Evaluation and Comparison of the Principal Component Analysis (PCA) and Isometric Feature Mapping (Isomap) Techniques on Gas Turbine Engine Data

Uduak A.Umoh+, Imoh J.Eyoh+ and Jeremiah E. Eyoh*

+Department of Computer Science,
University of Uyo, Uyo, Akwa Ibom State, Nigeria

*Department of Turbo Machinery (Reliability and Maintenance),
Exxon Mobile, QIT, Eket, Akwa Ibom State, Nigeria

Abstract

This paper performs a comparative analysis of the results of PCA and ISOMAP for the purpose of reducing or eliminating erratic failure of the Gas Turbine Engine (GTE) system. We employ Nearest-neighbour classification for GTE fault diagnosis and M-fold cross validation to test the performance of our models. Comparison evaluation of performance indicates that, with PCA, 80% of good GTE is classified as good GTE, 77% of the average GTE is classified as average GTE and 67.6% of bad GTE is classified as bad GTE. With ISOMAP, 67% of good GTE is classified as good GTE, 70.8% of the average GTE is classified as average GTE and 81% of bad GTE is classified as bad GTE. PCA produces 26% error rate with nearest neighbour classification and 17% error rate with M-fold cross validation. While ISOMAP produces 35% error rate with nearest neighbour classification, and 26.5% error rate with M-fold cross validation. Results indicate that PCA is more effective in analyzing the GTE data set, giving the best classification for fault diagnosis. This enhances the reliability of the turbine engine during wear out phase, through predictive maintenance strategies.

1.0 Introduction

Maintenance of complex engineering systems such as GTE has posed a serious challenge to systems engineers, as this affects the GTE subsystems and entire system reliability and performance. Monitoring the health of a system is part of the predictive maintenance approach that seeks to extend the reliability and life of the system. Principal Component Analysis (PCA) and Isomeric Feature Mapping (ISOMAP) are dimensionality reduction techniques employed to transform a high-dimensional data space to a low-dimensional space with information and local structure of the data set being preserved as much as possible. Principal Components Analysis, PCA has been proven to be good in transforming high dimensional linear data set to lower dimensional space, with much lose of information contained in the original data. Applying linear techniques of dimensionality reduction to a nonlinear data such as GTE data set is sure not going to give a much success story as when linear techniques are applied to a linear data set. Isometric Feature Mapping, ISOMAP is a nonlinear dimensionality reduction method that maps from the high dimensional space to a low-dimensional Euclidean feature space. Also, the projected observation with reduced dimensions preserves as much as possible the intrinsic metric structure of the observation [9]. In this work, we evaluate and compare analyzed signal characteristics and extracted features based on PCA and ISOMAP data-based analysis techniques. We explore Matlab and C++ programming tools for the implementation. .

2.0 Literature Review

Gas turbine engines have proven to be very efficient and are widely used in many industrial and engineering systems. They are used in systems such as Aircrafts, Electrical power generation Systems, Trains, Marine

vessels, as drivers to industrial equipment such as high capacity compressors and pumps. In most cases, areas of application of gas turbine engines are safety critical which require very high reliability and availability of these systems. To maintain high system reliability and availability, critical system parameter variables such as engine vibration, bearing temperature, lube oil pressure, etc, must be continuously monitored for prompt detection of deviation from normal operation values. To design a system for high reliability means, increasing the cost of the system and its complexity [4]. More so, monitoring, control and protection subsystems of the Gas Turbine Engines further add more cost and complexity to the overall system. The application of a classical maintenance approaches has been proven over the years, to be unsuitable for engineering systems such as Gas turbine engines [7] [6]. The health state of a GTE is determined by its functional state or characteristics of the parameter variables. Depending on the characteristics of these parameter variables, the GTE health state can be in a particular state [7]. In PCA, data can be transformed to a new set of coordinates or variables that are a linear combination of the original variables [8]. Researchers have developed various systems' health condition monitoring strategies in which the state of the system is expected to operate under designed operating conditions. Thus, condition based predictive maintenance has significant cost reduction implications [7].

The health state of a GTE is determined by its functional state or characteristics of the parameter variables. Depending on the characteristics of these parameter variables, the GTE health state can be in any of the following states [7]. Basic fault models are due [6] [7] [1] [10]. Most of the turbine engine diagnostic and prognostic systems are based on model-based and, or knowledge-based approaches, in which artificial neural networks techniques are used. Some of the disadvantages of this approach are that it adds more cost to the system life cycle and further physical and architectural structure of this complex system greatly reduces the reliability of the entire system [5].

