Pattern Recognition in Large-Scale Data Sets: Application in Integrated Circuit Manufacturing

Choudur K. Lakshminarayan[1], Michael I. Baron[2]

[1]Hewlett Packard Research, USA
{Choudur.Lakshminarayan}@iith.ac.in
[2]University of Texas at Dallas, USA
mbaron@utdallas.edu

Abstract. It is important in IC manufacturing to identify probable root causes, given a signature. The signature is a vector of electrical test parameters measured on process control bars on a wafer. Linear discriminant analysis and artificial neural networks are used to classify a signature of test electrical measurements of a failed chip to one of several pre-assigned root cause categories. An optimal decision rule that assigns a new incoming signature of a failed chip to a particular root cause category is employed such that the probability of misclassification is minimized. The problem of classifying patterns with missing data, outliers, collinearity, and non-normality are also addressed. The selected similarity metric in linear discriminant analysis, and the network topology, used in neural networks, result in a small number of misclassifications. An alternative classification scheme is based on the locations of failed chips on a wafer and their spatial dependence. In this case, we model the joint distribution of chips by a Markov random field, estimate its canonical parameters and use them as inputs for the artificial neural network that also classifies the patterns by matching them to the probable root causes.

Keywords. Integrated-Circuit, Failure Analysis, Signature Analysis, Pattern Recognition, Linear Discriminant Analysis, Mahalanobis Distance, Neural Networks, Back Propagation, Markov Random Fields and spatial signatures.

1 Introduction

SIGNATURE analysis (SA) is a statistical pattern recognition program designed to assign failed parts to one of several pre-determined root cause categories. Engineers invest lots of time tracing back test probe/electrical parameter failures to probable root causes. It is desired to have an automated program based on sound statistical theory that enables the classification of a failing signature to a root cause category such that the probability of misclassification is minimized. Linear discriminant analysis (LDA) is a well established parametric procedure that minimizes the probability of misclassification and allows the failure analysis engineer to state "The probability that a failing wafer with a specific signature belongs to the kth root cause category is p %."

Signature analysis is intended as an aid and not a replacement for sound engineering analysis and judgment. It is not meant for situations where the root cause is quite obvious, for it does not contribute anything that the engineer does not already know. It is really intended for the more subtle situation where the root cause of the failure is not readily apparent and so could throw light on the "probable" root cause.

Instead of, or in addition to the signatures, one can classify wafers and assign them to known root causes based on the locations of defective chips on a wafer. Analysis of production wafers indicates that patterns of failing clusters, their direction, size, and shape differ from one root cause to another, hence they can be used as inputs to the classification scheme. Information on spatial location of failed chips is summarized in ten parameters of a suitable Markov random field and used as input layer node elements for an artificial neural network to classify wafer patterns into root cause categories.

This paper is organized in the following manner. In section 2 we introduce linear discriminant analysis as a methodology for classification of patterns to root causes, decision rules that maximize the probability of correct classification, and the assumptions required for its implementation. Section 3 deals with issues related to missing data, outliers, collinearity within the features of an input pattern, and non-normality. In section 4 we present the artificial neural network approach to pattern classification. In section 5 we present some examples of the implementation of linear discriminant analysis and artificial neural networks in IC failure analysis. Finally, in section 7 we state the conclusions derived from our work.

2 Linear Discriminant Analysis and Mahalanobis Distance

Assume there are k root cause categories, and let j denote the j^{th} category where $j = 1,2,3,\ldots,k$. Let p be the number of electrical parameter measurements defining each signature denoted by $\boldsymbol{z}$; thus a signature can be expressed as a $1 \times p$ row vector. Let n_k be the number of signatures in the k^{th} root cause category. Let $\boldsymbol{\mu_k}$ denote the mean vector (centroid) of dimension $1 \times p$. The elements of $\boldsymbol{\mu_k}$ are the averages of the p test parameters defining the signature. Let $\boldsymbol{\Sigma}$ be the common covariance matrix, whose elements are variances and covariances of the p electrical test parameters. It is also assumed that the signature constitutes a sample random vector from a *multivariate normal distribution*. A multivariate normal distribution is a p-dimensional extension of the one variable Gaussian distribution. The decision rule that maximizes the probability of correct classification under the assumption of multivariate normality of an incoming failing signature $\boldsymbol{z} = (z_1, z_2,\ldots,z_p)$ into root cause category j is given by $argmax_j\ d_j(x)$, where

$$d_j(x) = \ln(p_j) - (1/2)\,(\boldsymbol{z}-\boldsymbol{\mu_j})^T \Sigma^{-1} (\boldsymbol{z}-\boldsymbol{\mu_j}), \quad j = 1,2,\ldots\ldots,k.$$

