{X}$ is a matrix that contains the data for each random variable in a column, as in eqn. (1). Then, the eigenvalue/eigenvector pairs $(k, \mathbf{w})$ of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ are estimates of the eigenvalue/eigenvector pairs $(l, \mathbf{a})$ of $\mathbf{C}$, which implies that

(a) the eigenvectors $\mathbf{w}$ of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ (hence the estimates of eigenvectors of $\mathbf{C}$) are equal to the singular vectors $\mathbf{v}$ of $\mathbf{X}$ (eqn. (4)) (i.e. loadings; cf. discussion about the interpretation of loading in p. 2), and

(b) the eigenvalues of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ (hence the estimates of eigenvalues of C) are equal to $(m-1)$ times the squares of the singular values of $\mathbf{X}$

Consequently, one can look at the values of

$$\frac{s_i^2}{s_1^2 + \cdots + s_r^2} = \frac{s_i^2}{E[\mathbf{x}^T\mathbf{x}]} = \frac{l_i}{l_1 + \cdots + l_r} \quad i = 1, \cdots, r \qquad (28)$$

to assess what percentage of the total variance of $\mathbf{x}$, $E[\mathbf{x}^T\mathbf{x}]$, is captured by each of the principal components. By looking at the few largest principal components, one can monitor (in the SPC sense) the system that produces the data

(a) visually, e.g., by plotting PC1 vs. wafer #, PC2 vs. wafer #, etc. or PC1 vs. PC2 vs. PC3 (recall that principal components are ideally independent of one another).

(b) numerically, by monitoring statistics such as the Hotelling statistic discussed below.

❑ *Principal components are directly related to multivariate SPC*

If the zero-mean vector random variable $\mathbf{x}$ has (non-degenerate) covariance $\mathbf{C}$, then one can construct the Hotelling (scalar) random variable

$$\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} = \underbrace{\mathbf{x}^T \mathbf{A}}_{\mathbf{y}^T} \Lambda^{-1} \underbrace{\mathbf{A}^T \mathbf{x}}_{\mathbf{y}} \hat{=} \mathbf{y}^T \Lambda^{-1} \mathbf{y} = \sum_{i=1}^{n} \frac{y_i^2}{l_i} \qquad (29)$$

i.e. the Hotelling random variable is the sum of n independent random variables, y_i^2 / l_i. If y_i are normally distributed, then $\sum\left(y_i^2 / l_i\right)$ is chi-square distributed. We stress that the matrix $\mathbf{C}$, as stated above, is assumed to be non-degenerate, so that all eigenvalues of $\mathbf{C}$ are non-zero and $\mathbf{C}^{-1}$ exists in eqn. (29). If some eigenvalues are zero, then we stop the summation in eqn. (29) at r, the rank of $\mathbf{C}$, to ensure $l_i \neq 0$.

Note that the new variable vector $\mathbf{y}$ defined in eqn. (29) is precisely the vector of principal components, as defined in eqn. (25). Therefore, when using PCA to monitor random variables, one can use the Hotelling T^2 statistic to perform a multivariate chi-square test [3]

$$\mathbf{x}^T \mathbf{P} \Lambda_r^{-1} \mathbf{P}^T \mathbf{x} = \sum_{i=1}^{r} \frac{y_i^2}{l_i} \leq c^2 \qquad (30)$$

where the matrix $\mathbf{P}$ consists of as many columns of $\mathbf{V}$ as the number of principal components retained (cf. eqn. (15)). The values of the principal components y_i for which $\sum_{i=1}^{r} y_i^2 \big/ l_i \leq c^2$ are inside an r-dimensional ellipsoid with axes $l_i c^2$. If up to 3 principal components are retained, then one can plot these ellipsoidal bounds and visually observe whether subsequent values of the principal components fall inside the ellipsoid, for multivariate SPC. An example will be shown with the actual wafer data in the sequel.