3. Research Methodology

Data-based health condition monitoring of GTE employs dimensionality reduction techniques to analyze the systems parameter variable data in order to extract hidden features which are useful in fault detection and diagnosis. This is achieved by exploring different data classification techniques for fault diagnosis. We first applied PCA to the EngData training set to project the high dimensional nonlinear data to a low-dimensional subspace [2]. The low dimensional data obtained shows that over 90 % of the information contained in the original high dimensional data is found in just the first ten principal component of the analysis. The ISOMAP technique, which is nonlinear method, is also applied to the data and the reduced dimensional data is further analyzed [3]. We evaluate, and compare the results of PCA and ISOMAP on the training data, using nearest-neighbour classification and cross validation techniques.

4.0 Performance Evaluation of PCA and ISOMAP

a. PCA

Though many techniques are available to test the performance of the data model developed using PCA, its performance is in a way, dependent on the nature of the data set being analyzed. PCA will perform much better analysis if the data set is normally distributed around the median. Before the PCA is applied on the data, it is first of all pre-processed to standardize the data for better results. The data was standardized to have zero mean, unit standard deviation and unity variance [2].

The analysis of the GTE training data set produces 15 PCs, Eigen values as shown on Table 1. The low-dimensional basis based on the principal components minimizes the reconstruction error, which is given by:

$$e = \| x - \hat{x} \| \quad (1)$$

This error e can be rewritten as;

$$e = \frac{1}{2} \sum_{i=K+1}^{N} \lambda_i$$

(2)

Where N = 98; K = 10, 11, 12, 13, 14, 15.

Throughout the analysis of this work, K is chosen to be 12.

Calculating error when k = 10 is as follows;

$e = ½ (98 – 88.8294)$

$e = 4.5853$

For K = 12;

$e = ½ (98 – 90.8989)$

$e = 3.551$

The residual error is relatively small as can be seen from the calculation when K = 12, as used in this analysis. This also indicates that PCA has been able to analyze the data comparatively well, though the GTE data is nonlinear and the distribution of the data is not perfectly around the median.

The classification of GTE classes is shown in Table 2. Here 80% of good GTE are classified as good GTE, 77% of the average GTE was classified as average GTE and 67.6% of bad GTE are classified as bad GTE. No bad GTE are classified as good GTE and no good GTE are classified as bad GTE. This achievement by PCA is vey commendable as it is very paramount in safety critical systems such as GTE.

We employ cross validation method to test the performance of the data-based model developed using PCA.

Table 1 showing 15 PCs, Eigen values

Principal Components (PCs)	Rival (latent)	Camus of Rival	Colum of Rival (%)
PC#1	36.6533	36.6533	37.4013
PC#2	13.9509	50.6042	51.6369
PC#3	8.5086	59.1128	60.3191
PC#4	7.2647	66.3774	67.7321
PC#5	6.4723	72.8498	74.3365
PC#6	4.8586	77.7083	79.2942
PC#7	3.7902	81.4985	83.1617
PC#8	3.2723	84.7708	86.5008
PC#9	2.3949	87.1657	88.9445
PC#10	1.6638	88.8294	90.6423
PC#11	1.2393	90.0688	91.9069
PC#12	0.9210	90.9898	92.8467
PC#13	0.8787	91.8685	93.7434
PC#14	0.7817	92.6502	94.5410
PC#15	0.7240	93.3743	95.2799

Table 2 Percentage of classification result with PCA

	PREDICTED CLASSIFICATION			
		Good GTE (class 1)	Average GTE (class 2)	Bad GTE (class 3)
KNOWN CLASSIFICATION	Good GTE (class 1)	12 (80%)	3	0
	Average GTE (class 2)	11	37 (77%)	0
	Bad GTE (class 3)	0	12	25 (67.6%)

Total number of test cases = 100

Total number of Good GTE = 15; percentage of good GTE classification = 80%

Total number of Average GTE = 48; percentage of average GTE classification = 77%

Total number of Bad GTE = 37; percentage of bad GTE classification = 67.6%

Table 2 shows that 12 good GTE out of 15 were classified as good GTE, 3 good GTE out of 15 were classified as average GTE and no good GTE was classified as bad

GTE. Also, from the table, it can be seen that no bad GTE was classified as good GTE. This is very reasonable for safety critical system such as GTE.

