Mechanical Systems and Signal Processing (1989) 3(4), 405-424

STATISTICAL PROCESS CONTROL OF ACOUSTIC EMISSION FOR CUTTING TOOL MONITORING

ALI A. HOUSHMAND AND ELIJAH KANNATEY-ASIBU, JR†

Department of Industrial and Operation Engineering and † Department of Mechanical Enginering and Applied Mechanics, The University of Michigan, Ann Arbor, MI 48109, U.S.A.

The problem of cutting process monitoring has been investigated in recent years, with encouraging results, using pattern recognition analysis of acoustic emission (AE) signals. The analyses are based on linear discriminant functions, which assume that the observed data (from each class) are independent random samples from multivariate normal distributions with equal covariance matrices. However, in a number of practical situations some (or all) of these assumptions may not necessarily hold, resulting in errors in the analysis.

In this paper, the distributions of AE spectra generated in earlier work are first analysed, and the results indicate departure from the assumptions, although the lack of normality was not too severe. Relaxing the assumption of equality of the covariance matrices, quadratic discriminant function analysis produced improved results for tool wear and chip noise monitoring while degrading tool fracture detection. The latter is due to inadequacy of the amount of data used in training the system. It is expected that increasing the data base would improve the results for all classes.

The analysis until now has focused on reducing the dimensionality of the feature space by eliminating the features with the least discriminatory power. Even though this inevitably reduces the performance of the system, it is a necessary compromise for increased computational speed. To make use of the entire feature set with a reduced matrix rank, a principal component analysis is investigated. The result is a substantial improvement in correct classification of AE signals, even under different cuting conditions.

1. INTRODUCTION

Automation of metal cutting processes have always been a principal goal of the manufacturing industry. However, the inability to monitor completely the condition of the cutting tool in real time has been a major obstacle to achieving this goal. Tool breakage and tool wear can result in substantial cost through damage to machinery and parts produced, as well as the cost associated with machine downtime.

Much research has recently been done to monitor the conditions of cutting tools in machining processes. In recent years, acoustic emission (AE) has been investigated as an effective sensing technique for machining processes. AE refers to elastic stress waves generated as a result of the rapid release of strain energy within a material due to a rearrangement of its internal structure. It has been applied to a variety of situations including weld flaw detection, fracture and crack propagation in pressure vessels, and mechanical equipment and material property evaluation during tensile testing. AE can be used to obtain direct information about the major activities of metal cutting, including plastic deformation, frictional contact, and fracture of both chip and tool. However, a major problem area has been the development of signal decomposition schemes that can identify and separate the contributions from these sources.

Statistical techniques have been used to develop a methodology for AE signal decomposition. The methods generally used to analyse AE signals corresponding to tool wear and

0888-3270/89/040405+20 $03.00/0

tool breakage were based on descriptive statistics, e.g. [1-4]. Kannatey-Asibu and Dornfeld [5] list several statistical methods which have been used in the past to analyse AE signals. Among them are: "count and count rate" (which is a record of signals whose amplitude exceed a pre-set threshold voltage), "amplitude distribution analysis" (which gives an indication of the number of signals whose amplitude fall within a predefined range), and "frequency spectrum" (which shows the contribution of each frequency component to the total power). Also, Kannatey-Asibu and Dornfeld [3] carried out experiments and analysed the data by assuming that the distribution of the rms value of AE signals follows a beta-distribution with some parameters. They found that the skewness and the kurtosis which are functions of the parameters of the assumed beta-distribution, were sensitive to progressive tool wear.

Kannatey-Asibu [6] proposed the use of linear discriminant function analysis as a technique for AE signal decomposition, in order to identify sources of AE signals in metal cutting processes. Following this work, others have studied the use of this technique to monitor the condition of cutting tools in metal cutting processes. See, for example [7, 8]. Emel and Kannatey-Asibu [9] conducted controlled experiments on a lathe under three different cutting conditions, and obtained data for four different classes, viz, chip noise, tool fracture, sharp tool, and worn tool. The data was used to develop linear discriminant functions for classifying future incoming signals. They concluded that the results indicate an 84-94% reliability for detecting tool failure of any type.

Although classification seems to be a good statistical technique for signal decomposition, much work needs to be done before one can use it to analyse AE signals for on-line cutting tool monitoring. This paper discusses the use of appropriate statistical techniques for analysis of frequency domain AE signals to monitor cutting tools. It will be shown, using the experimental data of Emel and Kannatey-Asibu [9], that substantial improvement over the results to date is achieved by the use of these techniques. The next section gives a general statement of the problem that we are investigating.

2. PROBLEM STATEMENT

Consider a process that can be in one of s ($s \geq 2$) mutually exclusive states. One of these states is defined as *in control*, and the rest are defined as *out of control*. For example, in a metal cutting process the in control state corresponds to a sharp and healthy cutting tool, whereas the out of control states correspond to a worn tool, fractured tool, or both. Suppose that the process can be monitored, at some regular intervals, through measurements of k-dimensional vectors of observations (attributes). In the above metal cutting example, these vectors of observations can each be a k-dimensional vector of spectral power components of AE signals, generated at regular intervals from the cutting process. Suppose that before actual monitoring of the process begins, controlled experiments have been conducted, where the process has been set at each possible state i ($i = 1, 2, \ldots, s$), and n_i vectors of attributes (we refer to these vectors of attributes as the training sets) have been obtained from each of these states. For $i = 1, 2, \ldots, s$ states, let us denote these vectors of attributes by $\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i$. The superscript i specifies that an observation belongs to state i, and the subscripts $1, 2, \ldots, n_i$ specify the order (time interval) in which the observations have been taken. For example, in the cutting process, $i = 2$ may denote a worn tool. Then $\mathbf{x}_1^i, \mathbf{x}_2^2, \ldots, \mathbf{x}_{20}^2$ denote 20 vectors (e.g. k-dimensional spectral power components of AE signals) of observations which have been taken at regular intervals from a controlled experiment where a worn tool has been used for machining.

Now let us assume that the actual cutting process starts in control, i.e., with a sharp and "healthy" tool. As cutting continues, the tool experiences wear. At some point in the process, we observe a single k-dimensional vector of attributes, $\mathbf{x}' = (x_1, x_2, \ldots, x_k)$. If it is determined that x is generated by an out of control state, i.e., worn or broken tool, then the process will be stopped for inspection and corrective action taken, otherwise it will continue. The problem we will be concerned with is to determine which one of the s mutually exclusive states has generated x, and whether to stop the process for inspection and correction.

3. BACKGROUND

The problem of classification may be considered as a "statistical decision function" problem, where we have a number of hypotheses each stating that the subject under investigation belongs to a particular distribution. We must accept one of these hypotheses and reject the others. The classification problem was first studied by Fisher [10], followed by Smith [11]. It has since been studied by many authors, see, for example [12-19]. In this section, we briefly discuss the classification functions as well as the principal components analysis as the basis for our data analysis.

