Nonlinear PCA for Transient Monitoring of an Automotive Engine

Xun Wang[1], George W. Irwin[1]*, Geoff McCullough[2], Neil McDowell[2], Uwe Kruger[3]

[1]Intelligent Systems and Control Research Group, [2]Internal Combustion Engines Research Group, Queen's University, Belfast BT9 5AH, UK.
[3]Department of Electrical Engineering, The Petroleum Institute, PO Box 2533, Abu Dhabi, UAE

Abstract: This paper reports on the application of non-linear principal component analysis to the detection of faults in an automotive gasoline engine during transient operation. An auto-associative neural network is trained on experimental data recorded from an identification cycle in which the engine speed and throttle position inputs were varied over a wide range of the operating map at rates similar to those experienced during normal operation. The model shows good generalisation to the New European Drive Cycle, an absence of unwanted false alarms under fault-free engine conditions, and successful detection of air leaks of varying magnitude in the inlet manifold.

Keywords: Nonlinear statistical process control, automotive engine, fault detection and diagnosis, neural networks, dynamic models.

1. INTRODUCTION

The introduction of emissions legislation has added to the complexity of modern automotive engines. Exhaust after treatment devices, such as catalytic converters or particulate filters, are necessary for automotive manufacturers to meet current and proposed emissions legislation. The control of modern engines is highly complex as many systems interact, each playing a significant role in emission reduction or engine performance. In addition, all must function correctly over a wide operational range, including transient conditions.

The detection of faults within the engine environment is covered by additional specific legislation and is commonly known as On-Board Diagnostics (OBD). OBD legislation details which component parts of the engine are to be tested and at what frequency.

The automotive industry currently uses a combination of signal and model-based diagnostic techniques, the latter being based mainly on physical models of the system. However, as the emissions thresholds have reduced, the OBD fault detection thresholds have decreased accordingly, thereby increasing the challenge in engine modelling and monitoring (Stobart, 2003). In addition, the expense of physical model- based OBD, in terms of mathematical complexity and computational intensity, has led to research on more practical alternatives. A number of other techniques for fault detection and diagnosis (FDD) have therefore been investigated and applied to automotive engines over the last decade (Nyberg, 1999; Grimaldi et al., 2001; Kimmich et al., 2005.).

Statistical techniques have also been applied, with principal component analysis (PCA) being one of the most popular. Here, a reduced set of statistically independent score variables are generated for process monitoring. Unfortunately, PCA in its original form is only applicable to linear data. Nonlinear extensions of PCA can take the form of auto-associative neural networks (AANNs) (Kramer, 1992), principal curves, or a combination of both. Because of their conceptual simplicity and close alignment to linear PCA, our previous work has relied on the use of AANNs for detecting and diagnosing automotive engine faults during steady-state operation (Antory et al., 2005, Wang et al., 2008a). Here nonlinear PCA (NLPCA) refers to an AANN model where the score variables are produced in the network's bottleneck layer. Further research has introduced additional statistical analysis to AANN, which increased the sensitivity in detecting minor faults (Wang et al., 2008a).

This contribution reports results from applying NLPCA to experimental data recorded from the intake system during transient operation of an automotive gasoline engine. The important practical significance of this work is that the resulting model proved suitable for more general use with unknown drive cycles as would be the case in actual deployment. Thus, rather than using the same operational cycle for training and testing the model, two completely different test cycles were employed, as will be detailed in the next section.

The paper is organised as follows. An introduction to the experimental petrol engine and the data collection regime is given next in Section 2. This is followed by a brief review of NLPCA in Section 3. Section 4 then presents the experimental condition monitoring results. The paper ends with a brief discussion and some conclusions.

*Corresponding author: g.irwin@qub.ac.uk, tel:+44(0)2890975439, fax:+44(0)2890664265.

2. AUTOMOTIVE ENGINE TESTS

This section briefly describes the experimental engine test-bed, followed by detailed explanations of the data collection regime under both normal and faulty operating conditions.

