Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Fisher Discriminant Analysis for Semiconductor Batch Tool Matching

Gregory Cherry * and **S. Joe Qin** **,[1]

* *Advanced Micro Devices, Inc., Austin, TX 78741 USA*
** *Departments of Chemical Engineering and Material Sciences, Electrical Engineering, and Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, 90089, USA*

Abstract: We propose in this paper how Fisher discriminant analysis can be applied for differentiating classes of semiconductor data from different tools or chambers. The tool and chamber matching analysis can be useful not only from a process characterization standpoint, but also for identifying the proper fault detection and classification strategy. If the FDA analysis shows that the chambers or tools are not matched, it is likely that either preventive maintenance will be needed, or a separate fault detection and diagnosis model will be required for each tool or chamber. Alternatively, if the match fraction is large enough, a global model should be considered for monitoring a plurality of tools or chambers to reduce the model maintenance effort. Contribution analysis is also applied to identify variables responsible for tool mismatching.

Keywords: process monitoring; batch process matching; Fisher discriminant analysis; semiconductor manufacturing

1. INTRODUCTION

An important characteristic of semiconductor fabrication is the rather significant amount of multiplicity present in both the devices being produced and the machines that are used to produce them. For example, multiple products are manufactured in the same facility, and the product mix is always changing as new technologies are introduced. Wafers for each product undergo multiple operations to create devices that meet certain electrical specifications. At any given operation, multiple processing tools are often available in order to meet productivity requirements based on customer demand. Additionally, many of these tools contain multiple processing chambers that have their own independent subsystems for processing individual wafers.

During the manufacturing cycle, all of these sources of multiplicity create obstacles to the fabrication of a consistent product from lot to lot and wafer to wafer. While multivariate analysis with principal component analysis (PCA) has been applied to provide early warnings when product or process information no longer agrees with historical behavior, the multiplicity present in the factory creates challenges related to the creation and maintenance of the multivariate analysis models. In the most extreme case where a unique model would be needed for every product in the fab, every critical operation in the line, every tool at each operation, and every chamber in those tools, the model maintenance effort required would be overwhelming. In response to these challenges, we present in this paper how Fisher discriminant analysis (FDA) can be used to characterize the differences between data from multiple factory contexts and alleviate the risks of model adaptation. The results will identify tools that are not adequately matched, and indicate when the multiplicity problem can be reduced by sharing model parameters across contexts.

Previous work has been published in which PCA scores have been used for characterizing differences between processing chambers, Skumanich et al. [2004], Smith et al. [2004]. However, we recommend FDA analysis in this work because it has the objective of finding the optimal directions to discriminate between the different classes of data. In contrast, PCA generates a model by identifying the main sources of variation, taking the entire data population as a whole regardless of the class contexts. While a review of the PCA scores may facilitate discrimination when the class groupings are not known a priori, Krzanowski [1979], FDA will be more reliable for discrimination.

The organization of this paper is as follows. The next section presents the basics of the FDA algorithm and provides an illustrative example of its use. Section 3 then provides case study simulation results for etcher chamber matching and rapid thermal anneal chamber matching. The last section gives conclusions.

2. FDA ALGORITHM

As described in the next few subsections, the FDA algorithm as used for comparing multiple classes of data is broken down into three steps. The first step is to define the classes that are to be compared with one another and to characterize the multivariate distributions of each of those classes. Next, the optimal directions for discriminating

[1] Corresponding Author: sqin@usc.edu.
Supported by AMD through sponsorship of the Texas-Wisconsin-California Control Consortium.

between the classes are calculated, thus establishing the Fisher discriminant rule. Finally, the maximum likelihood rule is applied to create a single metric that can be used to assess the similarities and differences between the classes.

2.1 Class Definition and Distributions

Semiconductor manufacturing facilities make multiple products on multiple tools. There are also many tools that have more than one chamber and adaptive models may be required to track the normal drift of certain processes. In the algorithm that follows, the term 'class' is used to distinguish between each of these subgroups. Each class would then be described by its own unique set of statistics that describe its multivariate distribution.