The second term is called the *Mahalanobis Distance statistic* denoted by D_j, and p_j in the first term is the prior probability of class C_j. It is quite common to assume equal prior probabilities for the k root cause categories. Unequal prior probabilities

may be assigned to root cause categories based on experience and engineering knowledge of semiconductor failure analysis. Such knowledge was not easily available and we also feared any unequal probability assignment may bias the classifications and hence the assumption of equal probabilities in our modeling. The new signature vector $\boldsymbol{z}$ is classified into category i if $d_j(x) > d_i(x)$ for $j \neq i$. Equivalently, vector $\boldsymbol{z}$ is classified into category C_j for which D_j is minimum. In order for the inverse of the covariance matrix $\boldsymbol{\Sigma}$ to exist, it must be of full rank p; thus there must be at least as many signatures in each root cause category as there are test probe parameters ($p \leq n_j$). In practice, $\boldsymbol{\mu_j}$ and $\boldsymbol{\Sigma}$ are unknown, and are estimated from the sample. The sample estimates of the mean vector and the covariance matrix are given as $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$. The constituents of the sample mean vector and the sample covariance matrix are sample analogs of their population counterparts. The application of linear discriminant analysis requires the assumption of *equicovariance* among the k root cause categories. Equicovariance property requires equal covariance structure across the k root cause categories. Combining the sample covariance matrices results in a pooled covariance matrix $\boldsymbol{S}_p$. Mathematically the pooled covariance matrix is given as

$$\boldsymbol{S}_p^{-1} = \frac{(n_1-1)s_1+.....+(n_k-1)s_k}{n_1+...+n_k-k},$$

which is a weighted average of the individual covariances of the k root cause categories.

An estimate of the Mahalanobis distance metric is given by:

$$\hat{D}_j = \left(\boldsymbol{z}-\boldsymbol{x}_j\right)^T \boldsymbol{S}_p^{-1}\left(\boldsymbol{z}-\boldsymbol{x}_j\right)$$

where for a fixed value of p, the estimated Mahalanobis distance is asymptotically distributed as a chi- square with p degrees of freedom as each n_j ,$j=1,2,….,k$, becomes large.

It can be shown that classification by LDA minimizes the total probability of misclassification. We observe *en passant* that classification by LDA is tantamount to maximizing the posterior probability of a root cause C_j given a signature with respect to a zero-one loss function. See [1, 2] for a detailed discussion of discriminant rules and Bayes classifiers.

There are no definite guidelines for the number of sample signatures in each root cause category, but it is recommended that each n_j be at least five times larger than p. The asymptotic chi-square property of the Mahalanobis distance statistic allows one to associate a probability of misclassification or its complement. A new incoming failing pattern $\boldsymbol{z}$ will be classified in category j if $\hat{D}_j$ is the minimum over all j=1,2,….,k. The Mahalanobis distance alone is used for the relative ranking among the root cause categories. A small value of $\hat{D}_k$ in conjunction with a large probability of correct classification suggests that the pattern does belong to the root cause k.

3 Implementation

The actual computation of the Mahalanobis distance for a *p*-variate pattern vector is a simple matter. However, implementation of an automated pattern recognition program in failure analysis (FA) is an involved proposition. Several methodological as well as logistical issues need to be addressed to successfully execute a pattern recognition program. In the following sub-sections we discuss the source of electrical test data, data formats, filtering data, missing data, and constructing root cause databases etc., in elaborate detail.

3.1 Sources of Electrical Test Data

Test data is obtained from a Keithley parametric tester. The Keithley is a parametric tester that obtains measurements on electrical test parameters commonly used in Semiconductor manufactuting. The tester is interfaced to a prober that is used to probe the process control bars on a wafer that contain test structures to measure the desired parameters. Typical test structures are resistors, capacitors, and transistors. Electrical contact to the structures is established via bond pads. When contact is established by the probes to the test structures, computer programs instruct the tester to measure the desired test structure characteristics. This data is analyzed to determine the device parameters on the process control bars on a wafer, prior to testing, visual inspection for superficial defects, and on to packaging for shipping to the end user.

3.2 Data Format

The resultant data from Keithley is formatted in a columnar form for readability by the pattern recognition program. The data stream consists of lot number, wafer number, site number, and the test probe measurements. A *p*-dimensional row vector of electrical test measurements constitutes a signature. A typical signature is given by $\boldsymbol{x}=\left(x_1,x_2,...,x_p\right)^T$. We will refer to a signature also as a pattern, and the constituents of a signature as features. Typical electrical test parameters are: drive current, breakdown voltage, sheet resistance, P+SD resistance, and N+SD resistance etc. It is typical in IC failure diagnostics that electrical test parameters in excess of 300 are studied to capture anomalies. Layout of the architecture of the SA program is given in Figure 1. Each root cause set consists of signatures and the corresponding root causes including the lot number, the wafer number, and the site number.