❑ *Lumping apples and oranges as "fruits" is OK but should be done with caution*

The variables x_1, x_2, x_3 in the preceding example all refer to etch rate. Therefore it is natural to express their values in the same units. It is possible, however, to consider sets of random variables of different nature, e.g., etch rate, power, pressure, flowrate, etc. In that case, the units of measurement (scaling) become important when performing PCA for these variables, in that different scalings can produce arbitrarily different eigenvalues of the covariance matrix $\mathbf{C}$.

One straightforward way to avoid this ambiguity is to perform PCA on the correlation matrix, instead of on the covariance matrix. But it has to be stressed that the correlation matrix weights each variable according to its variance, whether the latter is large or small.

Another criterion for obtaining meaningful PCA results when dealing with variables for which there is experimental measurement error is to ensure that all errors are independent and of the same magnitude (cf. eqn. (19)).

3 CASE STUDY

Etch profiles (49 measurement points $x_1,\ldots,x_{49}$) from 9 different tools were collected, thus creating a 9×49 matrix $\mathbf{X}$. Figure 13 indicates that 2 or 3 principal components result in less that 10% or 5% error, respectively. Corresponding scores are shown in Figure 14. Loadings are shown as weights in Figure 15 and as basis surfaces in Figure 16. The quality of reconstruction of the original data by 3 principal components is excellent, in that it captures curvature characteristics, as indicated by the samples shown in Figure 17. Figure 18 shows the Q-test (eqn. (23)) and Hotelling T^2-test (eqn. (30)), revealing no outliers. These tests can be used to monitor future wafers, i.e. if future points fall within the bands indicated in Figure 18, then future wafers are etched "similarly" to those contained in the original set.

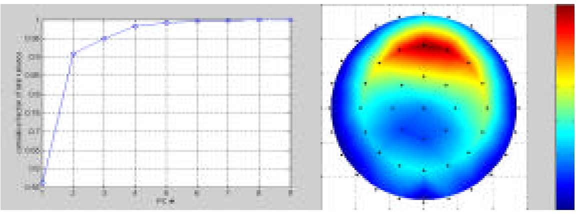


Figure 13 – Cumulative fraction of total variance captured by principal components (left) for variables $x_1,\ldots,x_{49}$ *scaled by subtraction of sample averages* $\bar{x}_1,\ldots,\bar{x}_{49}$ *(right).*

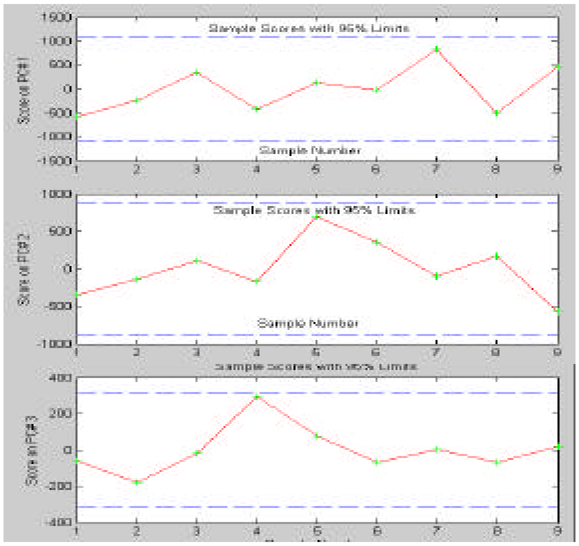


Figure 14 – Scores for the first 3 principal components (cf. Figure 5). (Confidence bounds by PLS-toolbox® [3]*.)*