Data that follow the multivariate normal distribution can be described by the variable means and covariance matrix, or equivalently by the variable means, variable standard deviations, and correlation matrix. These quantities are required in order to build a principal component analysis model that can then be used for process monitoring using Hotelling's T^2, the squared prediction error (SPE), or the combined index (φ), each with their respective control limits, Jackson [1991], Nomikos and MacGregor [1995], Yue and Qin [2001].

If a set of tools perform a certain operation, then each tool could be considered an individual class. Let class k be used to describe one of those tools. A raw data matrix gathered from historical operation of that tool would be given as $\mathbf{X}_k^0 \in \mathbb{R}^{n_k,\times m}$, with n_k samples and m variables. The variable means, $\mathbf{b}_k$, and variable standard deviations, σ_k, would be used to scale the data matrix to zero-mean and unit variance, thus giving the normalized matrix $\mathbf{X}_k$. The correlation matrix would then be calculated by

$$\mathbf{R}_k = \frac{1}{n_k - 1}\mathbf{X}_k^T\mathbf{X}_k. \tag{1}$$

Thus, the multivariate distribution for tool k is parameterized by $\mathbf{b}_k$, σ_k, and $\mathbf{R}_k$. The subsequent application of FDA will therefore indicate the level of similarity between all tools in the set.

Figure 1 graphically illustrates this and the remaining steps in the analysis procedure. For this example data set, data are artificially created for two variables and two classes. The leftmost plot shows a scatter plot with •'s for data from class 1 and ∘'s for data from class 2. The means for classes 1 and 2 are then plotted with the ♦ and ■, respectively. The standard deviations and correlation matrices, which complete the parametrization of the distributions, are then used to construct the elliptical regions that are centered about the means for each class. In the figure, these ellipses are based on a 99% confidence limit and are denoted with dotted and dash-dotted lines for classes 1 and 2, respectively. The subplot on the right and the remaining four elements of the leftmost plot will be described in the subsections that follow.

2.2 Fisher Directions and the FDA Rule

Once the means, standard deviations, and correlation matrices have been collected for a set of c chamber classes or adaptive model classes, Fisher discriminant analysis can then be applied, Chiang et al. [2000], Hardle and Simar [2003]. Weighting all classes equally, the overall mean vector for the entire population is calculated by

$$\bar{\mathbf{b}} = \sum_{k=1}^{c} \frac{\mathbf{b}_k}{c}. \tag{2}$$

The overall mean can alternatively be defined by taking into account the number of samples in each class, but in this work we assume that all classes make an equivalent contribution. For our example displayed in Figure 1, the overall mean is depicted with a ★ situated directly between the two class means.

The within-class scatter matrix for class k, which is equivalent to the covariance of that class, is then determined by

$$S_k(i,j) = R_k(i,j)\sigma_k(i)\sigma_k(j) \qquad \text{for all } i,j \tag{3}$$

where $S_k(i,j)$ denotes row i and column j of $\mathbf{S}_k$.

Summing up all of the within-class scatter matrices for all classes then gives the overall within-class scatter matrix,

$$\mathbf{S}_w = \sum_{k=1}^{c} \mathbf{S}_k, \tag{4}$$

which quantifies the typical correlation between the variables when considering the classes separately. Conversely, the between-class scatter matrix provides a measure of how the classes relate to each other as

$$\mathbf{S}_b = \sum_{k=1}^{c} (\mathbf{b}_k - \bar{\mathbf{b}})(\mathbf{b}_k - \bar{\mathbf{b}})^T. \tag{5}$$

The within-class and between-class scatter matrices are respectively depicted with the solid ellipse and solid line in the leftmost plot in Figure 1. Because only two classes of data are being compared, the between-class scatter matrix is depicted by a line, whereas an additional class of data would be represented by an ellipse.

