2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

Identification and Monitoring of Automotive Engines[1]

Wallace E. Larimore* and Hossein Javaherian†

*Adaptics, Inc, 1717 Briar Ridge Road, McLean, VA 22101 USA
Phone: 703 532-0062, Fax: 703 536-3319, Email: larimore@adaptics.com
†General Motors R&D, Warren, MI
Email: hossein.javaherian@gm.com

Abstract— The objective of this paper is to extend and refine the nonlinear canonical variate analysis (NLCVA) methods developed in the previous work for system identification and monitoring of automotive engines. The use of additional refinements in the nonlinear modeling are developed including the use of more general bases of nonlinear functions. One such refinement in the NLCVA system identification is the selection of basis functions using the method of Leaps and Bounds with the Akaike information criterion AIC. Delay estimation procedures are used to considerably reduce the state order of the identified engine models. This also considerably reduces the number of estimated parameters that directly affects the identified model accuracy. This increased accuracy also affects the ability to monitor changes or faults in dynamic engine characteristics. A further objective of this paper is the development and use of nonlinear monitoring methods as extensions of several previously used linear CVA monitoring procedures. For the case of linear Gaussian systems, these monitoring methods have optimal properties in detecting faults or system changes in terms of the general maximum likelihood method. In the nonlinear case, departures from optimality are investigated, but the procedure is shown to still work quite effectively for detecting and identifying system faults and changes.

Index Terms— Nonlinear subspace system identification, automotive engine fault detection, feedback.

I. Approach to Engine Modeling and Monitoring

The identification and monitoring of automotive engines has been a difficult problem. The engine dynamics are nonlinear and depend on a host of variables that may change considerably at different engine operating conditions. While the identification and monitoring of linear time-invariant systems has become routine using advanced subspace methods that are automated, the problem for nonlinear systems has been more difficult.

In this paper, a number of extensions (Larimore, 1999, 2003, 2005, 2006) of subspace modeling (Van Overschee and DeMoor, 1994; Verhaegen, 1994) and monitoring (Larimore, 1997a; Wang et al, 1997; Juricek et al, 2004) are discussed and applied to automotive engine data. Those methods are implemented in Matlab ©based on the ADAPTx software (Larimore, 1992a) to demonstrate the substantial improvements that can be obtained:

- Efficient Computation. The CVA system identification offers a computationally stable and efficient way to identify a high-order multivariable system.
- Optimal Open-loop Baseline Model. CVA provides optimal maximum likelihood (ML) parameter estimation (Bauer, 2005; Deistler et al, 1994) of the open-loop dynamics in the presence of feedback for a broad class of systems (Larimore, 1996, 1997b, 2004, 2006) to provide an optimal baseline model computed for no-fault conditions in large samples.
- Near Optimal Test for a Fault. The method of scores computes an approximately optimal likelihood ratio (LR) test of hypothesis for any specified fault relative to the ML baseline model determined by CVA. This approximation is exact in linear gaussian processes or in large samples.
- No Optimization Required. Only the computation of first and second derivatives of the log likelihood function (LLF) relative to the fault parameter(s) is required.
- Linear Growth following a Fault. Characteristic linear growth in the score test statistic following the occurrence of the fault.
- Handles Outliers and Mismodeling. Outlier detection methods may be useful even in the case of intermittent mismodeling in isolated portions of the state space.
- Near Optimal Fault Isolation for Simultaneous Faults. Provides approximately optimal resolution of which fault(s) have occurred and which have not occurred for simultaneous faults. The approximation is exact for linear gaussian processes for large samples
- Minimum Time/Samples for Detection. A fault is detected in the minimum possible number of samples.

Various aspects of the monitoring and fault detection are discussed below using the data from a 5.3L V8 engine. The case of the one output, air-fuel ratio afr, was considered. The output afr is known to involve a lot more delay than the output, *torque*, often used in engine modeling. Primarily the case of 9-inputs are considered with particular variables shown in Table I. The inputs T_{cool} and T_{exh} were found early in the study

[1]Financial support for this research was provided by General Motors Corporation which is gratefully acknowledged.