3.1. DISCRIMINANT FUNCTIONS

Theorem. Suppose x is an observation from one of the s multivariate populations Π_i with density $f_i(\mathbf{x}|\mathbf{\Omega}_i)$, where $\mathbf{\Omega}_i$ represents the known matrix of the parameters of the distribution. If p_i is the *a priori* probability of x belonging to Π_i, and if C_{ij} is the cost of misclassifying x as belonging to Π_i when it actually belongs to Π_j, then the regions of classification $R_1, R_2, \ldots, R_s$ that minimise the expected cost of misclassification are defined by assigning x to R_r if

$$\sum_{\substack{i=1\\(i\neq r)}}^{s} p_i C_{ri} f_i(\mathbf{x}|\mathbf{\Omega}_i) < \sum_{\substack{i=1\\(i\neq j)}}^{s} p_i C_{ji} f_i(\mathbf{x}|\mathbf{\Omega}_i) \quad \text{for all } j = 1, 2, \ldots, s; j \neq r. \tag{3.1}$$

For $s = 2$, the above theorem states: classify x into R_1 if

$$p_2 C_{12} f_2(\mathbf{x}|\mathbf{\Omega}_2) < p_1 C_{21} f_1(\mathbf{x}|\mathbf{\Omega}_1). \tag{3.2}$$

Now suppose that Π_i is a multivariate normal distribution with known mean vector $\boldsymbol{\mu}_i$, and covariance matrix Σ_i, $(i = 1, 2)$. By the above theorem, in order to minimise the expected cost of misclassification, we classify x into Π_1 if:

$$\frac{|\Sigma_1|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\}}{|\Sigma_2|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\}} \geq \frac{p_2 C_{12}}{p_1 C_{21}}. \tag{3.3}$$

Taking the natural logarithm (ln) of both sides and simplifying gives

$$[(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - (\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)] \geq 2 \ln \left[\frac{|\Sigma_1|^{1/2} p_2 C_{12}}{|\Sigma_2|^{1/2} p_1 C_{21}}\right]. \tag{3.4}$$

Equation (3.4) is the quadratic discriminant function of x. If $\Sigma_1 = \Sigma_2$, $p_1 = p_2$, and $C_{12} = C_{21}$, then (3.4) reduces to the simple linear discriminant function of x given by:

$$\mathbf{x}'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0. \tag{3.5}$$

If there are more than two populations $(s > 2)$, and their misclassification costs are equal, then the rule for classifying x into one of these populations is to calculate the *a posteriori* probability for each population and assign x to that population with the largest *a posteriori* probability. In the case of unequal normal population coveraince matrices, the *a posteriori* probability for population i can be obtained by exponentiating the expression

$$-\tfrac{1}{2}\mathbf{x}'\Sigma_i^{-1}\mathbf{x}+\mathbf{x}'\Sigma_i^{-1}\boldsymbol{\mu}_i-\tfrac{1}{2}\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i+\ln p_i-\ln|\Sigma_i|^{1/2} \tag{3.6}$$

for each i. If the covariance matrices are equal the *a posteriori* probability for the population i can correspondingly be obtained from

$$\mathbf{x}'\Sigma^{-1}\boldsymbol{\mu}_i+\ln p_i-\tfrac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i, \tag{3.7}$$

which is the linear discriminant function of $\mathbf{x}$.

The results described above are based on the assumption that the population parameters are all known. However, in practice these are seldom known, and therefore must be estimated from data. This means that the above results are valid if large samples are available to give reliable estimators.

3.2. PRINCIPAL COMPONENT ANALYSIS

Suppose we have a k-dimensional random vector $\mathbf{X}'=(X_1, X_2, \ldots, X_k)$ having a multivariate (not necessarily normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We can find linear combinations $Y_1, Y_2, \ldots, Y_r$ of $X_1, X_2, \ldots, X_k$ $(r<k)$ which convey approximately the same amount of variance as expressed by the original variables. The variables $Y_1, Y_2, \ldots, Y_r$ are called the first r principal components of the dependence structure among $X_1, X_2, \ldots, X_k$, and are evaluated as follows:

Let

$$Y_i=\sum_{j=1}^{k}\alpha_{ij}X_j \quad \text{for } i=1,2,\ldots,r.$$

Then

$$\text{maximise}_{\{\alpha_{1k}\}}\ \text{var}(Y_i)=\text{var}\left(\sum_{j=1}^{k}\alpha_{1j}X_j\right)$$

subject to the constraint

$$\sum_{j=1}^{k}\alpha_{1j}^2=1.$$

This gives $\boldsymbol{\alpha}_1'=(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k})$ as the set of weights which maximise $\text{var}(Y_1)$. $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$ which is equal to $\text{var}(Y_1)$. Next we solve

$$\text{maximise}_{\{\alpha_{2k}\}}\ \text{var}(Y_2)=\text{var}\left(\sum_{j=1}^{k}\alpha_{2j}X_j\right)$$

subject to the constraint

$$\sum_{j=1}^{k}\alpha_{2j}^2=1 \quad \text{and} \quad \text{covar}(Y_1, Y_2)=0.$$

This gives $\boldsymbol{\alpha}_2'=(\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k})$ as the set of weights which maximise $\text{var}(Y_2)$. $\boldsymbol{\alpha}_2$ is the eigenvector corresponding to the second largest eigenvalue of $\boldsymbol{\Sigma}$ which is equal to $\text{var}(Y_2)$. After $Y_1, Y_2, \ldots, Y_{r-1}$, have been obtained, we solve

$$\text{maximise}_{\{\alpha_{rk}\}}\ \text{var}(Y_r)=\text{var}\left(\sum_{j=1}^{k}\alpha_{rj}X_j\right)$$

subject to the constraint

$$\sum_{j=1}^{k} \alpha_{rj}^2 = 1,$$

and covar $(Y_m, Y_r) = 0$ $(m = 1, 2, \ldots, r-1)$.
This gives $\boldsymbol{\alpha}_1' = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k})$ as the set of weights which maximise var (Y_1). $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the r-th largest eigenvalue of $\boldsymbol{\Sigma}$ which is equal to var (Y_r). Therefore, the principal component analysis rotates the axes of $X_1, X_2, \ldots, X_k$ to obtain the new orthogonal axes $Y_1, Y_2, \ldots, Y_r$ which account for a large portion of the total variance of the X_is.

4. RESULTS AND DISCUSSION

In this section, we present and discuss the result of our analysis, but first a brief description on how the data [9] were obtained.

Controlled experiments were conducted on a lathe machine under three sets of cutting conditions, which resulted in three sets of data. We refer to these three sets of data as "set 1", "set 2", and "set 3". The cutting conditions corresponding to data sets 1 and 2 are described in Table 1 of Emel and Kannatey-Asibu [9] (they refer to sets 1 and 2 as tests 11 and 13 respectively). Below, we briefly describe how the frequency domain AE signals were obtained for set 1. Throughout, we use set 1 to report our findings, and use the other two sets for confirmation.

For each of the four possible states (tool fracture, chip noise, sharp tool, and worn tool), separate recordings of time domain AE signals were made. For example, for the tool fracture state, 29 separate experiments were conducted where each time the tool was deliberately fractured during the machining process and in each case the corresponding AE signals were recorded. One-millisecond time portions of each of these signals were then sampled, corresponding to 4096 data points. The time domain data were then transformed to the frequency domain using the fast Fourier transform (FFT). This resulted in 2048 data points in the frequency range of 1-1000 kHz. The sample means of every 40 data points were then calculated to represent 51 (4048 divided by 40) features each with a 20 kHz frequency range. The first five features were then discarded due to the low signal-to-noise ratio, in essence filtering off the lower 100 kHz frequencies. This resulted in 46-dimensional vector of features (variables). These 46 power spectral components were further normalised with respect to the total power of the spectrum, using the expression

$$X_k = 10 \log \left(\frac{Z_k}{\sum_{k=1}^{46} Z_k} \right) \quad \text{for } k = 1, 2, \ldots, 46,$$

where Z_k represents the contribution of feature k. This makes the power components invariant to the absolute energy. The end results were 29 independent vectors, each consisting of 46 power components.