The next step in the procedure is to find the Fisher optimal discriminant directions, which are based on maximizing the ratio of the between-class scatter to the within-class scatter,

$$J(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{S}_b\mathbf{a}}{\mathbf{a}^T\mathbf{S}_w\mathbf{a}}. \tag{6}$$

This is equivalent to the following generalized eigenvalue problem that is solved for the Fisher directions using

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{a} = \lambda\mathbf{a}. \tag{7}$$

Several Fisher directions corresponding to the significant generalized eigenvectors can be used to separate the classes by projecting the original data for a sample, $\mathbf{x}^0$, into the Fisher space using the linear transformation

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}^0. \tag{8}$$

where columns of $\mathbf{A}$ are composed of the generalized eigenvectors. It follows that the means and covariances of the transformed data are represented by

$$\bar{\mathbf{y}}_k = \mathbf{A}^T\mathbf{b}_k \tag{9}$$

and

$$\mathbf{\Omega}_k = \mathbf{A}^T\mathbf{S}_k\mathbf{A}, \tag{10}$$

respectively. The transformed data from our example problem are represented in the plot on the right in Figure 1. As shown, the original bivariate data have been transformed to a single variable in the Fisher space, with

the probability densities generated from the means and variances from Equations (9) and (10). The two classes are clearly separable, with only a small region of overlap between them that is shaded in the figure.

Given a new data point, $\mathbf{y}$, the Fisher discriminant rule classifies it into class j according to

$$j = \arg\min_k \mid \mathbf{y} - \bar{\mathbf{y}}_k \mid, \tag{11}$$

which is equivalent to finding the class mean in the Fisher space that has the smallest Mahalanobis distance from the sample of interest. This rule is depicted by the vertical line with •'s at its endpoints in the plot on the right in Figure 1. The Fisher discriminant rule would classify samples projected to the left of this line into class 1, while samples projected to the right would be classified into class 2.

2.3 Maximum Likelihood Rule and Class Overlap

In contrast to the Fisher discriminant rule, the maximum likelihood (ML) rule classifies according to

$$j = \arg\max_k \mid f_k(\mathbf{y}) \mid, \tag{12}$$

which identifies the class with the largest probability density for the point in the Fisher space, Anderson [2003]. In (12), the probability density function follows a multivariate normal distribution as

$$f_k(\mathbf{y}) = (2\pi)^{-\frac{1}{2}(c-1)} \mid \mathbf{\Omega}_k \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}}_k)^T \mathbf{\Omega}_k^{-1} (\mathbf{y}-\bar{\mathbf{y}}_k)}. \tag{13}$$

For our example problem, the maximum likelihood rule is denoted by the vertical line with ×'s at its endpoints. It is only slightly to the left of the Fisher discriminant rule, which can be accounted for by the slight differences in variance between the two distributions. If the covariances of the classes after transformation are equivalent, then the Fisher discriminant rule will provide an equivalent classification boundary as the maximum likelihood rule.

While the Fisher discriminant and maximum likelihood rules could be useful for classifying samples, that is not our final objective. Instead, we proceed to calculate the actual error rate to define a single measure for comparing the different classes in more general terms, Hardle and Simar [2003]. After defining a classifier function using the maximum likelihood rule as

$$\phi_j(\mathbf{y}) = \begin{cases} 1 \text{ if } j = \arg\max_k \mid f_k(\mathbf{y}) \mid \\ 0 \text{ otherwise} \end{cases}, \tag{14}$$

the probability of classifying $\mathbf{y}$ into class i even though it belongs in class j is determined by

$$p_{ij} = \int \phi_i(\mathbf{y}) f_j(\mathbf{y}) d\mathbf{y}. \tag{15}$$

The actual error rate, which is defined by

$$AER = \frac{\sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} p_{ij}}{c}, \tag{16}$$

then gives a measure of the 'overlap' between the classes. If the classes are completely separable, the AER will be close to zero. On the other hand, classes that are indistinguishable will have an AER that approaches a maximum of $\frac{c-1}{c}$. By dividing by the maximum, we define a match fraction as

$$MF = \frac{c}{c-1} AER, \tag{17}$$

which takes a value between 0 and 1 and can be used to assess the overall overlap among the classes.