Despite the fact that the GTE data set is noisy and nonlinear, the result from PCA is very impressive because of the following achievements: The residual error is reasonably small. The high dimensional data space is projected to low-dimensional subspace without much lost of information contained in the original data. 80% of good GTE was classified as good GTE, 77% of the average GTE is classified as average GTE and 67.6% of bad GTE is classified as bad GTE. No bad GTE is classified as good GTE and no good GTE was classified as bad GTE. This achievement by PCA is vey commendable as it is very paramount in safety critical systems such as GTE. The cross validation of the training model of the data base also recorded an impressive result; that is 83% of the training data model is classified while only 17% of the training data model is misclassified. Therefore PCA has been able to detect 80% of the good GTE, 77% of the average GTE and 67.6% of the bad GTE, though PCA is not always an optimal dimensionality reduction procedure for classification purposes.

b. ISOMAP

As stated in the case of PCA, the effectiveness or performance of ISOMAP depends on the nature of the data set. ISOMAP give a better result for manifolds of moderate dimensionality, since the estimates of manifold distance for a given graph size degrades as the dimensionality increases. The data set whose classes or features are sparsely distributed without defined uniformity, such as engineering data obtained from practical systems, may not give a better result when analyzed using ISOMAP.

The performance of ISOMAP can be evaluated using nearest neighbour classification of the test data set and cross validation of the training data set. In this project work, the performance of ISOMAP is seriously affected by the choice of neighbourhood factor, k for the algorithm. This may be due to the nature of the data set. The neighbourhood factor above 8 gives a comparatively bad result while a value of k below 7 leads to discontinuity and the Y.index (which contains the indices of the points embedded), produced is less than 98 indices. When k = 6 or 5 was used, the Y.index was 35 and k = 3 gave much lower indices. This made the ISOMAP analysis limited to only neighbourhood factor values. That is 7 or 8. Table 3 presents percentage of classification result with ISOMAP when k = 7. Table 4 shows percentage of classification result with ISOMAP when K = 8

Table 3 Percentage of classification result with ISOMAP when K = 7

		PREDICTED CLASSIFICATION		
		Good GTE (class 1)	Average GTE (class 2)	Bad GTE (class 3)
KNOWN CLASSIFICATION	Good GTE (class 1)	0 (0%)	14	1
	Average GTE (class 2)	0	33(68.75%)	15
	Bad GTE (class 3)	0	5	32 (86%)

Total number of test cases = 100
Total number of Good GTE = 15; percentage of good GTE classification = 0%
Total number of Average GTE = 48; percentage of average GTE classification = 68.75%
Total number of Bad GTE = 37; percentage of bad GTE classification = 86%

Table 4 Percentage of classification result with ISOMAP when K = 8

	PREDICTED CLASSIFICATION			
		Good GTE (class 1)	Average GTE (class 2)	Bad GTE (class 3)
KNOWN CLASSIFICATION	Good GTE (class 1)	1 (6.7%)	14	0
	Average GTE (class 2)	3	34 (70.8%)	11
	Bad GTE (class 3)	0	7	30 (81%)

Total number of test cases = 100
Total number of Good GTE = 15; percentage of good GTE classification = 6.7%
Total number of Average GTE = 48; percentage of average GTE classification = 70.8%
Total number of Bad GTE = 37; percentage of bad GTE classification = 81%

With K = 8, (though, even number is not a good choice for K), the classification gives a slightly good result as no good GTE was classified as bad GTE and no bad GTE is classified as good GTE. It is still not generally good approach because 14 out of 15 good GTE are classified as average GTE.

Figure 1 presents Residual Variance vs Isomap dimensionality with K = 7. ISOMAP technique applied to GTE data set is able to correctly recognize its intrinsic three-dimensionality as indicated by the arrow in Figure 1.