For the sharp and worn tool states, separate tests were conducted where sharp and worn tools were used, respectively. In recording continuous AE signals generated from these experiments, transient signals due to chip breakage were also periodically recorded. Chip noise signals were thus obtained from the same tests conducted for sharp and worn tools. For each of these tests, a total of 20 samples (46-dimensional vectors generated in exactly the same way as above) were obtained. Also, from each experiment, up to 20 chip noise samples were obtained. The tests with sharp and worn tools were repeated four times each, so a total of 80 vectors of observations were obtained for each of sharp and worn tool states, and a total of 74 vectors of observations were obtained for the chip noise.

Let us assume now that the data from each state are independent random samples from multivariate normal distributions with unknown parameters, and that the sample sizes are large enough so that good estimates of the parameters of these distributions are possible. Furthermore, assume that the population covariance matrices are equal (even though the test for equality of these matrices revealed that they are not). Substituting the appropriate estimates for the parameters in expression (3.7), and using the forward stepwise procedure described in Appendix A, we obtained the linear discriminant functions of data set 1 for the four states (tool fracture, chip noise, sharp tool, and worn tool). We assumed that the *a priori* probability of an observation belonging to state i, p_i, as well as the misclassification costs, c_{ij}, are the same for all i and j. Also, we develop discriminant functions for all four states in order to perform simultaneous classification. This will give results comparable to those in Table 4 of Emel and Kannatey-Asibu [9]. The stepwise procedure selected 14 features (variables) as the most discriminatory features. These features, together with their corresponding frequency bands, are given in Table 1.

Table 2 gives the number of the vectors classified into each state, and the percentage of correct classification which are represented by the boldface numbers along the diagonal. The percentage of correct classifications for the chip noise and the tool fracture states is reasonable but that of the sharp and the worn tool states is very poor. The first possible remedy is to test for the equality of the covariance matrices to see if the quadratic discriminant functions are more appropriate. The percentage of correct classifications in Table 2 can be compared with those obtained by Emel and Kannatey-Asibu [9], shown in Table 3. They used 20 features which were selected by ranking the F statistics corresponding to one-way Anova (see, for example [20]) and selecting the first 20. This feature selection was based on the assumption that the 46 features (variables) were independent, which is not necessarily the case, as illustrated by the data in Table 4 which gives the Pearson correlation coefficients between each pair of eight arbitrarily selected

TABLE 1

The list of 14 features, selected by stepwise discriminant analysis, and their corresponding frequencies

Feature	Frequency (kHz)
3	140
4	160
7	220
8	240
9	260
10	280
12	320
15	380
16	400
34	760
38	840
39	860
44	960
46	1000

TABLE 2

The results of LD analysis of the 14 features selected by stepwise procedure. The top entries are number of vectors classified into a population. The bold entries are percentage correct classification

True population	Classified population			
	Chip noise	Tool fracture	Sharp tool	Worn tool
Chip noise	67 **90**	4	0	3
Tool fracture	2	27 **93**	0	0
Sharp tool	0	0	66 **82**	14
Worn tool	2	0	23	55 **69**

TABLE 3

The results of the linear discriminant analysis of Emel and Kannatey-Asibu [9] *for data set* 1 (*Table* 4 *in their paper*)

	State			
	Chip noise	Tool fracture	Sharp tool	Worn tool
Percentage correct classification	84	85	75	61

features in each state. This means that the inclusion of a given feature in the discriminant function may very well depend on the inclusion or exclusion of some other features. This fact has been taken into account in our feature selection procedure (see Appendix A for discussion on how to account for this correlation structure in feature selection). The correlations between each pair of the remaining 38 features were similar to those in Table 4. The four values in each box represent the correlations between the feature specified in the row and the feature specified in the column for each of the four states. For example, the correlation between feature 8 and 7 is 0·248 for the chip noise, 0·34 for the fracture, 0·539 for the sharp tool, and 0·732 for the worn tool state. Most of these correlations are highly significant, which can be compared with the threshold values corresponding to 95% level of significance, given at the bottom of the table.

The test for equality of population covariance matrices (see, for example [17] for the derivation of the test statistics) was rejected at any level of significance less than 0·001, so expression (3.6) was used to obtain the quadratic discriminant functions for data set 1. Table 5 gives the number of vectors classified into each state and the percentages of correct classification along the diagonal. As expected, the quadratic procedure performs better overall. However, for the tool fracture state, the percentage of correct classifications drops from 93 to 79. This is clearly due to the small sample size (29) in this state (Foley [21] suggests that there should be at least four times as many samples as there are features in order to obtain reliable discriminant functions). Also, the covariance matrices must

TABLE 4

Pearson correlation coefficients ρ ($\rho = \text{covar}(X_i, X_j)/(\text{SQRT}(\text{var}(X_i) \cdot \text{var}(X_j)), i \neq j)$ *between the indicated features*

Feature	State	Fr_3	Fr_4	Fr_7	Fr_8	Fr_9	Fr_{10}	Fr_{12}
Fr_4	CN	0·304						
	FR	0·405						
	SH	0·726						
	WR	0·667						
Fr_7	CN	0·289	0·253					
	FR	−0·621	−0·219					
	SH	0·609	0·645					
	WR	0·448	0·638					
Fr_8	CN	−0·121	0·080	0·248				
	FR	−0·080	0·342	0·340				
	SH	0·276	0·338	0·539				
	WR	0·356	0·449	0·732				
Fr_9	CN	−0·208	−0·156	−0·156	0·389			
	FR	−0·158	0·311	0·260	0·724			
	SH	0·106	0·058	0·325	0·392			
	WR	0·331	0·361	0·684	0·717			
Fr_{10}	CN	0·067	0·154	0·277	0·283	0·218		
	FR	0·155	0·692	0·066	0·550	0·541		
	SH	−0·281	−0·266	−0·115	0·196	0·447		
	WR	0·134	0·252	0·604	0·677	0·624		
Fr_{12}	CN	−0·252	−0·028	0·133	0·239	0·136	0·169	
	FR	−0·547	−0·475	0·187	−0·071	−0·041	−0·468	
	SH	−0·633	−0·534	−0·432	−0·102	0·243	0·542	
	WR	−0·269	−0·182	0·145	0·316	0·436	0·429	
Fr_{15}	CN	−0·296	−0·523	−0·560	−0·141	0·034	−0·379	−0·259
	FR	0·054	−0·207	−0·240	−0·683	−0·501	−0·455	0·074
	SH	−0·802	−0·760	−0·639	−0·292	−0·372	0·372	0·720
	WR	−0·713	−0·699	−0·425	−0·262	−0·150	−0·078	0·537

CN: chip noise; FR: tool fracture; SH: sharp tool; WR: worn tool.
Significant at 5% if entries are: >0·229 for CN, >0·367 for FR, and >0·22 for SH and WR.

be estimated by available observations. This means that with 14 variables, a total of 98 parameters (the elements of a lower triangular covariance matrix for tool fracture state) must be estimated from a total of 406 data (29 × 14) points. This clearly reduces the number of degrees of freedom available for good estimation, resulting in poorer classification.