Variable	Name	Number
	Outputs	
afr	air-fuel ratio	1
$torque$	output torque	
	Inputs	
TPS	throttle position sensor	1
$VIgnition$	ignition voltage	2
fpw	fuel pulse width	3
T_{cool}	coolant temperature	
T_{exh}	exhaust temperature	
maf	manifold air flow	4
map	manifold absolute pressure	5
rpm	engine speed	6
SA	spark angle	7

TABLE I
Variables Used in Engine Models

to have little effect on afr and were removed from the analysis.

II. State Space Model Structure

The ultimate model structure desired is a state space form because it can have a more parsimonious structure. This is achieved in a number of steps:

- A NARX (Nonlinear AutoRegressive with eXogenous inputs) model is fitted to the data involving nonlinear functions of the inputs (Larimore, 1990b, 2002; Rao, 1966).
- Only linear functions of the outputs are included in the autoregressive (AR) terms of the NARX model to insure stability of the model.
- Various nonlinear functions involving moving averages in the inputs (MX terms) are added and deleted from the NARX model using subset selection (Furnival and Wilson, 1974) and the AIC (Akaike, 1973) to find a good fit to the data.
- The delays between the various inputs and outputs of the NARX model are determined by hypothesis testing (Larimore, 2003).
- The inputs are advanced to removed these delays to potentially reduce the state order of the state space model.
- The final model includes the state space model with delay blocks at the inputs.

The NARX model is of the form (Tong, 1990)

$$y_t = g(y_{t-1}, \ldots, y_{t-j}, u_{t-1}, \ldots, u_{t-k}) + e_t = g(p_t) + e_t \tag{1}$$

where y_t is a vector of outputs, u_t are a vector of inputs, and e_t is a white noise vector with some specified probability density function. The past p_t is the vector of past lagged outputs and inputs of finite dimension. In much of this paper, the maximum likelihood (ML) problem is considered where e_t is a zero mean gaussian random variable with covariance matrix Σ_{ee}.

The nonlinear function $g(p_t)$ of the past can be approximated as a linear combination of basis functions such as monomial functions

$$p_t^{(\iota)} = y_{t-1}^{(\iota_1)} y_{t-2}^{(\iota_2)} \cdots y_{t-j}^{(\iota_j)} u_{t-1}^{(\iota_1)} u_{t-2}^{(\iota_2)} \cdots u_{t-k}^{(\iota_k)} \tag{2}$$

where $\iota = (\iota_1, \iota_2, \ldots, \iota_{j+k})$ is a vector of nonnegative integers composed of the indices ι_ℓ, for $\ell = 1, \ldots, j+k$, specifying the powers of the respective outputs and inputs of the past vector p_t. Here $\mathcal{B}_p = \{p_t^\iota \text{ for } \iota \in I\}$ for some set I of indices. Then the approximating linear combination is

$$g(p_t) \cong \sum_{\iota \in I} a_\iota p_t^{(\iota)} \tag{3}$$

where I is an index set specifying the power product terms in the sum and a_ι for $\iota \in I$ are the unknown coefficients.

As discussed above, the selected model structure retains linear variables in the autoregressive terms and contains nonlinear variables only in the moving average terms. This is due to nonlinear AR terms causing unstable dynamics in nonlinear propagation of the dynamics. On the other hand, for linear AR terms, if the AR polynomial has stable roots, then the dynamics are stable no matter what the MX terms are. This does impose some limitations on the potential dynamics of the NARX model, but it does insure a stable model if the AR terms are stable.

III. CVA and Selection of State Order

For a given NARX model structure developed below, the states of the system are first determined by the Canonical Variate Analysis (CVA) (Hotelling, 1936) method of system identification between the past and corrected future (Larimore, 1983, 1999, 2004, 2006). This is done to determine and order linear combinations of nonlinear functions of the past in terms of their predictability for the future. A plot of the Akaike information criteria (AIC) (Akaike, 1973, 1976; Larimore, 1983b, Larimore and Mehra, 1985; Hurvich et al, 1991) versus the model state order gives a concise description of how the choice of the state order affects prediction of the future. Figure 1 shows a comparison of using the various NARX models for developing a state space model. The NARX models compared have the following form:

- the NARX model with quadratic terms and no delays removed (minimum at 16 states) – the top curve,
- quadratic terms with delays removed (minimum at 7 states) – the lowest curve at 6 states,
- quadratic and cubic terms with delays removed (minimum at 10 states) – labeled "cubic, delay modeled", and
- cross product, quadratic and cubic terms with delays removed (minimum at 15 states but 10 states very close second) – the lowest curve at 3 states.