3.3 Databases

It is evident from Figure 1 that databases of root causes are a key to the implementation of the SA program. A file consisting of lot number, wafer numbers, site numbers, vector of electrical test measurements, and root causes is compiled. The file is separated by root causes and the *k* individual root cause databases is established. Upon

building the root cause databases, after working through the issues of test data, the root cause specific statistics such as mean vectors and covariance matrices are computed which are the ingredients that go to compute the Mahalanobis distance metric.

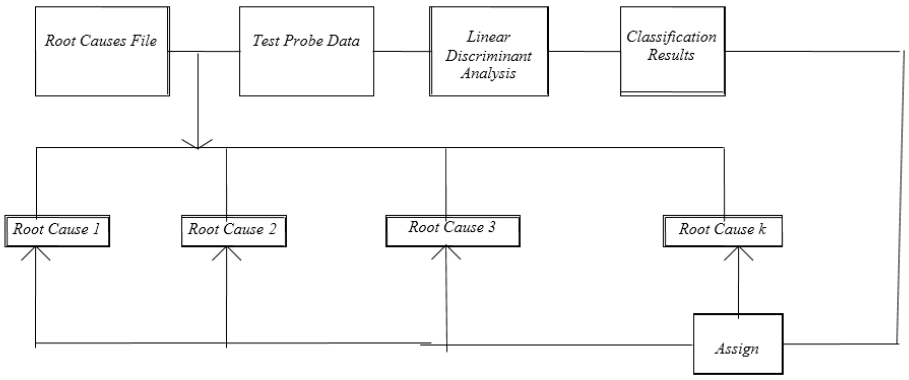

Fig. 1. Signature Analysis Program Layout

3.4 Handling Missing Data

Missing data occurring sporadically is not a problem and is readily handled. The term "sporadic" means that no more than five percent of the data is missing for any given parameter for each root cause set. The missing data is replaced by the corresponding component sample mean from the mean vector for that root cause. This substitution induces certain *artificialness* in the data, but enables the computation of the covariance matrix. Missing data is an ugly phenomenon that rears its head in a real-world application. Missing data could correspond to sparsely populated features in a signature, and also features being completely absent. Eliminating sparse features (>5% missing) in a given root cause could lead to signatures of unequal dimension residing in different root cause categories. If this is the case, the largest intersection of the components of the two vectors is taken. For example, sample mean vector $\overline{\boldsymbol{x}}_1$ for root cause set 1 consists of components $(x_1,-,x_3,x_4,x_5)$ and the mean vector $\overline{\boldsymbol{x}}_2$ for root cause 2 consists of components $(x_1,x_2,-,x_4,x_5)$, then the intersection of these two four dimensional vectors (four features present) would be the three dimensional vector with components $(x_1,-,-,x_4,x_5)$. There are several reasons why $\overline{\boldsymbol{x}}_1$ is missing its second component: either no data is available for this parameter (i.e., no data is available for test number 2 from the electrical test data) or data is available, but more than five percent is missing, or available data is bogus. Bogus data is due to failing tests by Keithley, or some artificial value is included in the electrical test data to indicate some other abnormality in testing. The greatest intersection method allows measurements of the same quantities to be compared. The resulting

decrement in the dimension of the pattern decreases the Mahalanobis distance metric, but the probability is accounted for in the degrees of freedom, which is the parameter of the chi-square statistic.

The greatest intersection is employed across all *k* root cause categories. For this reason, the root cause categories are carefully chosen so they are of the same dimension in the signatures inhabiting them. For example, if there are seven root cause categories where the dimension of the signatures are all the same, say 60 features, then the addition of an eighth root cause category with only fifty features would eliminate the information from 10 electrical test parameters. It should be noted that signatures are technology dependent since each technology consists of different test parameters. It is therefore essential to build databases of signatures and root causes by technology for SA implementation.