The results of Table 5 are not particularly good, especially in the case of detecting and discriminating between sharp tools and worn tools with up to 18% error in classification. Further examination of the data set 1 provides some explanations for these poor results. Table 6 gives the lag one serial correlations of the 14 variables selected for the discriminant analysis above. Similar serial correlations were present for the remaining 32 variables in each state. The serial correlations are significant for sharp and worn tool (this clearly is a violation of the underlying assumption of independence of observations in discriminant analysis). They are less strong for chip noise, and almost non-existent for tool fracture. The reason for lack of serial correlations between the observations for tool fracture in

TABLE 5

Results of QD analysis of the 14 selected features listed in Table 1. Top entries are number of vectors classified into a population. Bold entries are percentage correct classification

True population	Predicted population			
	Chip noise	Tool fracture	Sharp tool	Worn tool
Chip noise	74 **100**	0	0	0
Tool fracture	6	23 **79**	0	0
Sharp tool	0	0	66 **82**	14
Worn tool	3	0	10	67 **84**

TABLE 6

Lag 1 serial correlations of 14 selected variables (features) selected in the stepwise discriminant analysis

Variables (features)	Chip noise	Tool fracture	Sharp tool	Worn tool
3	0·155	0·018	0·538	0·461
4	0·230	0·146	0·473	0·570
7	0·137	−0·088	0·263	0·314
8	0·015	0·053	0·176	0·286
9	−0·055	−0·089	0·223	0·509
10	−0·251	0·028	0·160	0·187
12	−0·021	−0·187	0·198	0·255
15	0·259	−0·193	0·543	0·436
16	−0·088	−0·097	0·561	0·549
34	0·329	0·362	0·606	0·636
38	0·444	0·101	0·647	0·606
39	0·547	−0·024	0·550	0·574
44	0·516	0·095	0·414	0·580
46	0.458	−0·033	0·438	0·621
Significant at				
5%	0·229	0·367	0·220	0·220
1%	0·298	0·471	0·286	0·286

this study was the independent experiments which generated the 29 vectors (xs) of observations.

Another problem with the data is the lack of normality which was assumed for the above linear and quadratic discriminant function analyses. Figures 1–8 give histograms of eight features, two for each state. The shapes and the patterns of the histograms of the remaining variables in each state were similar to these eight histograms. For normally distributed observations, skewness and kurtosis must equal zero. Departure of these two measures from zero is sufficient indication of departure from normality. Nakanishi and

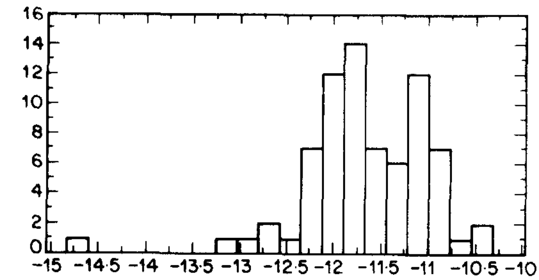

Figure 1. Histogram of feature 4 from chip noise. Skewness = −1·41, kurtosis = 5·40.

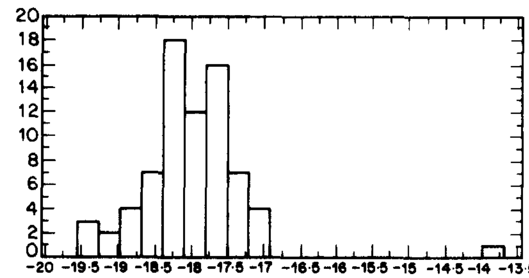

Figure 2. Histogram of feature 16 from chip noise. Skewness = 2·16, kurtosis = 12·33.

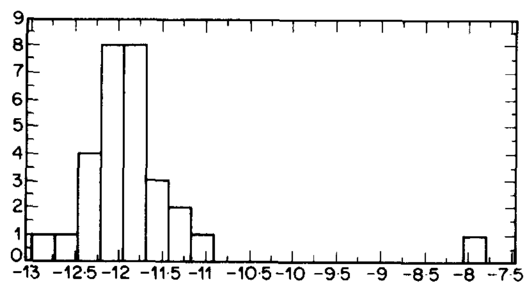

Figure 3. Histogram of feature 3 from tool fracture. Skewness = 3·40, kurtosis = 13·70.

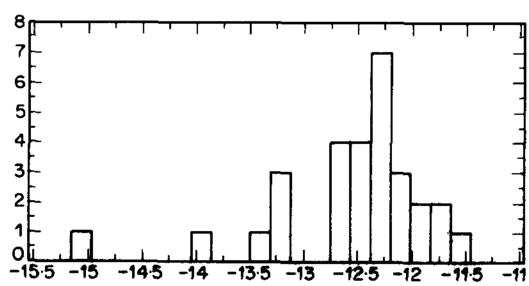

Figure 4. Histogram of feature 4 from tool fracture. Skewness = −1·75, kurtosis = 3·97

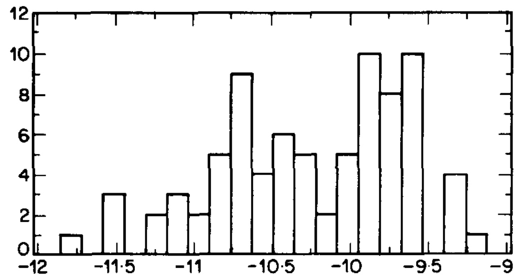

Figure 5. Histogram of feature 3 from sharp tool. Skewness = −0·42, kurtosis = −0·58.

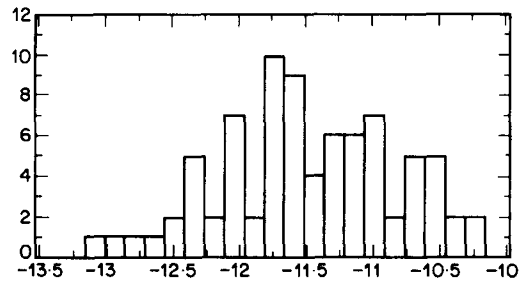

Figure 6. Histogram of feature 4 from sharp tool. Skewness = −0·21, kurtosis = −0·43.

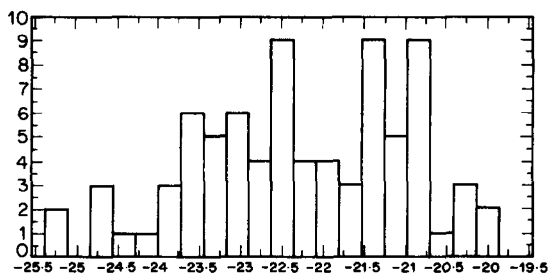

Figure 7. Histogram of feature 39 from worn tool. Skewness = −0·29, kurtosis = −0·67.