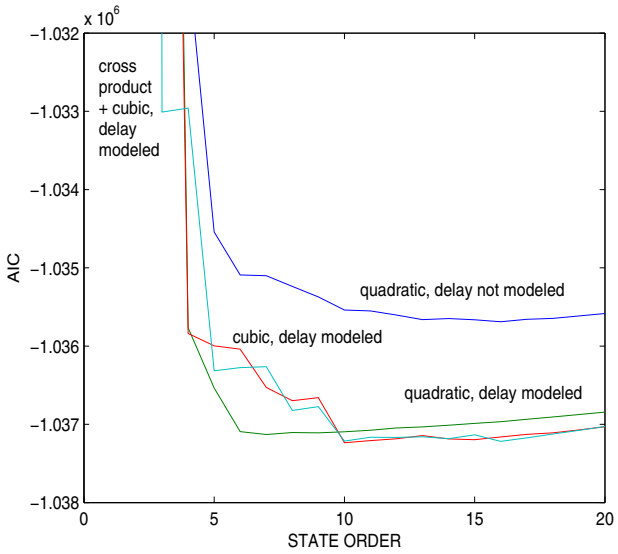

Fig. 1. Comparison of AIC versus State Order

The quadratic model with no delays removed is decidedly inferior to the other models. In the other models, a 4-state model captures most of the prediction for the future. The quadratic model with 6 states does almost as well as any of the other models, and it is only slightly surpassed by the cubic or the cubic plus cross product models with 10 or more states. The baseline model selected was the 7-state quadratic model with delays removed with 14 inputs consisting of the 7 inputs in Table I plus their squares.

IV. Score Statistic Approach to CVA

The approach to monitoring changes and faults in the engine is based on the score statistic (Cox and Hinkley, 1974; Cox, 2006), also called the local statistics approach (Basseville and Nikiforov, 1993), involving a particular form of the likelihood ratio test. For the nonlinear state space structure above, the nonlinear CVA procedure provides approximate ML estimates for large sample. The one-step prediction errors ν_t are used to evaluate the log likelihood function (LLF) as

$$logp(Y_1^N;\theta) = -\frac{1}{2}\sum_{t=1}^{N}[ln(2\pi|R| + \nu_t^T R^{-1}\nu_t \qquad (4)$$

where R is the covariance of ν_t, and Y_1^N usually abbreviated as Y represents generically the input and output data (see Larimore (2004)).

In the case where a parameter θ describing a fault is 1-dimensional, the likelihood ratio statistic for comparison of the null hypothesis θ_0 versus a particular alternative $\theta_0 + \delta$ is

$$log\frac{p(Y;\theta_0+\delta)}{p(Y;\theta_0)} = \delta\frac{\partial logp(Y;\theta_0)}{\partial\theta_0} + o(\delta) \qquad (5)$$

where $o(\delta)$ are higher order terms that go to zero in large samples. The partial derivative $\partial logp(Y;\theta_0)/\partial\theta_0$ of the log likelihood function is called the efficient score for the observations Y and denoted as $U(\theta_0)$, and is fundamental in the asymptotic theory of maximum likelihood estimates.

In general situations, a baseline model of the process may involve a large number of parameters θ. However, for the purpose of monitoring, a much smaller set of parameters γ may be of interest for addressing issues of fault detection and isolation. In this case, γ may specify a reparameterization $\theta(\gamma)$ of the original parameters θ. The one requirement is that the null hypothesis parameter vector γ_0 be associated with the baseline model $\theta_0 = \theta(\gamma_0)$. This is necessary since all computations of the score statistic in the method of scores must be evaluated at the baseline model with parameters θ_0. In most discussions below the fault detection and isolation will be expressed directly in terms of the original parameters θ, but for additional clarity the more explicit and general parameterization $\theta(\gamma)$ will be used.