2.1 Engine test cell

The application was a four-cylinder 1.8 litre spark ignition engine, manufactured by Nissan. This engine represents current technology with devices such as variable valve timing, inlet swirl plates, exhaust gas recirculation and a close-coupled catalyst. The engine installation can be seen in Fig. 1.

The engine was installed in a state-of-the-art test bed at the Queen's University Belfast. An AC dynamometer, equipped with a Ricardo S3000 controller, was used to control the engine throughout the simulated transient drive cycles. Sensor signals were recorded using the testcell data acquisition hardware – a Ricardo TaskMaster 500/2000 system, capable of recording up to 32 analogue input channels simultaneously.

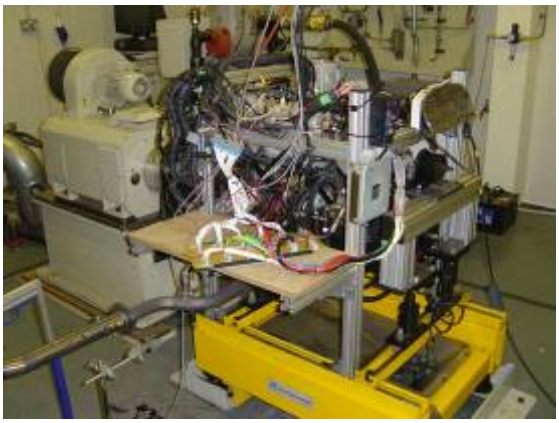

Fig. 1. Engine installed in test laboratory.

The intake subsystem of this engine was investigated. In order to simplify the air intake model, the exhaust gas recirculation (EGR) function was disabled. The following four variables were used to analyse this subsystem: crankshaft rotational speed (rev/min), pedal position (%), mass air flow (kg/h), inlet manifold pressure (bar). Rotational speed and pedal position formed the engine inputs, while the other two represent the dynamic behaviour of the intake system. Significantly, these four variables are all available from sensors fitted to the standard vehicle and so the modelling and fault detection process outlined in this paper requires no additional hardware.

2.2 Drive Cycles

The New European Drive Cycle (NEDC), for which the two engine inputs appear in the top two plots of Fig. 2, was used in this work. All the data sets were collected at a sampling rate of 10Hz. This drive cycle is used for emission certification of light duty vehicles in Europe. It is composed of various sections which simulate both city and highway driving conditions. The measurement of exhaust emissions forms one component of the Type Approval test, which is compulsory for any new vehicle model entering the European market. Other similar drive cycles include the FTP 75 used in the USA and the 10-15 Mode Cycle found in Japan.

It is clear from Fig. 2 that this is a highly transient cycle that includes phases of rapidly varying inputs of engine speed and pedal position as gear changes are simulated. During vehicle deceleration, when the pedal position is zero, the engine speed is frequently higher than idle speed as the engine is motored through the transmission and contributes to the vehicle's braking requirement. These phases of the NEDC are simulated during testing by supplying a torque input from the dynamometer to maintain the commanded engine speed.

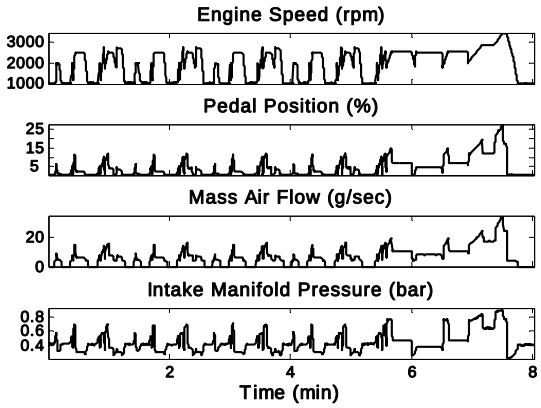

Fig. 2. Engine signals recorded during the NEDC.