3. INDUSTRIAL CASE STUDIES

In the case studies that follow, the method of He et al. [2005] will be employed. As previously published, the method consists of two steps. The first step is to apply an overall FDA model to visualize all class clusters in the Fisher space, which can be useful for assessing whether or not a fault has occurred. Next, that analysis is followed by additional FDA for every individual pair of classes, which can be used for fault diagnosis. Although the same techniques will be implemented in this paper, the objective is not to identify and diagnose faults. Instead, the objective of this work can be more generally stated as the quantification of the differences between the classes and the identification of the variables that contribute most to those differences.

3.1 Plasma Etch Chamber Matching

To demonstrate how the techniques of Section 2 can be applied to identify differences between tools and chambers, we consider a set of three dual-chamber etch tools. The reasoning behind tool matching comes from the expectation that consistency between processing tools will facilitate the fabrication of a more consistent product. For this study the means, standard deviations, and a correlation matrix were calculated from 16 sensor summary statistics for each of the six chambers for a single recipe. The summary statistics were calculated from two main processing steps for the sensors that were deemed important. The first nine summary statistics were calculated from one step, while the remaining six were calculated from another.

Application of FDA simultaneously to all classes of the etcher data set yielded the results shown in Figure 2. In addition to the data samples that have been projected onto the first two Fisher directions, elliptical regions defined using Equations 9 and 10 are also displayed in the figure using 99% confidence thresholds. The plot makes it very easy to distinguish between all of the chambers using FDA, which was also evident by a calculated match fraction of 0%. But even though all chambers are distinct, it is interesting to observe that for all three tools, the chambers from the same tools are adjacent to each other in the plot, which indicates at least some degree of similarity between chambers on the same tool.

Figure 3 shows the results of applying FDA pair-wise to a few of the chambers. The axes in the plots on the left show the probability distributions of each of the classes in the Fisher space, which can be used to asses how similar each set of data is to the other. The FDA and ML rules are also provided in the probability plots. It is observed that when the distributions have similar variances, the FDA and ML rules are closely matched. However, when the variances differ, the FDA and ML rules are shifted away from each other.

The right set of plots in the figure provides bar charts of the Fisher direction, which indicates the main contributing variables that discriminate between the two classes, He et al. [2005]. Of the 16 variables included, either one

or both of variable 7 and variable 9 were the main differentiators between all chambers.

This etcher case study showed how the FDA algorithm was able to quickly indicate that the sensor summary statistic data collected were clearly not matched from chamber to chamber. Based on a final evaluation of these differences, the chamber differences were judged to be irrelevant with respect to product performance, and no chamber matching would be required. Instead, the models deployed for fault detection and diagnosis would need to be maintained independently for each chamber to provide adequate sensitivity to future excursions.

3.2 Rapid Thermal Anneal Chamber Matching

The second case study for the FDA algorithm has the objective of matching chambers from a set of dual-chamber rapid thermal anneal (RTA) tools. As with the etcher study, data from three tools were available, so a total of six chambers will be compared against each other. However, in contrast to the etcher study, the variables that make up the data matrices are not summary statistics. Instead, interpolation was used to construct multi-way data matrices by extracting 10 samples from a single processing step for 8 sensors, yielding a total of 80 variables after unfolding. Preprocessing was performed with the shape-preserving piecewise cubic spline interpolation method, Mathworks [2005], de Boor [1978].