3.5 Collinearities

The dimension of the signatures should be reduced whenever possible to alleviate cumbersome and often unnecessary calculations. The device analysis engineer should eliminate those parameters that duplicate information; for instance, voltages of the same element measured from different reference points. Inclusion of highly *collinear* (correlated) measurements results in a singular or nearly singular covariance matrix. A *singular* covariance matrix is one which is not invertible. Notice that $\boldsymbol{S}_p^{-1}$ in $\hat{D}_j$ requires the inversion of the covariance matrix $\boldsymbol{S}_p$. High collinearities among the features in the signature cause *near singularities* that throw off the inverse of the covariance matrix $\boldsymbol{S}_p$. The greatest intersection method automatically reduces the dimension but further judicious reductions in the dimension of the signature should be constantly sought to avoid collinearities. Also, we employ the method of *step-wise discriminant analysis* to find an optimal subset of features in a signature. A brief overview of step-wise discriminant analysis is presented in Section 4.4.

3.6 Effects of Outliers

Classification by linear discriminant analysis is sensitive to *outliers*. An outlier is a quantity that does not belong to the assumed statistical distribution of that random signature. Outliers have harmful effects on classification by the Mahalanobis distance metric. While the covariance matrix gives lesser weight to parameters with high variation, it is not sensitive enough to compensate for outliers. It is desirable that intra-group variation within a root cause set be as small as possible so that a signature can be classified accurately. Thus outliers within a root cause set must be eliminated. In the present application, any measurement greater than 4.5σ (standard deviations) away from each component mean is deemed an outlier. Using this outlier window, harmful effects of outliers are mitigated.

3.7 Effects of Non-Normality

Electrical test parameter measurements in a signature do not satisfy the assumption of multivariate normality. A random vector is not multivariate normal when it is not Gaussian in *p*-dimensions. Severe departures from multivariate normality have a large effect on the Mahalanobis distance statistic. Since the statistic follows the chi-square distribution, which is a consequence of the assumption of multivariate normality, any departure of the data from the assumed statistical distribution results in probabilities of classification being extremely low. It in fact results in non-classification. However, the Mahalanobis distance metric is the smallest from the known root cause category for a new incoming signature. In the next section we introduce the artificial neural network approach as a pattern classifier.

4 Artificial Neural Networks

By definition a neural network is a massively distributed parallel processor that acquires knowledge by optimizing the inter-neuron synaptic strengths by a learning process. The learning process consists of an algorithm that systematically modifies the synaptic strengths to achieve the desired objective. The architecture of a neural network is a simplified analogue of the network of neurons within a human brain responsible for memory, pattern recognition and a host of other activities. See [7] for a comprehensive treatment of neurophysiology and neural networks. The computational elements that comprise a neural network, akin to the neuron are known as nodes, units, or processing elements. A neural network is an arrangement of neurons organized in layers. The simplest neural network is a single layer network consisting of neurons in the input layer feeding into an output layer. A schematic of a single layer *feedforward* network is given in Figure 2. The cumulative effect of the input layer nodes at the output node is a sum of the product of input values and their corresponding weights (synaptic strengths) plus a bias term.

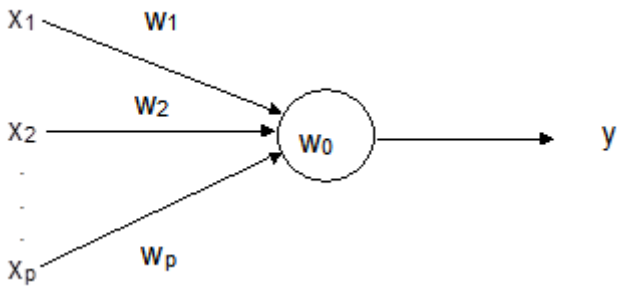

Fig. 2. Single layered feedforward network

It is feedforward in that the connections between inputs and outputs are in one (forward) direction only. A natural extension is a *multi-layered* feedforward network which consists of an input layer of neurons, a hidden layer of neurons feeding into an output layer of neurons. The number of hidden layers can be greater than one. A schematic of a fully connected, multi-layered feedforward neural network with one hidden layer is illustrated in Figure 3.

4.1 The Back Propagation Algorithm

The goal of a neural network algorithm is to establish a relationship between the inputs and their corresponding responses. Neural networks are chosen often when it is difficult to mathematically express a relationship between the inputs (signatures) and the outputs (classes). We will, in the following, formally define the constituents and the mechanics of the back propagation network that is usually effective in establishing a relationship or a mapping between the input signatures and the output classes. Suppose we have a set of data of p input-output pairs (the input-output pairs correspond to a vector of electrical parameter measurements and the probable cause of failure), $(x_1, y_1), (x_2, y_2),\ldots, (x_p, y_p)$. The input-output pairs are such that $x \in \Re^p$, *and* $y \in \Re^K$. Back propagation algorithm is used to train the neural networks to develop an approximate relationship $y \cong f(x)$. Training the neural network corresponds to finding the appropriate set of weights also known as synaptic strengths. The back propagation method usually achieves this objective given a sufficient number of training samples and correctly chosen input-output pairs. A detailed discussion of back propagation algorithm is given in [4]. Back propagation algorithm is a method to minimize the objective function $E_p = \frac{1}{2}\sum_{j=1}^{p}(dpj - opj)^2$, where E_p is the error due to the p^{th} signature vector, d_{pj} is the desired value for the j^{th} output neuron, and o_{pj} is the actual output of the j^{th} output neuron. We observe that each term in the sum is the contribution to the total error from a single output neuron. The minimization of E_p is achieved by finding its derivative with respect to the weights w_{ji} and computing the changes in weights in the direction of its negative gradient, in other words: $\Delta pwji\alpha - \frac{\partial Ep}{\partial wji}$, where w_{ji} is the weight connecting the j^{th} node in layer l and the i^{th} node in layer l-1.