While the NEDC is indeed a highly dynamic cycle, and is representative of typical vehicle use, Fig. 4 shows that the engine control inputs do not in fact cover the whole operational map. For example, engine speed does not exceed 3500rpm while the throttle pedal position is less than 15% for the majority of the cycle, briefly reaching a maximum of around 28% at 3500rpm during the highway driving phase. This represents only 55% of the peak torque available at that engine speed. It is clear from this that data from the NEDC would not be suitable for training an AANN as the resulting model would produce inaccurate predictions, and hence generate false alarms, when the engine is used outside the areas of the map accessed during training.

An alternative drive cycle, the Kimmich identification cycle (KI cycle), was considered for the engine inputs to cover a much wider range of the operating region (Kimmich *et al.*, 2005). However, although their cycle indeed covers a broader range of engine speeds and throttle positions, the *rate* at which these variables change is significantly lower than that experienced in real driving. The NEDC, on the other hand, is more dynamic in nature and captures these realistic transients. A modified identification (MI) cycle was therefore produced for our work in order to derive a suitable model. This was developed by examining the sections of the NEDC where the engine speed was undergoing the greatest transients, which occurs during accelerations in 1st gear. The timescale of the KI cycle was then reduced by a factor of 5.7 such that the rate of change of engine speed matched that of the most transient section of the NEDC. The resulting drive

cycle therefore combines the benefits of both the KI cycle and the NEDC as it covers a wide range of the engine's operating map, while also simulating realistic dynamics. The two inputs of the MI cycle are shown in Figure 3, together with the two engine response variables.

Comparing the coverage of the operating regions of the MI and NEDC cycles in Fig. 4, it is clear that any model built on engine data from the former cycle will better represent a much broader range of operation, as required.

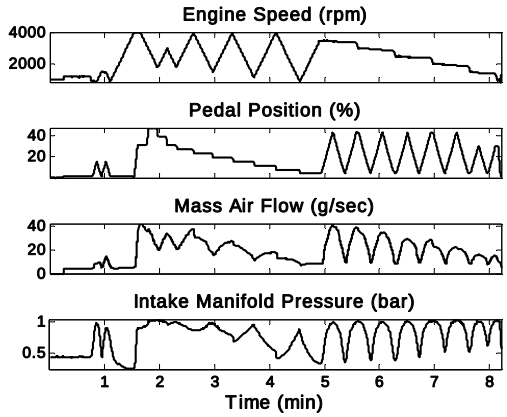

Fig. 3. Engine signals from the modified identification (MI) cycle.

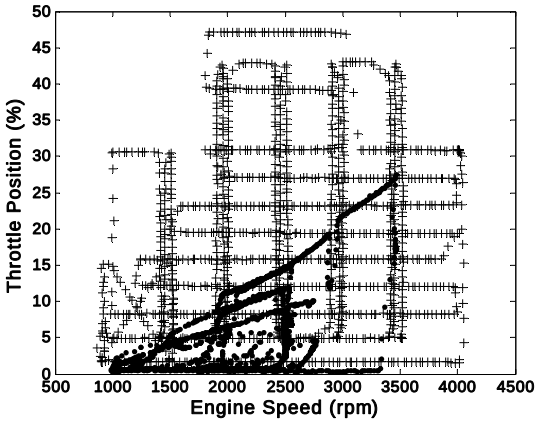

Fig. 4. Comparison of operating regions accessed by the two drive cycles (MI cycle in crosses, NEDC in circles).

2.3 Data collection

The MI cycle last about 8 minutes, providing 4785 data points. This cycle was repeated three times under fault free conditions, with two sets of data used for model building and the third reserved to validate the model. A further single set of NEDC data was recorded during normal 'fault-free' operation in order to assess the model's generalisation capability on unseen data.