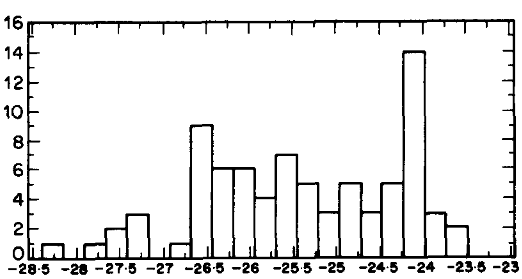

Figure 8. Histogram of feature 44 from worn tool. Skewness = −0·32, kurtosis = −0·77.

Sato [22] pointed out that large skewness and kurtosis have important effects on the performance of the linear and quadratic discriminant functions.

The above results indicate that the standard linear or quadratic normal theory discriminant analysis do not produce optimum results for tool health monitoring using frequency domain AE signals. This may be due to the fact that the generated acoustic emission signals are mixed with various corrupting signals such as instrumentation noise, digitisation noise, vibration, etc. [23, 24]. We should therefore find a method of separating and isolating these signals to distinct components, and use those components for classification which represent the acoustic emission. The principal component analysis is a useful technique for separating data into distinct parts, since it performs an orthogonal transformation of the axes which characterise the data with each axis representing a distinct component. The concept of principal components was discussed in the preceding section. We applied this technique to our data and obtained the first six principal components for each of the three data sets.

Table 7 gives the cumulative percentage of total variance explained by the first six principal components of the data set 1 for each of the four states. Tables 8 and 9 give the corresponding cumulative percentages for data sets 2 and 3 respectively. Notice that for these two sets no data on the "tool fracture" state were available. It can be seen in the first column of Table 7 that the first principal component accounts for a large portion of the total variability in these data (between 49 and 83%). Also, these six principal components, together, account for up to 91% of the total variance (sum of the elements along the diagonal of the variance covariance matrix). This means that instead of the 46 features we can use these six components, thereby reducing the problem size by 40 dimensions, but only lose up to 24% of the total variance. Furthermore, these components are orthogonal to each other (correlation between each pair is zero). This means that

TABLE 7

Cumulative percentage of variance explained by the first six principal components of the 46 variables in data set 1

State	Cumulative % of variance explained, Set 1					
	1	2	3	4	5	6
Chip noise	48·7	59·7	66·3	72·1	74·4	76·4
Tool fracture	52·2	63·8	72·4	76·0	79·5	82·4
Sharp tool	83·1	85·7	88·0	89·2	90·0	90·7
Worn tool	81·2	86·8	88·4	89·5	90·3	91·0

TABLE 8

Cumulative percentage of variance explained by the first six principal components of the 46 variables in data set 2

State	Cumulative % of variance explained, set 2					
	1	2	3	4	5	6
Chip noise	62·4	72·2	76·5	79·2	81·5	83·6
Sharp tool	51·2	69·7	73·8	77·5	79·6	81·5
Worn tool	86·2	87·8	88·9	89·9	90·7	91·4

TABLE 9

Cumulative percentage of variance explained by the first six principal components of the 46 variables in data set 3

State	Cumulative % of variance explained, set 3					
	1	2	3	4	5	6
Chip noise	70·1	81·5	84·7	86·2	87·5	88·6
Sharp tool	84·1	87·7	90·0	90·9	91·5	92·1
Worn tool	83·5	87·2	88·7	90·1	90·8	91·5

each component may represent a particular characteristic of the machining process. Similar results were found for the other data sets, as Tables 8 and 9 indicate.

Examination of these principal components revealed interesting results in terms of their descriptive statistics and their discriminatory powers. Table 10 gives the descriptive statistics of the six principal components for set 1. Those of sets 2 and 3 are given in Tables 11 and 12 respectively. As Table 10 indicates, the first principal component for each state has about the same distribution (similar minimum, maximum, mean, standard deviation), but for the remaining principal components, these distributions seem to have

TABLE 10

Descriptive statistics of the six principal components for the four states of set 1

State	Principal component	Descriptive statistics of the principal components, set 1					
		Min.	Max.	Mean	S.D.	Skewness	Kurtosis
Chip noise (74)	1	−125·9	−108·0	−118·1	3·7	0·17	−0·01
	2	−10·5	3·1	0·2	1·8	−3·45	17·16
	3	−49·5	−40·9	−43·5	1·4	−1·34	4·15
	4	−5·5	1·9	−2·5	1·3	0·82	2·16
	5	−5·0	−1·0	−2·9	0·8	0·05	−0·27
	6	0·4	3·8	1·9	0·7	0·10	−0·23
Tool fracture (29)	1	−127·6	−105·6	−119·7	3·6	1·78	7·37
	2	−1·1	10·0	4·2	1·7	0·34	6·30
	3	−6·0	0·6	−1·4	1.4	−1·88	3·77
	4	−3·5	0·2	−1·6	0·9	−0·07	−0·51
	5	8·9	12·6	11·0	0·9	−0·21	−0·40
	6	6·6	9·9	8·2	0·8	−0·07	−0·49
Sharp tool (80)	1	−123·1	−99·9	−112·5	6·2	−0·12	−1·18
	2	20·0	26·0	22·7	1·1	0·39	0·14
	3	16·4	21·5	19·0	1·0	0·09	−0·28
	4	−34·4	−30·5	−32·9	0·7	0·36	0·64
	5	−16·6	1·8	0·5	0·6	−0·38	0·56
	6	−5·4	−2·9	−4·2	0·6	−0·04	−0·35
Worn tool (80)	1	−126·9	−100·8	−112·2	6·3	−0·15	−0·92
	2	−25·7	−17·1	−22·5	6·7	0·30	0·18
	3	41·6	45·6	44·0	0·9	−0·61	0·32
	4	−1·9	2·7	−0·3	0·8	0·60	1·71
	5	−4·0	−0·7	−2·4	0·6	0·26	0·17
	6	−5·1	−1·8	−3·1	0·6	−0·32	0·48

Sample size in parentheses.

TABLE 11

Descriptive statistic of the six principal components for the three states of set 2

State	Principal component	Descriptive statistics of the principal components, set 2					
		Min.	Max.	Mean	S.D.	Skewness	Kurtosis
Chip noise (40)	1	−122·2	−102·7	−113·7	4·4	0·43	−0·10
	2	−34·7	−25·2	−31·2	1·7	1·62	4·16
	3	−43·6	−37·0	−40·3	1·2	−0·12	1·40
	4	1·6	5·7	3·0	0·9	0·88	0·45
	5	1·0	5·5	3·3	0·8	0·21	1·2
	6	−16·3	−12·7	−14·0	0·8	−0·62	0·54
Sharp tool (40)	1	−109·3	−94·7	−101·8	3·2	−0·31	0·43
	2	−3·8	10·5	0·4	1·9	3·32	16·7
	3	−39·0	−34·2	−36·3	0·9	−0·53	0·96
	4	−10·8	−6·6	−8·2	0·9	−0·46	0·42
	5	−1·5	1·2	−0·4	0·6	0·50	−0·19
	6	−8·6	−5·8	−7·3	0·6	0·09	−0·42
Worn tool (60)	1	−120·4	−95·4	−107·8	6·5	−0·54	−0·88
	2	−13·2	−9·3	−11·3	0·9	−0·23	−0·30
	3	−34·5	−30·9	−32·9	0·7	0·44	0·06
	4	−26·9	−24·0	−25·5	0·7	0·22	−0·51
	5	−3·9	−1·0	−2·4	0·6	−0·16	−0·13
	6	25·0	27·5	26·3	0·6	−0·02	−0·63

Sample size in parentheses.