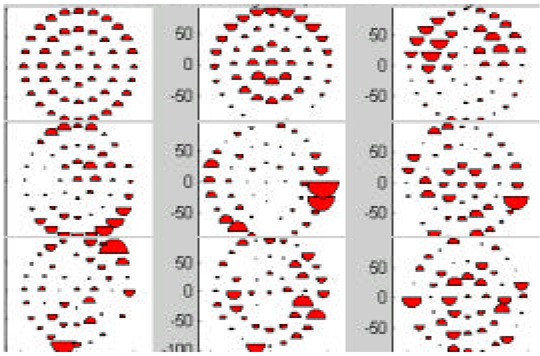
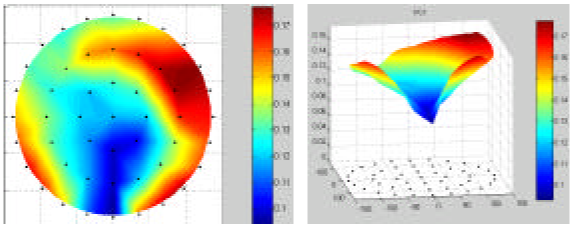


Figure 15 – Loadings as weighting coefficients (cf. Figure 7) for all 9 principal components. Semidisk size and orientation denote magnitude and sign, respectively.



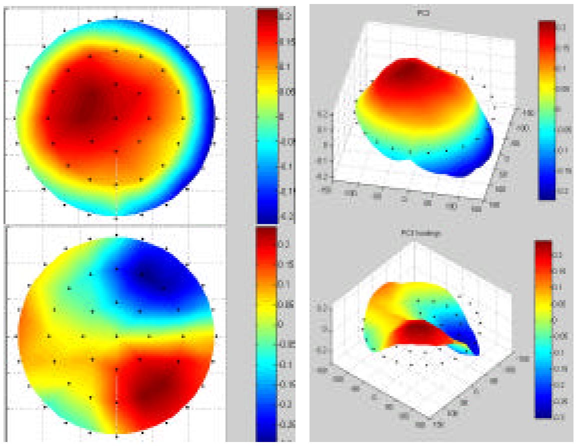


Figure 16 – Loadings as contour surfaces (cf. Figure 4) for the first 3 principal components. Each loading is viewed from the top and from an angle.

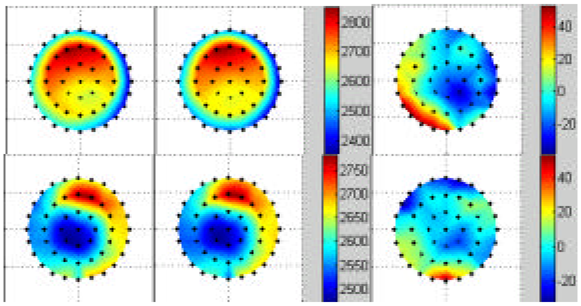


Figure 17 – Original etch profile (column 1), etch profile reconstructed from 3 principal components (column 2) and approximation error (column 3) for two sample wafers (cf. Figure 1)

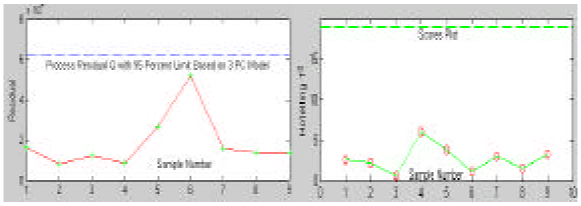


Figure 18 – Residual square errors and Q-test (cf. eqn. (23)) and values of the Hotelling statistic and T^2*-test (cf. eqn. (30))*

Acknowledgement – The first author gratefully acknowledges support for this work from the National Science Foundation and Lam Research Corporation through a GOALI grant.

4 REFERENCES

[1] Horn, RA, and CR Johnson, *Matrix Analysis*, Cambridge University Press (1985).

[2] Dewilde, P., and Ed. F. Deprettere, "Singular Value Decomposition: An Introduction," in *SVD and signal processing: algorithms, applications, and architectures*, edited by Ed. F. Deprettere, North-Holland, 3-41. (1988).

[3] http://www.eigenvector.com/.

[4] Stewart, G., "Perturbation Theory for the Singular Value Decomposition", in *SVD and Signal Processing, II*, R. J. Vacarro ed., Elsevier, Amsterdam (1991).

[5] Jolliffe, IT, *Principal Component Analysis*, Springer-Verlag (1986).