4.2 Training with Back Propagation

A multi layered feedforward neural network with one hidden layer was chosen for training with the following characteristics:

- The number of elements in the input signature was chosen to be 34 after a careful review of significant parameters for the technology under investigation.

• The data is scaled such that the domain of any given electrical parameter is the unit interval (0,1).

• Sufficient training data relative to fourteen known root causes was established prior to training.

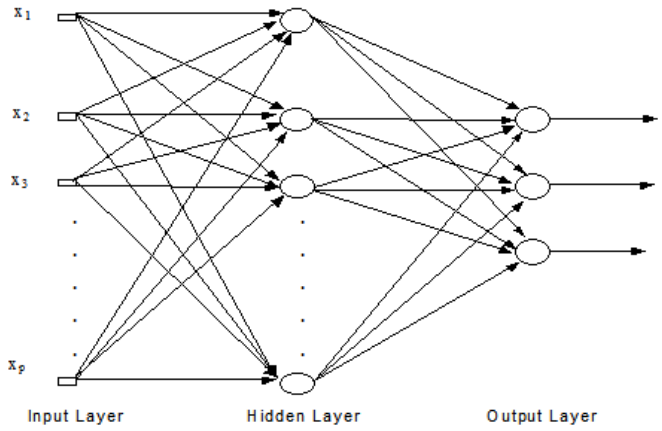

Fig. 3. Architecture of a multi-layer feed forward neural network

• The number of hidden layer units was chosen to be a third of the sum of input and output layer units. The number of hidden layer units is recommended by [4]. The computation amounts to 16 hidden layer units.

• The hidden layer and output layer weights (synaptic strengths) were generated from a Gaussian distribution with mean 0 and variance 0.3. It is recommended a small value for the variance is chosen.

• All the bias units were set to zero in the entire network.

• The fourteen output classes were set to an identity matrix of dimension 14x14 the largest element (1) replaced by 0.9 and the smallest element (0) replaced by 0.1.

• The training data was split into three sets, namely the training set, validation set, and the test set as recommended by [6].

• The training set was used to train several preliminary network architectures

• The validation set was used to identify the network with the least sums of squares of error.

• The third test set was used to assess the performance of the chosen network with examples independent of data used in training.

- The network was trained using the pattern by pattern approach. The patterns were selected randomly from each root cause category to eliminate any biases during learning.

4.3 Comments

In spite of a careful and judicious choice of electrical parameters that comprise the input signature, due to their impact on the assessment of the performance of a process control bar, some mildly correlated parameters were included in the signature vector. In the technology we chose for implementation, mild correlations among the features were not a major problem. To extract best performance, we needed to remove mild correlations within the signature, and achieve data parsimony. *Principal components analysis* (PCA) is a viable approach to enable both objectives. In addition to eliminating correlations among the features of the signature vector, PCA enables reduction in the dimension of the signature vector while retaining bulk of the information contained in the original vector of dimension 34. PCA is a statistical transformation procedure that eliminates correlations among the electrical parameters within a signature by expressing each parameter as a linear combination of test parameters. A detailed account of PCA is given in Section 4.4. We applied PCA, mainly to achieve reduction in the dimensionality of the signature vector. Upon its application, the signature dimension was reduced to 14. The neural network architecture within the PCA context included 14 input nodes, 9 hidden nodes and 14 output nodes corresponding to the classes of known failure mechanisms. Dimensionality reduction is desirable for a variety of reasons. Smaller signature vector results in fewer nodes within a neural work resulting in faster convergence to the desired solution. We urge the user to consider PCA for the dual purpose of signature vector orthogonalization and dimensionality reduction.

4.4 Step-Wise Discriminant Analysis

When the dimension of the incoming signature is very large, an approach found suitable to achieve parsimony is *step-wise discriminant analysis* [3]. In this procedure the signature is partitioned into two groups, such that the union of two groups yields the original signature. The procedure is described here briefly. Let $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2)$ be a partition of the original signature $\boldsymbol{X}$.