Consider the general case where the fault parameters are d-dimensional with the score given by

$$U(\theta) = \nabla logp(Y;\theta) \qquad (6)$$

where $\nabla = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_d)$ denotes the gradient of the LLF with respect to the parameters θ.

The following properties of the efficient score $U(\theta_0)$ are easily shown (Cox and Hinkley, 1974, pp. 107-109 and 311-330; Cox, 2006, pp. 96-105):

- the expected value is zero so $E\{U(\theta_0);\theta_0\} = 0$.
- The covariance matrix is

$$Cov\{U(\theta_0);\theta_0\} = E\{U(\theta_0)U^T(\theta_0);\theta_0\} = I(\theta_0) \qquad (7)$$

  where $I(\theta_0)$ is the Fisher information matrix with the parameter estimation error covariance matrix $E(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T = I(\theta_0)^{-1}$.
- The Fisher information matrix is also expressible as the expected value of the second partial derivative matrix of the log likelihood function (LLF)

$$Cov\{U(\theta_0);\theta_0\} = I(\theta_0) = E\{-\nabla\nabla^T logp(Y;\theta_0);\theta_0\} \qquad (8)$$

Now consider the likelihood ratio test, of the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_A : \theta \in \Theta_A$ where Θ_A contains θ_0, using the test statistic

$$W_L = 2log\frac{sup_{\theta\in\Theta_A}p(Y;\theta)}{p(Y;\theta_0)} \qquad (9)$$

An asymptotically equivalent test is the ML test statistic

$$W_E = (\hat{\theta}-\theta_0)^T I(\theta_0)(\hat{\theta}-\theta_0) \qquad (10)$$

using the maximum likelihood estimate $\hat{\theta}$ under the alternative hypothesis H_A. Using the efficient score

$U(\theta_0)$, the score statistic

$$W_U = U^T(\theta_0)I^{-1}(\theta_0)U^T(\theta_0) \tag{11}$$

can be shown to be asymptotically equivalent to both W_L and W_E. The advantage of the score statistic W_U is that the maximum likelihood statistic $\hat{\theta}$ does not need to be computed, but only partial derivatives of the log likelihood function at θ_0. This is particularly useful in a monitoring problem where little data may be available following a change.

A recent paper (Juricek, Seborg, and Larimore, 2004) has used the score statistic for the monitoring of processes identified using CVA that was particularly sensitive in detecting a number of departures from the null hypothesis.

V. Distribution of the Likelihood Ratio Test

Under suitable regularity conditions, the asymptotic distribution of the LR statistic in the nested case is Chi-square with the degrees of freedom equal to the number of additional parameters in the alternative hypothesis that are not contained in the null hypotheses. Under the null hypothesis H_0, the distribution of the test statistic (11) is central Chi-squared. Under the alternative hypothesis H_A, the distribution of the test statistics is noncentral with the noncentrality parameter given by

$$(\theta^* - \theta_0)^T I(\theta_0)(\theta^* - \theta_0) = N(\theta^* - \theta_0)^T I_s(\theta_0)(\theta^* - \theta_0) \tag{12}$$

where $I_s(\theta_0)$ is the per sample Fisher information and N is the sample size in computing the test statistic.

In the tests for faults discussed below, the fault test statistic is the cumulative sum of the likelihood ratio computed from the score test statistic (11), so the distribution of the fault test statistic is that of a cumulative noncentral chi-square variable. Thus, the expected value of the fault test statistic grows linearly with the sample size. The slope of the fault test statistic when a fault is present is equal to the square of the change $\theta^* - \theta_0$ in the parameters under the alternative hypothesis H_A normalized by the per sample covariance matrix of the parameter estimation error (the inverse Fisher information matrix). This completely characterizes the asymptotic behavior of the fault test statistic based on the maximum LR test and equivalent test statistics.

VI. Air-Fuel Ratio Bias Fault

From the state space model form, an output bias b_o in the system is expressed as

$$y_t = \hat{y}_t + b_o + \nu_t \tag{13}$$

where $\hat{y}_t$ is the one-step prediction of the Kalman filter and ν_t is the innovation. To simulate an output bias fault in the afr variable of magnitude 0.2, the value 0.2 was added to the afr output. The resulting data set will be called the 'Fault' data whereas the original will be called the 'NoFault' data. As discussed above, because of the maximum likelihood nature of the identification method, the presence of feedback during the closed-loop operation of the engine will not impact the engine identification results.