2.4 Air leak fault

In this investigation, the faulty condition took the form of an air leak in the intake manifold. This is indicative of a process, rather than sensor fault, and is representative of a leakage past a gasket or fitting between the throttle plate and the intake valve. A minor air leak potentially may not be noticeable to the driver. Nevertheless, when a fault of this type occurs, the driver would adjust the throttle pedal until the desired torque is achieved. Thus, in this fault scenario it is imperative to preserve the values of the engine inputs, i.e. the engine speed and pedal position between the fault-free and faulty conditions. The fault was introduced by drilling a hole into a bolt which was subsequently screwed into the inlet manifold of the engine downstream of the throttle plate. There were in total four such bolts: a solid one without a hole to produce the fault-free condition, and three others with 2mm, 4mm, and 6mm diameter holes to provide faults of differing magnitude. Data representing all three faulty conditions were collected for both the MI cycle and the NEDC.

3. NON-LINEAR PCA

The AANN is a special neural network architecture with 3 hidden layers, called the mapping, bottle-neck, and demapping layers respectively as shown in Fig. 5.

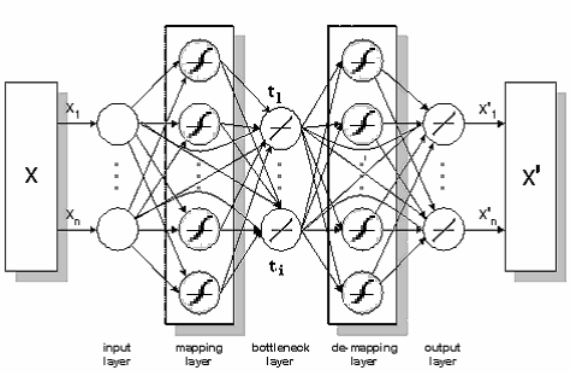

Fig. 5. NLPCA architecture.

The AANN represents an identity mapping for a given set of n variables, such that the network inputs and outputs are identical. The activation function used in the mapping and demapping layers was a hyperbolic tangent function, while the other two layers contained linear activation functions. Note also the presence of direct links from the inputs to the bottleneck layer and from the bottleneck layer to the outputs.

This network is regarded as a nonlinear version of PCA because of its similarity in producing 'scores', normally fewer in number than the original variables from a given process. The scores generated by linear PCA are based on the linear relationships among the original variables, and can effectively represent the variance in these variables. The scores can then be used for reconstructing the original variables or for monitoring any unknown data from the same process. In the case of NLPCA, the $i < n$ nodes in the bottleneck layer represent the nonlinear scores, $\mathbf{t}_1, \cdots, \mathbf{t}_i$. They are obtained by capturing the nonlinear relationship between the original variables in the input layer. The scores are then able to reconstruct the original variables by passing them through the demapping layer to the output layer. If there are only linear relationships between the original variables,

the scores from such a NLPCA model should in theory be the same as those from a PCA one.

The mathematical description of the identity mapping can be described in two parts. The scores are produced in Figure 5 by the mapping layer that constitutes a nonlinear transformation on the inputs. Thus, for the k^{th} score:

$$\mathbf{t}_k = G_k(\mathbf{x}) \tag{1}$$

The variables at the output layer $\mathbf{x}'$ can then be obtained using:

$$x'_j = H_j\begin{pmatrix} t_1 & t_2 & \cdots & t_i \end{pmatrix}^T \tag{2}$$

The AANN network parameters are trained by minimising the cost function:

$$J = \sum_{j=1}^{n} (x_j - x'_j)^2 \tag{3}$$

When the NLPCA has been trained from fault-free data, a Q statistic can be calculated based on the model prediction error. Thus,

$$Q = \mathbf{e}^T\mathbf{e} \tag{4}$$

Here, $\mathbf{e}$ refers to the difference between the model prediction and the measured value for one sample. Q follows a central χ^2 distribution and the confidence limits can be estimated as discussed in (Jackson, 1991). The values of the Q statistic from the training data were used to calculate the 95% and 99% confidence limits, which are used as benchmarks for monitoring unknown data from the process.