The mean vector and the covariance matrix corresponding to the partition are given as $\mu_i=(\mu'_{i1},\mu'_{i2})^T$, ($i=1,2,..k$), and $\Sigma=\begin{pmatrix}\Sigma 11 & \Sigma 12\\ \Sigma 21 & \Sigma 22\end{pmatrix}$ respectively under the usual assumption of the equicovariance normal model, $\boldsymbol{X}\sim N(\mu_i,\Sigma)$ in the categories C_i ($i=1,2,..k$). In terms of the conditional distribution of $\boldsymbol{X}_2 \mid \boldsymbol{X}_1$, the notion of addi-

tional information from the sub-vector $\boldsymbol{X}_2$ plays an important role in many variable selection methods. The hypothesis that $\boldsymbol{X}_2$ provides no additional information over $\boldsymbol{X}_1$ is statistically formulated as

$$H_0: \mu_{i2} - \mu_{h2} - \Sigma_{21}\Sigma_{11}^{-1}(\mu_{i1} - \mu_{h1}) = 0,$$

$i \neq h = 1,...,k$. Let $\tilde{B} = (k-1)B$ be the between group matrix sum of squares and $(n-k)S_p$ be the pooled within group sum of squares. Mathematically $\tilde{B}$ and S_p are given as $\tilde{B} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})(x_i - \bar{x})}{k-1}$ and $S_p^{-1} = \frac{(n_1-1)s_1 + ...+(n_k-1)s_k}{n_1 + ... + n_k - k}$.

Corresponding to a partition of $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, $\tilde{B}$ and S_p are partitioned as

$$\tilde{B} = \begin{pmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \tilde{B}_{21} & \tilde{B}_{22} \end{pmatrix} \quad \text{and} \quad (n-k)S_p = (n-k)S_p \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

The procedure for testing the hypothesis that $\boldsymbol{X}_2$ provides no additional information based on the matrices $\tilde{B}$ and S_p adjusted for $\boldsymbol{X}_1$ is

$$S_{2.1} = S_{22} - S_{21}S_{11}^{-1}S_{12} \quad \text{and}$$

$$\hat{B}_{2.1} = (\hat{B}_{22} + S_{22}) - (\hat{B}_{21} + W_{21})(\hat{B}_{11} + S_{11})^{-1}(\hat{B}_{12} + S_{12}) - S_{2.1}$$

with degrees of freedom $n-k-p_1$ and $k-1$, where for simplicity S_p is denoted by S, and p_1 is the number of electrical test parameters in the sub-vector $\boldsymbol{X}_1$. The test statistic for testing the hypothesis is given by $\Lambda_{2.1} = \frac{|S_{2.1}|}{|\hat{B}_{2.1} + S_{2.1}|}$, which is distributed as an F-statistic.

The step-wise discriminant analysis procedure is applied sequentially similar to standard step-wise procedures in multiple linear regression. Initially the signature vector $\boldsymbol{X}$ is partitioned into two groups such that one group consists on k-1 features and the other set only one electrical test parameter. Tests of hypotheses are sequentially conducted to add and delete features in the signature vector till an optimal set of features supplying additional information for pattern classification are identified.

- **Principal Component Analysis**

As we noted in the previous section, an important issue is the size of the input vector pattern. At the outset, it is advisable that one make a judicious choice of features based on the ability of the features to contribute to pattern recognition. *Principal component analysis* (PCA) alluded to as Karhunen Loeve transformation (KLT) in communication theory is a data reduction technique that enables dimensionality reduction while retaining intrinsic information in the original data set. In our context, it enables finding a subset of effective features from the original set. The details of the procedure are given in the following.

Algebraically principal components are particular linear combination of the p features $x_1, x_2,\ldots,x_p$ in the input pattern vector. Geometrically, these linear combinations represent a new coordinate system obtained by rotating the original system with features $x_1, x_2,\ldots,x_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a parsimonious description of the original data. Although p principal components are required to account for the total system variability, majority of the variation is captured by a smaller number m. The m principal components can then replace the original p features. Thus, the original data set consisting of p features with n measurements each, is replaced by k principal components with n measurements each.

Since the objective of PCA or KLT is to explain variability in a system, it depends entirely on the *variance-covariance* matrix (structure) of the data. Construction of principal components does not depend on the assumption of multivariate normality. However, principal components derived from data arising from multivariate normal populations have useful interpretations in terms of constant density ellipsoids. In the following we describe the construction of principal components of the features x_1, $x_2,\ldots,x_p$ in the training database.