TABLE 12

Descriptive statistics of the six principal components for the three states of set 3

State	Principal component	Descriptive statistics of the principal components, set 3					
		Min.	Max.	Mean	S.D.	Skewness	Kurtosis
Chip noise (70)	1	−131·2	−108·0	−121·3	5·8	0·41	−0·71
	2	−38·4	−22·9	−28·1	2·4	−1·12	3·98
	3	−50·2	−41·2	−43·4	1·2	−2·44	11·75
	4	3·6	10·0	8·6	0·8	−3·02	16·13
	5	13·5	−8·0	−10·4	0·8	−0·32	2·79
	6	−0·9	2·8	1·0	0·7	−0·28	0·39
Sharp tool (80)	1	−129·7	−99·3	−113·1	7·2	−0·23	−0·80
	2	13·9	20·8	17·4	1·5	0·32	−0·27
	3	−31·9	−25·8	−29·7	1·2	0·67	1·23
	4	−26·4	−22·8	−24·3	0·7	−0·62	0·27
	5	−10·7	−7·4	−8·9	0·6	−0·30	0·03
	6	−1·6	1·5	−0·3	0·6	0·28	0·37
Worn tool (80)	1	−124·8	−94·0	−107·9	6·9	−0·48	−0·50
	2	−28·1	−18·5	−21·4	1·4	−1·19	4·31
	3	−50·5	−45·6	−47·1	0·9	−1·20	1·96
	4	7·3	12·0	12·0	0·9	−0·30	0·10
	5	6·2	−2·7	−4·3	0·6	0·01	0·73
	6	0·6	3·8	2·5	0·6	−0·24	0·07

Sample size in parentheses.

completely different locations (means) in each state, as well as other differences. More importantly, by comparing Tables 10-12, the distribution of the first principal component seems to be the same, irrespective of the cutting conditions of the process. This means that the first principal component, despite its importance in terms of amount of variance it explains, has no discriminatory power. It also indicates that the first principal component is describing a particular characteristic of the system which is present irrespective of the condition of the cutting tool or the cutting conditions of the process.

Figures 9-11 give three plots of α_{ij} vs. j (one from each data set, 1, 2, and 3), for $j = 1, 2, \ldots, 46$, where $\boldsymbol{\alpha}_1' = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{146})$ are optimally calculated weights which maximise the variance of the first principal component Y_1, and

$$Y_1 = \sum_{j=1}^{46} \alpha_{1k} X_j$$

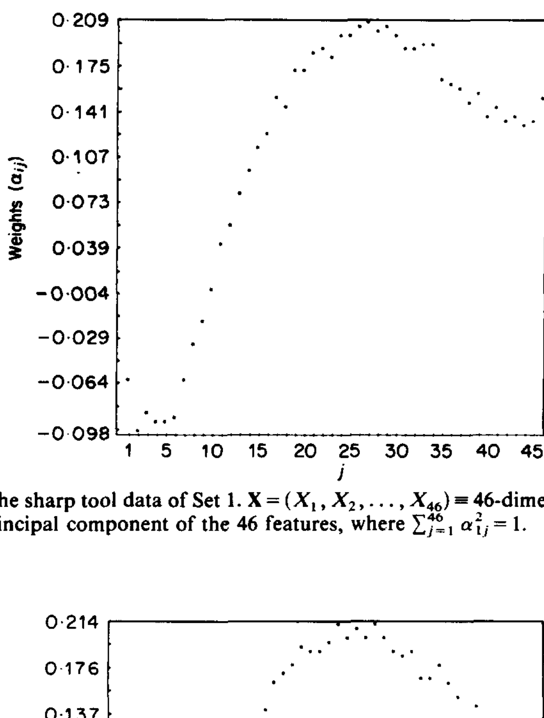

Figure 9. α_{1j} vs. j for the sharp tool data of Set 1. $\mathbf{X} = (X_1, X_2, \ldots, X_{46}) \equiv$ 46-dimensional vector of features. $Y_1 = \sum_{j=1}^{46} \alpha_{1j} X_j \equiv$ first principal component of the 46 features, where $\sum_{j=1}^{46} \alpha_{1j}^2 = 1$.

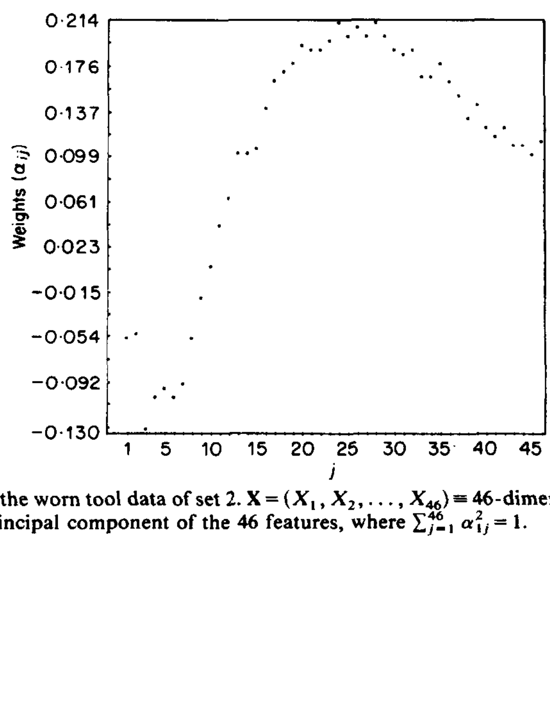

Figure 10. α_{1j} vs. j for the worn tool data of set 2. $\mathbf{X} = (X_1, X_2, \ldots, X_{46}) \equiv$ 46-dimensional vector of features. $Y_1 = \sum_{j=1}^{46} \alpha_{1j} X_j \equiv$ first principal component of the 46 features, where $\sum_{j=1}^{46} \alpha_{1j}^2 = 1$.

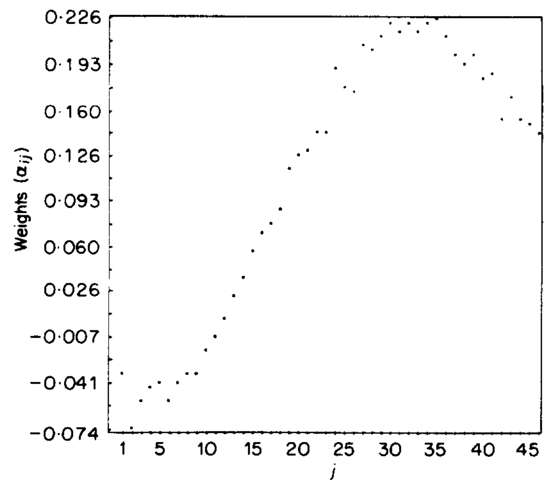

Figure 11. α_{1j} vs. j for the chip noise data of set 3. $\mathbf{X} = (X_1, X_2, \ldots, X_{46}) \equiv 46$ dimensional vector of features. $Y_1 = \sum_{j=1}^{46} \alpha_{1j} X_j \equiv$ first principal component of the 46 features, where $\sum_{j=1}^{46} \alpha_{1j}^2 = 1$.