Function	Outputs and Inputs		
	afr	fpw	fpw^2
bias function	$afr + b$		
$\partial/\partial b$	1		
$\partial^2/\partial b^2$	0		
gain function		$\gamma * fpw$	$(\gamma * fpw)^2$
$\partial/\partial\gamma$		fpw	$2\gamma * fpw^2$
$\partial^2/\partial\gamma^2$		0	$2 * fpw^2$
cross partial			
$\partial^2/\partial b\partial\gamma$	0	0	0

TABLE II
Derivatives of Inputs and Outputs w.r.t. Fault Parameters

The software for monitoring using the score statistic requires the first and second partial derivatives of the output and input data parameterized by the fault parameter b of the afr bias and the fault parameter γ of the fpw gain developed in Section VII. The parameterized functions and their derivatives of the outputs and inputs are given in Table II.

A plot of the score test statistic for detecting the afr bias fault using the NoFault data is shown in Figure 2 to have a maximum value or 23. Under the null hypothesis of no fault, the test statistic is distributed as a Chi-squared statistic on 1 degree of freedom. For a linear gaussian process with efficient estimates of the parameters, a Chi-squared statistic on 1 d.f. has mean 1 and variance 2. Clearly, there is significant mismodeling exhibited by the test statistic.

Figure 3 shows the first 100 points. The first 29 points are used for initialization of the state space model. As derived in (12), starting with point 30 the test statistic grows at a linear rate per sample as the ratio of the square of the bias change of $\Delta b_o = 0.2$ times the per sample information $I_s(b_0)$, so

$$(\Delta b_o)^2 I_s(b_o) = (0.2)^2 * 100 = 4 \tag{14}$$

The score test statistic at 10 samples after the fault exceeds twice the maximum value of 23 of the test statistic due to mismodeling under the null hypothesis in Figure 2. Inspection of the test statistic over the entire Fail data showed that it grows persistently and linearly as long as the fault is present. If the mismodeling can be bounded, then a reasonable threshold can be determined for reliable detection of the fault. In the present case, a detection at times greater than 10 sample times after the fault is plausible, although if there is no rush, a larger detection time such as 50 sample times could be used to prevent the possibility of the mismodeling causing false alarms.

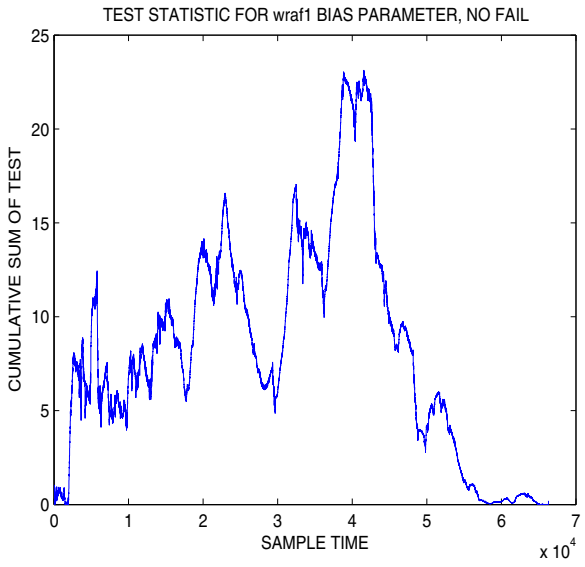

Fig. 2. Score Test Statistic for afr ($wraf1$) Bias for the NoFail data.