Let $X = [x_1, x_2,\ldots,x_p]$ be the input feature vector with the *variance-covariance* matrix Σ. Let Σ have the eigen value-eigen vector pairs $(\lambda_1, e_1), (\lambda_2, e_2),\ldots\ldots,(\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq,\ldots..,\ \lambda_p \geq 0$. The i^{th} principal component is given by $y_i = e_{1i}x_1 + e_{2i}x_2 +\ldots+e_{pi}x_p$, $(i = 1,2,\ldots\ldots,p)$, where $(e_{1i}, e_{2I},\ldots,\ e_{pi})$ are the components of the i^{th} eigen vector of the *variance-covariance* matrix Σ. With the choice that the coefficients of the i^{th} principal component being the components of the i^{th} eigen vector, the variance of $y_i = \lambda_i$, and correlation between y_i, and y_j is equal to zero. In other words, principal components which are linear combinations of original features are orthogonal (uncorrelated) to each other. Since the *variance-covariance* matrix Σ is seldom known, we estimate from the set of input patterns in the training database. To extract principal components from a set of input patterns, standard statistical packages such as SAS®, S-PLUS® can be utilized. For a detailed account of the theory and application of principal components, see [1]. From the discussion above it is clear that computation of principal components requires extraction of the eigen values and eigen vectors. Eigen value and vectors can be obtained by using such procedures as LU decomposi-

tion or QR decomposition of Σ. Numerical recipes in C [6] is a good resource for the source code required.

5 Results

The following are some examples of the application of linear discriminant analysis and artificial neural networks for integrated-circuit failure analysis. To validate the two techniques, we saved some signatures with known root causes for testing. The saved signatures did not contribute to the sample statistics computed or participate in neural network training.

Example. Two lots failing due to missing N+S/D implant were submitted to the automated signature analysis program for root cause identification. A signature of length 34 is applied to the program for pattern classification. The signature is from a certain device XXXXXXXXX belonging to technology YY. Tables 1 and 2 show the results of this analysis. The number of electrical test parameters measured for this technology is 123, but a signature of dimension 34 is applied for classification. The reason is that a careful selection of 72 test parameters for SA was chosen by engineering. Further, application of step-wise discriminant analysis reduced the dimension to 34.

The program classified signatures from each site into a root cause category. It is known that site 2 on wafer 17 was failing due to missed N+ S/D implant problem. LDA clearly identified the failing signature to belong to the known root cause. The probability column corresponds to the probability that the signature is from a population known to possess N+ S/D implant problem as the root cause. The low D value with high probability of classification suggests that the algorithm was able to accurately classify the new signature. The output summarizes the statistics relative to the top four probable root causes. Applying the neural network (standard and PCA) approach to classification, the results are consistent with those from LDA. Classification by a neural net is realized by choosing that root cause for which the error sum of squares is small. We reiterate that the error sum of squares is the square of the accumulated error between an input signature and the desired response. In the second example form lot 9749493, wafer 3, and site 2, although the distance metric D due to LDA was the smallest for the known root cause category, the probability of classification was very small. The computation of the probability of classification hinges on the distribution of sample D which is asymptotically chi-square under the assumption of multivariate normality of the signature vector. It is our suspicion that the statistical distribution of the signature vector is not multivariate normal. The data is perhaps from a heavy tailed distribution causing the probabilities to be small. The neural network found the appropriate root cause with the smallest error sum of squares.

Table 1. Classification by Linear Discriminant Analysis and Artificial Neural Networks

Lot Number	**9745158**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

LDA by Site

Wafer Number	Site Number	MD	Root Cause	probability
17	2	9.65	Missed N+S/D implant	0.980000
17	2	2.03.84	Missed Nwell Implant	0.000000
17	2	367.48	High Epi Doping	0.000000
17	2	408.77	Sidewall Overetch	0.000000

Lot Number	**9745158**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

ANN by Site

Wafer Number	Site Number	squarrred error	Root Cause
17	2	0.4857	Missed N+S/D implant
17	2	2.7752	Missed Nwell Implant
17	2	2.884	Missed DUF Implant
17	2	2.9477	Via

Lot Number	**9745158**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

ANN (principal components) by Site

Wafer Number	Site Number	squarrred error	Root Cause
17	2	0.6221	Missed N+S/D implant
17	2	0.8144	High Epi Doping
17	2	0.9815	Missed DUF Implant
17	2	1.0953	Via

Table 2. Classification by Linear Discriminant Analysis and Artificial Neural Networks