4. RESULTS

The NLPCA model was developed from the two normal sets of data resulting from the MI cycle. Subsection 4.1 provides details of how the model was trained, along with its validation on the MI cycle. The generalisation capability of the model is assessed in section 4.2, while its ability to detect air leak faults with the engine operating under the MI and NEDC cycle is presented in Subsection 4.3 and 4.4 respectively.

4.1 Training and validation

The NLPCA built on the training data had a 4-9-3-9-4 structure with 9 nodes in the mapping and demapping layers and 3 in the bottle-neck layer. Having trained the model it was subsequently validated using a new set of data recorded during a fault-free MI cycle. The performance of the model during this training and validation process is summarised by the variation in the Q statistic which is plotted in Figure 6.

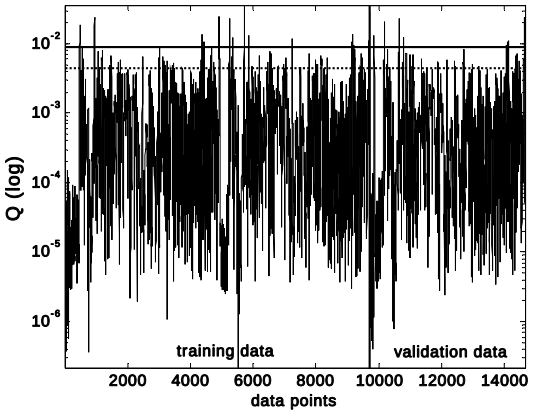

Fig. 6. Variation of the Q Statistic during model validation using normal the MI cycle.

The upper limit represents the 99% confidence, whereas the lower is for the 95% threshold, both of which were obtained from the Q statistic of the training data. The resulting levels of violations indicated are acceptable.

4.2 Generalisation

When a model is used in practice, the new operational inputs often take a form previously unseen by the model during its training. This is particularly the case with automotive engines, which impose a stringent requirement for generalisation if the results are to be of practical significance. This aspect of the NLPCA model was therefore challenged by comparing measured and predicted values of the two output variables, mass air flow and manifold air pressure, during the NEDC with the engine operating under normal, *fault-free*, conditions. The results are compared in Fig. 7 which reveals little difference between the measured data and the corresponding predictions. This is an important finding as the NEDC engine inputs of speed and pedal position vary at rates ranging from steady state through to the highly transient conditions found during 1st gear accelerations and motored deceleration phases. The model's excellent generalisation capability is further illustrated in Fig. 8, which shows the low values of the Q statistic produced during this analysis.

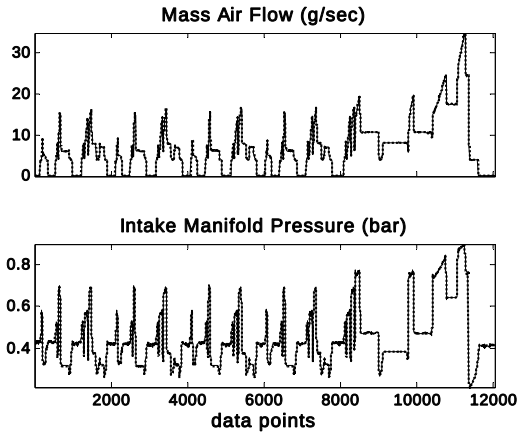

Fig. 7. Model predictions (dashed) and measured values (solid) on during fault-free NEDC operation.