(no apparent pattern was detected for any other α_{kj} vs. j for $k = 2, 3, \ldots, 6$). As these figures indicate, the weights are very small (less than 0·1) for small values of j. Now since j represents the features at different frequencies, this means that the features at frequencies less than 400 kHz or so contribute very little to the first principal component. This contribution seems to be largest in the frequency range of 400–800 kHz. Therefore, the first principal component must represent a particular characteristic of the system which is unaffected by the cutting conditions or the state of the cutting tool (e.g., electronics noise, digitisation noise, instrumentation noise, etc.). This is further supported by the fact that, discarding this component and performing linear discriminant analysis on the next five components, produce 100% correct classification for all three data sets. Table 13 gives the results of this analysis for data set 1. A comparison of this table with those of 1 and 2 reveals the improvement in the results obtained. It must be pointed out that in this paper we have resubstituted the same data points that were used to obtain the discriminant functions for validation (calculation of percentage correct classification). This is because we believe that classification techniques should be reliable enough to

TABLE 13

Number of vectors classified into a population. Bold entries are percentage of correct classification

	Classified population			
True population	Chip noise	Tool fracture	Sharp tool	Worn tool
Chip noise	74 **100**	0	0	0
Tool fracture	0	29 **100**	0	0
Sharp tool	0	0	80 **100**	0
Worn tool	0	0	0	80 **100**

classify correctly the same data points that were used to develop them. Also, since we did not have large data sets, cross validation of our results was not possible. The results may therefore appear optimistic. However, we believe that our results will be repeated with fresh data since the distribution of the principal components are very different for each state.

A further examination of Table 10 reveals an interesting result. Consider, for example, the third principal component. The range of numbers in this component are from −49·5 to −40·9 for chip noise, −6 to 0·6 for tool fracture, 16·4 to 21·5 for a sharp tool, and 41·6 to 45·6 for a worn tool. Clearly, there is no intersection between these four ranges of numbers. Therefore, it is possible to use only this component to classify signals into one of the four states by a simple inspection. It must, however, be noted that this inspection method does not always produce such good results. This is evident in Table 12 where no unique component has the characteristics of the third component in Table 10.

5. CONCLUSIONS

The use of AE methodology as a sensing technique for monitoring the condition of a cutting tool in metal cutting is examined in the context of statistical pattern recognition (linear or quadratic discriminant functions), as a tool for signal decomposition. It was found that the spectral power of acoustic emission signals are cross and serially correlated, and that they do not necessarily satisfy the required normality assumption. Departure from these major assumptions caused the classifier's performance to be less than optimum.

The principal component analysis of data in each state indicated that the spectral power of acoustic emission signals are contaminated by noise in the system. This noise was represented by the first principal component which had no discriminatory power. The amount of noise in the system was found to be maximum in the frequency range of 400 to 800 kHz. Separating the noise forms AE produced components with strong discriminatory power, so much so that only a few principal components produced 100% correct classification (under resubstitution of data). These results were particularly promising since they were repeated under three different cutting conditions. We suggest that further experiments under various cutting conditions can shed more light on the universality of these results.

ACKNOWLEDGEMENT

We would like to thank Dr Erdal Emel for providing the data used in this paper. The support of this research by the National Science Foundation under Grant Number DMC-8607198 and by the University of Michigan Industrial Consortium on Diagnostic Sensing and Control for Metal Cutting, as well as the Industrial Technology Institute, is greatly appreciated.

REFERENCES

1. Y. KAKINO 1980 *ASM Proceedings of an International Conference on Cutting Tool Materials*, In-process detection of tool breakage by monitoring acoustic emission. pp. 25–39.
2. I. INASAKI and S. YONETSU 1981 *Proceedings of the International Machine Tool Design and Research Conference*, **22**, pp. 261–268. In-process detection of cutting tool damage by acoustic emission measurement.
3. E. KANNATEY-ASIBU, JR and D. A. DORNFELD 1982 *Wear* **76**. A study of tool wear using statistical analysis of metal cutting acoustic emission.

4. E. N. DIEI and D. A. DORNFELD 1985 *Sensors and Controls for Manufacturing* (E. Kannatey-Asibu, Jr, A. G. Ulsoy and R. Komanduri, editors pp. 33-39. ASME Publication, PED-Vol. 18. A model of tool fracture generated acoustic emission during machining.
5. E. KANNATEY-ASIBU JR and D. A. DORNFELD 1981 *Journal of Engineering for Industry*, **103**, 330-340. Quantitative relationships for acoustic emission from orthogonal metal cutting.
6. E. KANNETAY-ASIBU, JR 1982 *Proceedings of the Tenth NAMRC*, pp. 487-492. On the application of the pattern recognition method to manufacturing process monitoring.
7. D. A. DORNFELD and C. S. PAN 1985 *Proceedings of the Thirteenth NAMRC*, Berkely, CA, pp. 299-303. Determination of chip forming states using linear discriminant function technique with acoustic emission.
8. E. KANNATEY-ASIBU, JR and E. EMEL 1987 *Mechanical Systems and Signals Processing* **1**, 333-347. Linear discriminant function analysis of acoustic emission signals for cutting tool monitoring.
9. E. EMEL and E. KANNATEY-ASIBU, JR 1988 *ASME Journal of Engineering Industry* **110**, 137-145. Tool failure monitoring in turning by pattern recognition analysis of AE signals.
10. R. A. FISHER 1936 *Annals of Eugenics* **7**, 179-188. The use of multiple measurements in toxonomic problems.
11. C. A. B. SMITH 1947 *Annals of Eugenics* **13**, 272-282. Some examples of discrimination.
12. D. F. MORRISON 1967 *Multivariate Statistical Method.* New York: McGraw-Hill.
13. S. J. PRESS 1972 *Applied Multivariate Analysis*, New York: Holt, Rinehart and Winston.
14. K. FUKUNAGA 1972 *Introduction to Statistical Pattern Recognition*, New York: Academic Press.
15. A. A. AFIFI and S. P. AZEN 1972 *Statistical Analysis. A Computer Oriented Approach.* New York: Academic Press.
16. R. I. JENNRICH 1977 *Statistical Method for Digital Computers*, Vol. III. New York: John Wiley. Stepwise discriminant analysis.
17. T. W. ANDERSON 1984 *An Introduction to Multivariate Statistical Analysis*, Second Edition. New York: John Wiley.
18. C. K. BAYNE and J. J. BEAUCHAMP 1984 *Communication in Statistics-Simulation and Computation* **13**(5), 669-682. Misclassification probabilities for second-order discriminant functions used to classify bivariate normal populations.
19. J. J. DAUDIN 1986 *Biometrics* **42**, pp. 473-481. Selection of variables in mixed variable discriminant analysis.
20. N. R. DRAPER and H. SMITH 1981 *Applied Regression Analysis*, Second Edition. New York: John Wiley.
21. D. H. FOLEY 1972 *IEEE Transactions on Information Theory* **IT-18**, 618-626. Consideration of sample and feature size.
22. H. NAKANISHI and Y. SATO 1985 *Communication in Statistics-Theory and Methods* **14**(5), 1181-1200. The performance of the linear and quadratic discriminant functions of three types of non-normal distributions.
23. L. J. GRAHAM and G. A. ALERS 1975 *Monitoring Structural Integrity by Acoustic Emission, ASTM STP 571*, pp. 11-39. Acoustic emission in frequency domain.
24. P. HORVATH and F. J. COOK 1981 *Quantitative Non-destructive Evaluation* **1**, 463-473. Establishing signal processing and pattern recognition techniques for inflight discrimination between crack-growth acoustic emission and other acoustic waveforms.