Figure 4 displays the results of applying FDA to all six chambers. In the first plot in the figure, the sample points are overlayed with the 99% confidence limits, while in the second plot, only the class means are overlayed with the limits. Of the six chambers, 1A and 1B were found to be completely separable from all other chambers, while 3B was somewhat separable. Alternatively, chambers 2A, 2B, and 3A were relatively well-matched for the variables captured from the tools. The overall match fraction was 31%.

To further investigate the similarities and differences between the RTA chambers, the pair-wise FDA match fractions are summarized in Table 1. As displayed in the table, we again observe that chambers 1A and 1B were completely separable, with pair-wise match fractions of 0% when compared against all other chambers. Chambers 2A, 2B, and 3A exhibited match fractions ranging from 46% to 59% when compared against each other. Finally, chamber 3B matched 2A, 2B, and 3A with match fractions of 6%, 9%, and 8%, respectively.

Table 1. Summary of pair-wise RTA chamber overlap

Chambers	1B	2A	2B	3A	3B
1A	0	0	0	0	0
1B	-	0	0	0	0
2A	-	-	0.53	0.59	0.06
2B	-	-	-	0.46	0.09
3A	-	-	-	-	0.08

An example of two chambers with a high degree of overlap is displayed in Figure 5 for chambers 2A and 3A. With a match fraction of 59%, this is as close as any of the chambers got to each other. Other comparisons between chambers 2A, 2B, and 3A showed similar results. The most significant contributions for the chamber difference was from sensor 6 at the first time index. However, the high degree of overlap indicates that the differences were not enough to clearly distinguish between them.

To conclude the RTA study, Figure 6 plots the chamber means of the raw data for sensor 6 across all 10 time indices in the batch. The data are centered against the recipe setpoint such that a value of zero is on target. This plot clearly shows discrepancies between the values observed for both chambers of tool 1. While there is some variability for the other chambers early on in the batch, they quickly converge to the setpoint at time index 2. In contrast, 1A and 1B are much slower to converge, and are even slightly offset from the setpoint at the very end of the processing step.

4. CONCLUSIONS

We demonstrate in this paper how Fisher discriminant analysis can be applied for differentiating classes of semiconductor data from different tools or chambers. The tool and chamber matching analysis can be useful not only from a process characterization standpoint, but also for identifying the proper fault detection and classification strategy. When applied to historical data from a set of etcher chambers, the algorithm identified discrepancies in sensor data that ultimately did not pose a risk to product performance. For this case, the differences indicated that a separate fault detection and diagnosis model would be required for each chamber. Conversely, the data captured from the RTA chambers highlighted issues with three processing chambers where they were not converging to their setpoints as quickly as they should have. The application of preventive maintenance on the chambers could be used to match them and thus facilitate sharing of a single multivariate analysis models across all of them.

REFERENCES

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc, 2003.

L.H. Chiang, E.L. Russell, and R.D. Braatz. Fault diagnosis and fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics Intell. Lab. Syst.*, 50:243–252, 2000.

C. de Boor. *A Practical Guide to Splines.* Springer-Verlag, 1978.

W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis.* Springer Verlag, 2003.

Q. P. He, S. J. Qin, and J. Wang. A new fault diagnosis method using fault directions in fisher discriminant analysis. *AIChE J.*,, 51(2):555–571, 2005.

J.E. Jackson. *A User's Guide to Principal Components.* Wiley-Interscience, New York, 1991.

W. J. Krzanowski. Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74:703–707, 1979.

The Mathworks. *MATLAB 7 User's Guide.* The Mathworks, Inc., 2005.

P. Nomikos and J.F. MacGregor. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37(1): 41–59, 1995.

A. Skumanich, J. Yamartino, D. Mui, and D. Lymberopoulos. Advanced etch applications using tool-level data. *Solid State Technology*, 47(6):47–52, 2004.

J.A. Smith, K.C. Lin, M. Richter, and U. LevAmi. Practical, real-time multivariate fdc. *Semicond. Int.*, 27(13): 751–56, 2004.