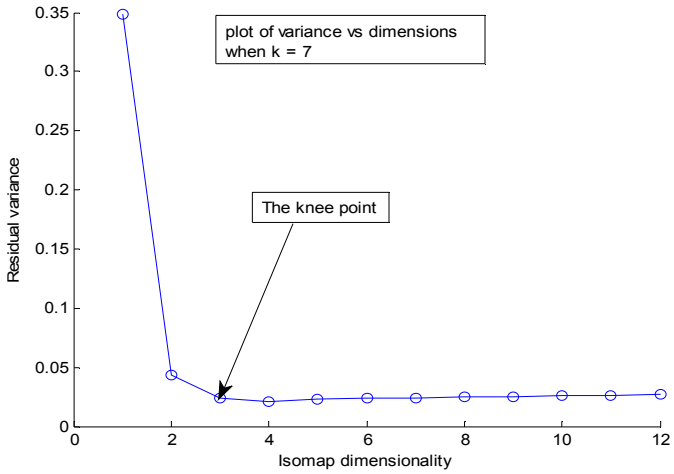

Fig. 1: Residual Variance vs Isomap dimensionality with K = 7

Other achievement by ISOMAP of the GTE data set includes the following:
ISOMAP generated a two-dimensional embedding with a neighbourhood graph which gives a visual information or characteristic of the data set. This is helpful in studying the geometrical structure of the GTE data. Also, the ISOMAP analysis preserves information contained in the data and the local structure of the data.

With k = 8, ISOMAP is achieve 6.7% of good GTE is classified as good GTE, 70.8% of the average GTE is classified as average GTE and 81% of bad GTE is classified as bad GTE. No bad GTE is classified as good GTE and no good GTE is classified as bad GTE. This achievement is reasonably good as no it is important in safety critical systems such as GTE. But the system availability and productivity is affected as over 93% of good GTE is classified as average GTE.
The cross validation of the training model of the data base using ISOMAP also recorded

an impressive result; that is 73.5% of the training data model was classified while only 26.5% of the training data model was misclassified.

5. Comparison of PCA and ISOMAP Analysis Results.

PCA and ISOMAP are dimensionality reduction techniques employed to transform a high-dimensional data space to a low-dimensional space with information and local structure of the data set being preserved as much as possible. Both techniques use the number of significant Eigen values to estimate the dimensionality.

ISOMAP is a graph-based, spectral, nonlinear method of dimensionality reduction approach with no local optima. It is parametric, non-iterative, polynomial time procedure which guarantees global optimality. PCA is non-parametric, linear method in which the direction of the greatest variance is the eigenvector corresponding to the largest Eigen values of the data set. PCA is guarantee to recover the correct or true structure of the linear manifolds while ISOMAP is guaranteed to recover the correct or true dimensionality and geometrical structure of a large class of non linear manifolds as shown in Figures 4 and 5. The knee point in the Figure 4 indicates the true dimensionality of the manifold, while in Figure 5; the PCA cannot recover the correct dimensionality. In this work, when the two methods are applied on the GTE data set, the results show that PCA best analyzed the data than ISOMAP. Table 5 compares the results obtained from both methods. Thus PCA performance for this analysis is better than ISOMAP. Figure 5 shows comparison evaluation of PCA and ISOMAP performance of the training data using nearest-neighbour classification and cross validation. PCA produced 26% error rate with nearest neighbour classification, and 17% error rate with M-fold cross validation. ISOMAP produced 35% error rate with nearest neighbour classification, and 26.5% error rate with M-fold cross validation.

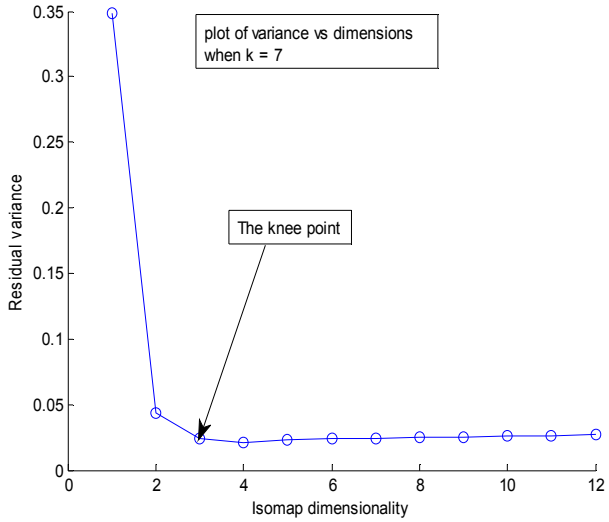

Fig. 4: ISOMAP Plot of variance (Eigen values) vs dimensionality:

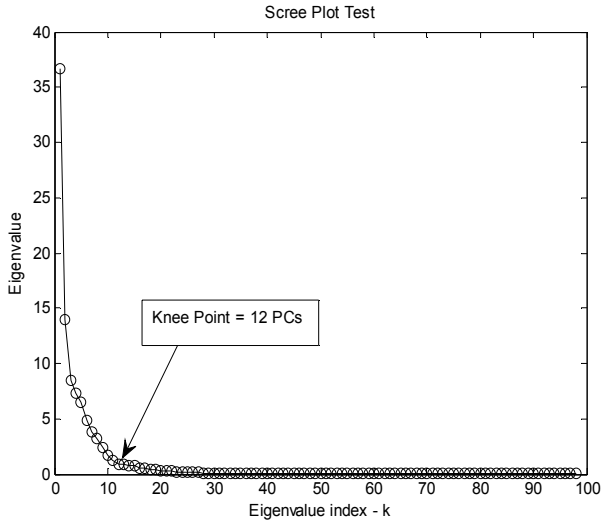

Fig. 5: PCA Plot of variance (Eigen values) vs dimensionality:

Table 5 comparison of PCA and ISOMAP Performance

	PCA Analysis		ISOMAP Analysis	
	Classified	Misclassified	Classified	Misclassified
NN Classify	74%	26%	65%	35%
M-Fold CV	83%	17%	73.5%	26.5%

6.0 Conclusions

Data-based techniques are simple and cost effective method of monitoring the health condition of a system, as part of the predictive maintenance strategy that seeks to improve and extend the reliability and life of the system. ISOMAP and PCA are employed to project the high-dimensional data space to the lower dimensional subspace. The low dimensional data set was analyzed to extract changes in the feature for fault detection and diagnosis. Data classification and visualization are very effective means of discovering characteristics or features encoded in a given data set. The GTE data set was visualized in two-dimension using scatter plot. The data-based model performance evaluation results indicate that PCA is very suitable and more effective in analyzing high-dimensional data such as GTE dataset than ISOMAP, giving the best classification for fault diagnosis. Thus PCA data based technique for health condition monitoring is an effective predictive maintenance strategy which can easily extract unknown or hidden features or geometrical structures of the system parameter variables. These features can be used to detect and diagnose system fault. The weakness of ISOMAP in this project may be due to the sparse nature of the GTE data set.

References

[1] Chiang, L. H., E.L. Russell, and R.D. Braatz. (2001). *Fault Detection and Diagnosis in Industrial Systems*. Springer-Verlag,

[2] Eyoh, J. E., Eyoh, I. J., Umoh, U. A. and Udoh, E. N. (2011a), Health Monitoring of Gas Turbine Engine using Principal Component Analysis Approach. Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 2 (4): 717-723

[3] Eyoh, J. E., Eyoh, I. J., Umoh, U. A. and Umoeka, I. J. (2011b), Health Monitoring of Dimensional Gas Turbine Engine (EngData) using ISOMAP Data-Based Analysis Approach. World Journal of Applied Sciences and Technology (WOJAST) 3(2), 112-119.
[4] Ghoshal, S., Roshan Shrestha, Anindya Ghoshal, Venkatsh Malepati, Somnath Deb, Krishna pattipati and David Kleinman, (1999)"An Integrated Process For System Maintenance, Fault Diagnosis and Support", Invited Paper in Proc. IEEE Aerospace Conf., Aspen, Colorado.
[5] Greitzer, F. L., Lars J. Kangas, Kristine M. Terrones, Melody A. Maynard, Bary W. Wilson, Ronald A. Pawlowski, Daniel R. Sisk and Newton B. Brown, (1999). "Gas Turbine Engine Health Monitoring and Prognostics", Paper presented at the International Society of Logistics (SOLE) 1999 Symposium, Las Vegas, Nevada, August 30 – September 2.
[6] Isermann, R. (2006). "Fault-Diagnosis Systems – An Introduction from Fault Detection to Fault Tolerance", " Springer, Berlin.
[7] Kadirkamanathan, V., (2008) "ACS 6304 – Health Care Monitoring", Department of Automatic Control & Systems Engineering, University of Sheffield, 21 – 25 January.
[8] Martinez, W. L. and Angel R. Martinez, (2004) "Exploratory Data Analysis with MATLAB", (Computer Science and Data Analysis), Chapman & Hall/CRC, 2004 **...**CRC Press.
[9] Tenenbaum, J. B. (1998). "Mapping a Manifold of Perceptual Observations", Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.
[10] Yang, P., Sui-sheng Liu, (2005)"Fault Diagnosis System for Turbo-Generator Set Based on Fuzzy Network", International Journal of Information Technology, Vol. 11 No. 12, 2005, pp. 76-84.