Lot Number	**9749593**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

LDA by Site

Wafer Number	Site Number	MD	Root Cause	probability
3	2	1333.1	Missed N+S/D implant	0.000000
3	2	1666.11	Thin Gate Oxide	0.000000
3	2	1817.46	Missed Nwell Implant	0.000000
3	2	1849.86	Low Epi Doping	0.000000

Lot Number	**9749593**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

ANN by Site

Wafer Number	Site Number	squarred error	Root Cause
3	2	0.2937	Missed N+S/D implant
3	2	1.619	Thin Gate Oxide
3	2	1.7003	Missed Nwell Implant
3	2	1.7286	Low Epi Doping

Lot Number	**9745158**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

LDA by Site

Wafer Number	Site Number	MD	Root Cause	probability
17	2	9.65	Missed N+S/D implant	0.980000
17	2	2.03.84	Missed Nwell Implant	0.000000
17	2	367.48	High Epi Doping	0.000000
17	2	408.77	Sidewall Overetch	0.000000

6 Conclusion

Linear discriminant analysis and artificial neural networks are viable tools for pattern classification in an integrated-circuit manufacturing environment. They aid the practicing engineer to identify potential root causes. Depending on the type of data collected from wafers, a practitioner chooses a method based on signatures (Sections 2-5). Methods based on LDA and ANN can be implemented during the course of production. For the implementation of these methods, it is paramount to build a database that consists of a plethora of root causes. This helps to suitably associate signatures to probable root causes. While the manual inspection of failed chips takes days to isolate root causes, automation of signature analysis by pattern classification methods can

reduce cycle times of failure analyses by orders of magnitude and furthermore improve the quality of product produced.

Manufacturing data does not behave in a manner that it can be conveniently modeled by a probability distribution. The linear discriminant technique depends on the assumption of multivariate normality of the signature vector. This assumption is often violated. It may be that the underlying random mechanism is a *heavy-tailed distribution*. Loosely speaking, a heavy-tailed distribution is one for which the probability distribution tapers off much slower than the Gaussian. Any probability that is computed based on the assumption of multivariate normality, which leads the D statistic to follow a chi-square distribution asymptotically, tends to be very small. Practitioners typically interpret this as the inability of the algorithm to classify correctly. While this argument may be valid, it is our experience that the D statistic value, although very large for all the root causes, was smallest for the root cause suspected for a new failing signature.

In the situation when the assumption of multivariate normality is feared to be violated, one may apply a *variance-stabilizing transformation* ([15], chap. 15) that will typically reduce or eliminate the effect of non-normality. Logarithmic, square root, and arcsine transformations applied to single components of signatures of are rather standard.

References

1. R.A. Johnson, and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 3rd edition, Englewood Cliffs, New Jersey: Prentice Hall, 1992.
2. R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons, 1973.
3. G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, New York: John Wiley & Sons, 1992.
4. J. A. Freeman and D.M. Skapura, *Neural Networks, Algorithms, Applications, and Programming Techniques*, Computation and Neural systems Series, Reading, Massachusetts: Addision Wesley, 1991.
5. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd edition, Cambridge: Cambridge University Press, 1996.
6. B. D. Ripley, *Pattern Recognition and Artificial Neural Networks*, Cambridge: Cambridge University Press, 1996.
7. S. Haykin, *Neural Networks. A Comprehensive Foundation*, New York: Macmillan College Publishing, 1994.
8. K.W. Tobin, S.S. Gleason, T.P. Karnowski, S.L. Cohen, Fred Lakhani, *Automatic Classification of Spatial Signatures on Semiconductor Wafermaps*, Private communication.
9. S.S. Gleason, K.W. Tobin, T.P. Karnowski, Spatial Signature Analysis of Semiconductor Defects for Manufacturing for Manufacturing Problem Diagnosis, Solid State Technology, July, 1996.

10. I. T. Jolliffe, *Principal Component Analysis*, New York, Springer-Verlag, 1986.
11. D.J. Hand, *Discrimination and Classification*, New York: John Wiley and Sons, 1981.
12. M. Baron, C. K. Lakshminarayan, Z. Chen, Markov Random Fields in Pattern Recognition for Semiconductor Manufacturing, Technometrics, 43, 66-72, 2001.
13. M. D. Longtin, L. M. Wein, R. E. Welsh, Sequential Screening in Semiconductor Manufacturing, I: Exploiting Spatial Dependence, Operations Research,44, 173-195,1996.
14. W. Taam, M. Hamada, Detecting Spatial Effects from Factorial Experiments: an Application from Integrated-Circuit Manufacturing, Technometrics, 35, 149-160.
15. G. W. Snedekor, W. G. Cochran. *Statistical Methods*, 8th edition, Ames: Iowa State University Press, 1989.