H. Yue and S. Joe Qin. Reconstruction based fault identification using a combined index. *Ind. Eng. Chem. Res.*, 40:4403–4414, 2001.

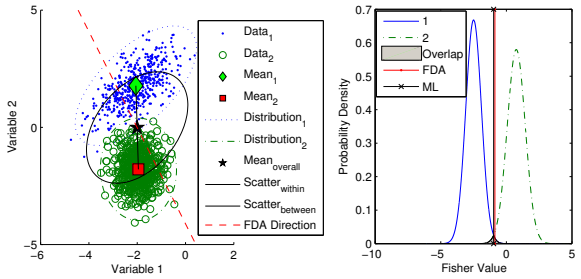

Fig. 1. FDA as applied to a bivariate data set with two classes.

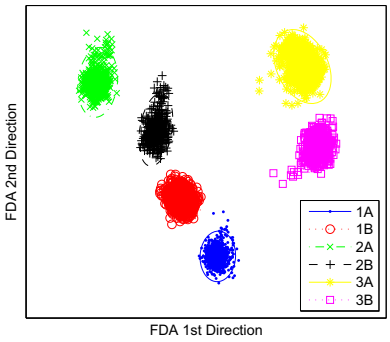

Fig. 2. Scatter plot of six etch chambers after projection onto the two most significant Fisher directions.

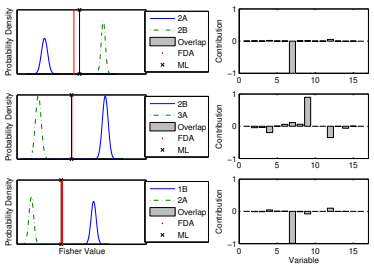

Fig. 3. Probability densities of etcher data projected into the Fisher space and the variable contributions to the FDA directions.

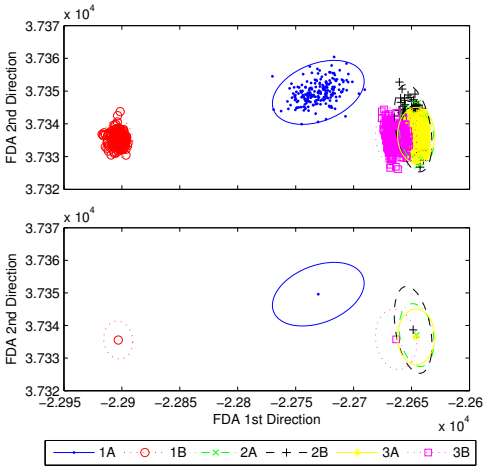

Fig. 4. Scatter plots of six RTA chambers after projection onto the two most significant Fisher directions.

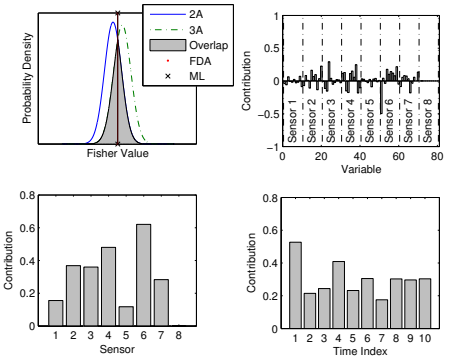

Fig. 5. Probability densities of RTA data for Chambers 2A and 3A and group contributions to the largest discriminant direction.

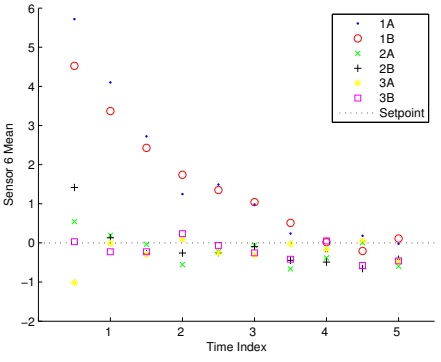

Fig. 6. Sensor 6 raw data means for the RTA chambers