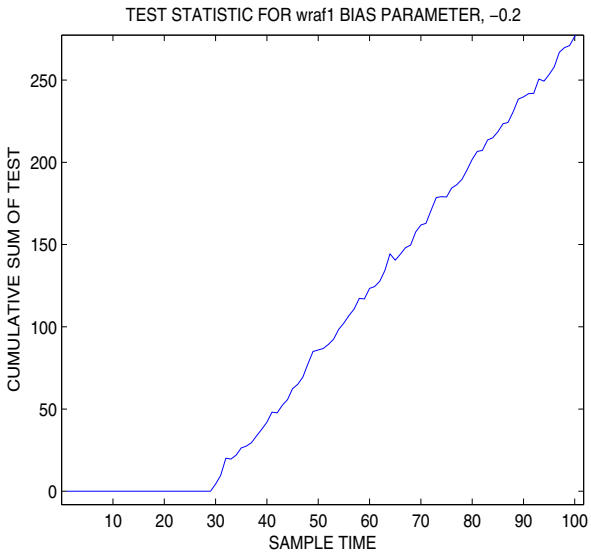

Fig. 3. Magnified Score Test Statistic for afr ($wraf1$) Bias for the Fail data with $\Delta b_o = 0.2$.

A key parameter in the test is the growth rate $(\Delta b_o)^2 I(b_o)$. This quantity times the number of observations since the occurrence of the fault gives the essential behavior of the test as derived in (12). If there is no mismodeling, then this growth rate (14) characterizes the test and specifies the detection probability as a function of sample size.

VII. Monitoring Injector Clogging

One fault mode of particular interest in this study is the fuel injector clogging. From the state space model form, the injector clog parameter γ of the system produces a change in the term $G * u_t$ to the form $G * g(\gamma) * u_t$ in the state equation

$$x_{t+1} = \phi x_t + G g(\gamma) u_t + K \nu_t \tag{15}$$

The diagonal matrix $g(\gamma)$ of gain factors $diag(g) = (g_{1,t}(\gamma), \ldots, g_{dimu,t}(\gamma))$ has elements equal to 1 for the case of no fault and less than 1 for a fault. Note that the parameters $g_{i,t}(\gamma)$ of the fault can be associated with a gain change in the gain matrix G or with a change in the inputs u_t. It will be convenient below to associate the clogging with a scaling of the inputs u_t associated with fpw, i.e. input variables

$$g_3(\gamma) * u_{3,t} = (\gamma * fpw_t) = \gamma * u_{3,t} \tag{16}$$

$$g_{10}(\gamma) * u_{10,t} = (\gamma * fpw_t)^2 = \gamma^2 * (u_{3,t})^2. \tag{17}$$

where $u_{3,t}$ is the linear term and $u_{10,t}$ is the quadratic term. The derivatives of the parameterized input functions for inputs fpw and $(fpw)^2$ are given in Table II.

To simulate an injector clog by a factor γ using the engine data set, the input variable fpw is multiplied by the factor $1/\gamma$. The resulting data set with the inputs fpw/γ and $(fpw/\gamma)^2$ will be called the 'Fault' data whereas the original will be called the 'NoFault' data. This produces the correct result since the the 'Fault' data input of fpw/γ is reduced by factor γ giving the actual input data in the 'NoFault' case. As discussed above, because of the maximum likelihood nature of the CVA identification method, the feedback or interaction of the engine with other variables does not effect any other input or output variables. Thus in simulating the injector clog fault it is only necessary to proportionately increase the input.

A plot of the score test statistic is shown in Figure 4 for the NoFault data. Under the null hypothesis of no fault, the test statistic is distributed as a Chi-squared statistic on 1 degree of freedom. For a linear gaussian process with exact estimates of the parameters, a Chi-squared statistic has mean 1 and variance 2. Clearly, there is significant mismodeling exhibited by the test statistic.

A plot of the score statistic with the fpw gain fault of $\gamma = 0.7$ is shown in Figure 5 for the data set consisting of 117,230 points. Note the abrupt steps in the test statistic near the sample times of 20,000 and 90,000. These steps are the result of outliers in the gradient of the LLF shown in Figure 6. It is possible to use robust outlier detection methods to remove these outlier effects.