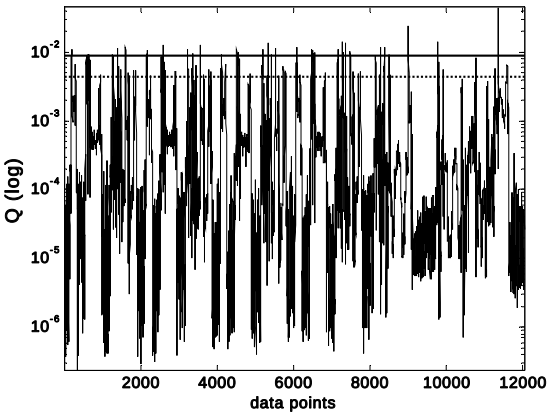

Fig. 8. Variation of the Q Statistic throughout the fault-free NEDC.

4.3 Fault detection on MI data

This section assesses the model's ability to detect a fault produced by running the MI cycle in the presence of a 2mm, 4mm, and 6mm air leak on the inlet manifold. It can be seen in Fig. 9 that the number of violations grows as the magnitude of the fault increased. Since each set of data contains 4780 data points, the model would indicate the fault if approximately 48 points exceed the 99% limit or 239 points fall outside the 95% limit. The 4mm and 6mm faults are clearly detected by the significant number of violations to both limits. In the case of the 2mm air leak, there are 297 violations for the 99% limit and 1330 for the 95% limit. This clearly indicates the successful detection of what is in fact a very small fault.

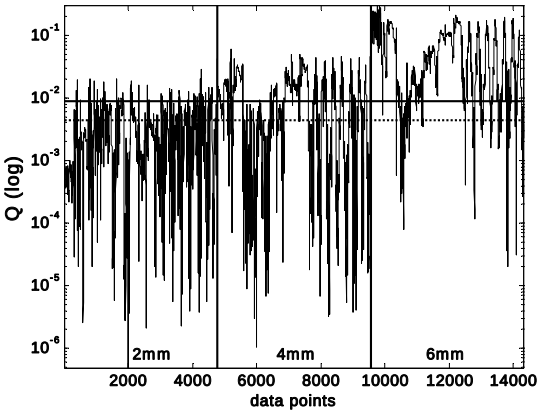

Fig. 9. Monitoring 2mm/4mm/6mm air leak faults with MI cycles.

It should be noted that there are certain regions in the faulty data set where the impact of the air leak fault may not be apparent in the monitoring statistic. This is expected, as the fault would not affect the engine when it is operating at high throttle openings. Under such circumstances, the manifold pressure is close to, or equal to, the atmospheric pressure. Consequently, the pressure difference across the leakage orifice is small and the flow rate of air through it is therefore negligible. This phenomenon is demonstrated by plotting the intake manifold pressure and the Q statistic during the MI cycle with a 4mm air leak in Fig. 10.

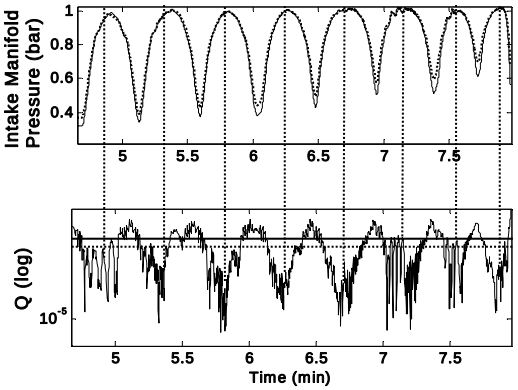

Fig. 10. Influence of 4mm air leak fault on engine variable with MI cycle and corresponding Q statistic.

This section of data corresponds to that collected from 4.5 minutes onwards in Fig. 3, where the pedal position presents large values periodically. The upper plot in Fig. 10 compares the engine signal recorded with a 4mm air leak fault (dotted trace) and the normal, fault-free, condition (solid trace). It can be seen that the phases at which the fault is not detected by the Q statistic correspond to points at which the manifold pressure is close to atmospheric.

4.4 Fault detection on NEDC

The variation in Q statistic for the NLPCA model shown in Fig. 11 includes the three faulty conditions with the engine operating under the NEDC. These results confirm successful detection of all faulty conditions, despite the model being trained using a substantially different identification cycle.