APPENDIX A. STEPWISE DISCRIMINANT ANALYSIS

Here we describe the stepwise method used for feature selection in our analysis of data. This method is described by Afifi and Azen [15].

Suppose $\mathbf{Y} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_Y$ is the k-dimensional vector of means, and $\boldsymbol{\Sigma}$ is the $k \times k$ variance-covariance matrix. Partition $\mathbf{Y}$, $\boldsymbol{\mu}_Y$, and $\boldsymbol{\Sigma}$ as:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{T} \end{pmatrix}, \qquad \boldsymbol{\mu}_Y = \begin{pmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{\mu}_T \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{A.1}$$

Where $\mathbf{Y}: k \times 1$, $\mathbf{Z}: u \times 1$, $\mathbf{T}: v \times 1$, $\Sigma_{11}: u \times u$, $\Sigma_{22}: v \times v$, $\Sigma_{12}: u \times v$, $\Sigma_{21}: v \times u$, and $u + v = k$. Now, if $\Sigma_{11} > 0$, and $\Sigma_{22} > 0$, then $\mathbf{Z} \sim N(\boldsymbol{\mu}_Z, \Sigma_{11})$, and $\mathbf{T} \sim N(\boldsymbol{\mu}_T, \Sigma_{22})$. It follows that the conditional distribution of $\mathbf{Z}$ given $\mathbf{T} = \mathbf{t}$ is also multivariate normal with mean vector a linear function of **t**, and variance-covariance matrix independent of **t**; that is

$$(\mathbf{Z}|\mathbf{T}=\mathbf{t}) \sim N[\boldsymbol{\mu}_Z + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{t}-\boldsymbol{\mu}_T), \Sigma_{11.2}]. \tag{A.2}$$

where $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Let $\mathbf{X}_j^i = (X_{j1}^i, X_{j2}^i, \ldots, X_{jk}^i)'$ be the j-th random vector from population i, and $\boldsymbol{\mu}^i = (\mu_1^i, \mu_2^i, \ldots, \mu_k^i)'$ be the mean vector of population i, (for $i = 1, 2, \ldots, s$). Let V_r denote the r-th feature; the procedure for stepwise discrimination is as follows.

First the *F to enter* along with its degrees of freedom is computed for each V_r, $r = 1, 2, \ldots, k$. This F to enter is the one-way analysis of variance F statistic for testing the null hypothesis H_0: $\mu_r^1 = \mu_r^2 = \ , \ldots, \mu_r^s$. That is,

$$F = \frac{\dfrac{\sum_{i=1}^{s} n^i(\bar{X}_r^i - \bar{X}_r)^2}{(s-1)}}{\dfrac{\sum_{i=1}^{s}\sum_{j=1}^{n^i}(X_{jr}^i - \bar{X}_r^i)^2}{(n-s)}}, \qquad (r = 1, 2, \ldots, k). \tag{A.3}$$

where n^i is the sample size for state i,

$$n = \sum_{i=1}^{s} n^i,$$

and $\bar{X}_r^i$ is the r-th sample estimate of μ_r^i, and $\bar{X}_r$ is the r-th sample estimate of feature mean, μ_r say. Now if all F to enter (F_r) are less than a prescribed inclusion level, called the *F to include* (the F value corresponding to α level of significance), the process terminates and we conclude that no feature significantly discriminates between the populations. Otherwise the feature V_{r1} having the largest F to enter is selected as the first feature. We then use (A.2) to find the conditional distribution of the remaining $k-1$ features conditioned on the feature selected, and use them to calculate the F to enter and its degrees of freedom for each feature not entered. This tests the null hypothesis H_0: $\mu_{r\cdot r_1}^1 = \cdots = \mu_{r\cdot r_1}^s$, where $\mu_{r\cdot r_1}^i$ is the mean of the conditional distribution in population i of V_r given V_{r_1}, $i = 1, 2, \ldots, s$, $r = 1, 2, \ldots, k$, $r \neq r_1$. If all the F to enter values of the remaining features are less than F to include, then we stop the process and conclude that only V_{r_1} discriminates, between the populations. Otherwise the feature V_{r_2} with the maximum value of F to enter is selected. Note that in the subsequent calculation of F statistics for any feature, we must take into account the fact that although the distributions of V_rs remain normal, the parametrisation changes because these distribution are now conditional. Therefore a different expression from (A.3) must be used to obtain these F values.

Now, we use (A.2) again, to obtain the conditional distributions of V_{r_1} given V_{r_2}, and V_{r_2} given V_{r_1}, as well as the conditional distributions of all the features not entered. Then we calculate two *F to remove* values and their degrees of freedom for V_{r_1} and V_{r_2}. These test null hypothesis H_0: $\mu_{r_1\cdot r_2}^1 = \cdots = \mu_{r_1\cdot r_2}^s$ and H_0: $\mu_{r_2\cdot r_1}^1 = \cdots = \mu_{r_2\cdot r_1}^s$, respectively. If either of the F to remove values are less than F to include value, then the feature with the smallest F to remove value is removed from the list and added to the excluded features. Otherwise the F to enter and its degree of freedom for each feature not entered are calculated. This tests the hypothesis H_0: $\mu_{r\cdot r_1 r_2}^1 = \cdots = \mu_{r\cdot r_1 r_2}^s$, where $\mu_{r\cdot r_1 r_2}^i$ is the mean of the conditional distribution in population i of V_r given V_{r_1} and V_{r_2}, $i = 1, 2, \ldots, s$, $r = 1, 2, \ldots, k$, $r \neq r_1$ or r_2. If all F to enter values are less than the F to include value, then the process stops, otherwise the feature with the largest F to enter is selected as the third feature. We continue this process until either the inclusion and the exclusion tests fail, or all the features are included.

The procedure described above is the forward selection discriminant function. The backward elimination procedure follows similar steps except that in that case we start with all the features included and try to exclude one feature at a time.

APPENDIX B. NOMENCLATURE

$\mathbf{x}_j^i$	j-th training sample from state i
$\mathbf{x}$	single vector of observation to be classified into one of s states, $(s \geq 2)$
Π_i	population (state) i
$f_i(\mathbf{x} \mid \boldsymbol{\Omega}_i)$	multivariate probability density function of the population Π_i with known parameter matrix $\boldsymbol{\Omega}_i$
C_{ij}	cost of misclassifying x into Π_i when it actually belongs to Π_j
p_i	*a priori* probability of x belonging to Π_i
μ_i	k-dimensional vector of the means for population Π_i
Σ	$(k \times k)$ variance-covariance matrix for each population under the assumption of equality of variance-covariance matrices
Σ_i	$(k \times k)$ variance-covariance matrix for population Π_i
$\mathbf{X}$	k-dimensional random vector of features
X_i	i-th component of the random vector $\mathbf{X}$
Z_i	i-th non-normalized feature
Y_i	i-th principal component (eigenvalue)
$\boldsymbol{\alpha}_i$	i-th eigenvector