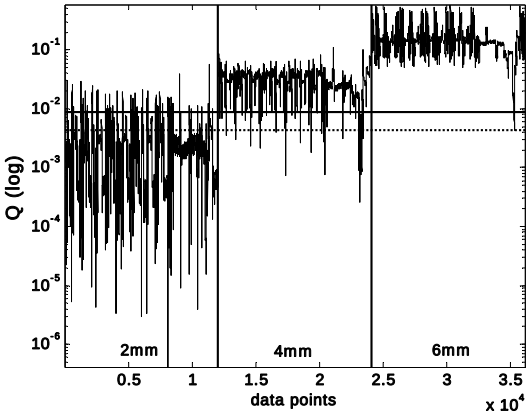

Fig. 11. Fault detection with NEDC data.

5. DISCUSSION AND CONCLUSIONS

This paper demonstrated that the NLPCA model gave good results when used for modelling and air-leak detection on noisy validation data produced by the MI cycle. Crucially for automotive applications it also provided excellent generalisation when used for modelling the unseen engine measurements from the NEDC. There was an absence of

unwanted false alarms under fault-free conditions, and successful detection air leaks of varying magnitude in the inlet manifold.

Although NLPCA has proved to be successful in this application, it has unavoidable drawbacks due to its structural complexity. If the number of engine variables increases, the training process can become difficult. In addition, if fault detection is required under faster transient conditions than those produced by the NEDC, then including time lagged variables in the AANN or using a recurrent network structure, may not be feasible due to the very significant increase in complexity involved. A new method is currently under investigation in which PCA is used on a set of lagged engine variables for dynamical modelling, from which the resulting smaller set of principal components then form the inputs to a PCA. Additionally, semi-physical models may be applied to the measured signals to produce a reduced set of input variables (Wang *et al*., 2008b)

The next stage of the research programme will also consider a broader range of both process and sensor faults, with an emphasis on fault diagnostics. The gradual degradation of the catalytic converter, which has a direct impact on tailpipe emissions, will be of special interest.

ACKNOWLEDGEMENT

The authors gratefully acknowledge support from the U.K. Engineering and Physical Sciences Research Council (Grant No. EP/C005457).

REFERENCES

Antory, D. U. Kruger, G.W. Irwin and G. McCullough (2005). Fault diagnosis in internal combustion engines using nonlinear multivariate statistics. *Proceedings of the Institution of Mechanical Engineers, Part I Journal of Systems and Control Engineering*, 219(4), 243-258.

Grimaldi, C.N., Mariani, F. (2001) OBD Engine Fault Detection Using a Neural Approach. SAE Paper No. 2001-01-0559.

Jackson, J.E. (1991). *A Users Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics. John Wiley, New York.

Kimmich, F., A. Schwarte and R. Isermann (2005). Fault detection for modern diesel engines using signal- and process model-based methods. *Control Engineering Practice* 13, 189-203.

Kramer, M.A. (1992). Autoassociative neural networks. *Computers and Chemical Engineering* 16(4), 313-328.

Nyberg, M., (1999). Model Based Diagnosis of Both Sensor Faults and Leakage in the Air-Intake System of an SI Engine. SAE Paper No. 1999-01-0860.

Stobart, R. (2003). Control Oriented Models for Exhaust Gas Aftertreatment; A Review and Prospects, SAE Paper No. 2003-01-1004.

Wang, X., U. Kruger, G.W. Irwin, G. Mc Cullough and N. Mc Dowell (2008a). Nonlinear PCA with the local approach for diesel engine fault detection and diagnosis. *IEEE Transactions on Control Systems Technology*, 16(1), 122-129.

Wang, X., N. Mc Dowell , U. Kruger, G. McCullough and G.W. Irwin (2008b). Semi-physical neural network model in detecting transient faults using the local approach. *Proceedings of the 17th IFAC World Congress, Seoul, Korea*, paper TuC28.6, 7086-7090