Figure 7 shows the first 400 points. The first 29 points are used for initialization of the state space model. Starting with point 30, the test statistic grows at an approximately linear rate per sample. The theoretical

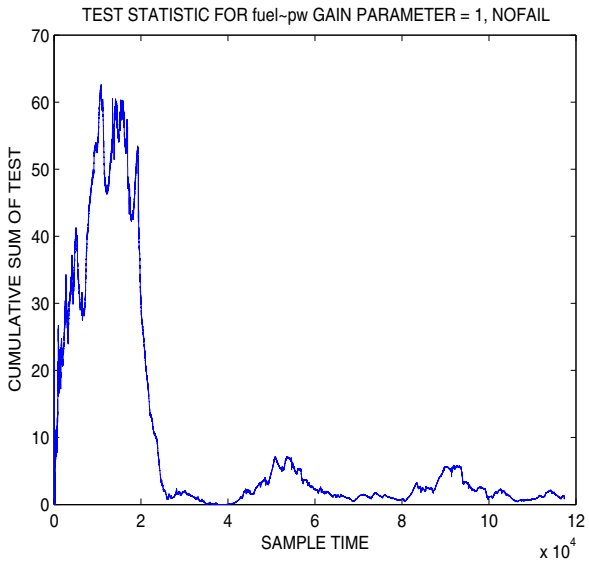

Fig. 4. Score Test Statistic using the NoFail data with fpw ($fuel\_pw$) gain parameter $\gamma = 1$.

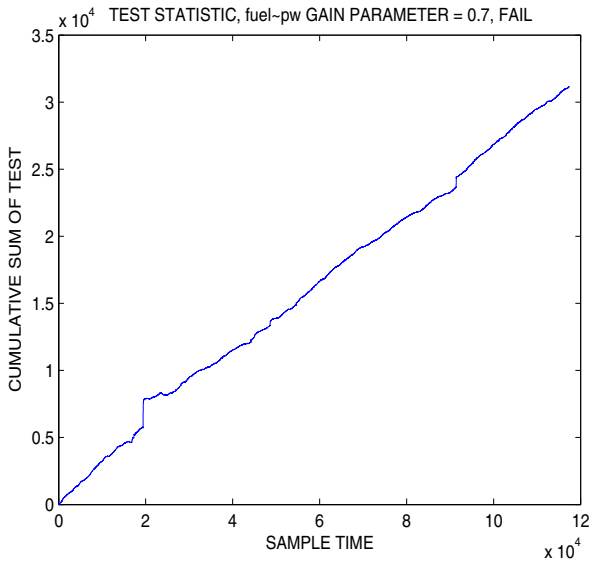

Fig. 5. Score Test Statistic using Fail data with with fpw ($fuel\_pw$) gain parameter $\gamma = 0.7$.

rate is the ratio of the square of the change in the gain parameter of $\Delta\gamma = 0.3$ times the per sample information $I(\gamma_0) = 1.3$, so

$$(\Delta\gamma)^2 I(\gamma_0) = (0.3)^2 * 1.3 = 0.117 \qquad (18)$$

The actual per sample growth rate observed in Figure 5 is 0.25, about double the theoretical. This is due to the nonquadratic behavior of the log likelihood function resulting from the nonlinear function $(\gamma * fpw)^2$. The growth rate (18) is based on the test statistic W_E given in terms of the parameter errors (10) whereas the actual score test statistic W_U is based on the partial derivatives (11) of the LLF. These test statistics are identical if the LLF is a quadratic function, i.e. gaussian, or are approximately identical for large samples. However, the quadratic input function $(\gamma * fpw)^2$ produces a significant departure of the LLF from quadratic.

The score test statistic at 500 samples after the fault exceeds three times the maximum value of 63 of the test statistic due to mismodeling under the null hypothesis in Figure 4. It is clear that the test statistic grows persistently and approximately linearly as long as the fault is present. If the mismodeling can be bounded, then a reasonable threshold can be determined for reliable detection of the fault. In the present case, detection using greater than 500 samples is plausible, although if there is no rush, a larger detection time could be used to prevent the possibility of mismodeling causing false alarms, such as 1000 or 2000 samples.

VIII. Detection and Isolation of Simultaneous Faults

The likelihood ratio tests for simultaneous faults is easily computed. In the case of the simultaneous afr bias and fpw gain faults, since from Table II the second cross partial derivative of the LLF is zero, the two tests are independent. Thus there is no correlation between these two tests for faults and they can be determined as two independent Chi-squared tests from Eq (11).

IX. Time/Samples Required for Detection

The time or equivalently the number N of samples required to detect a given size $\Delta\theta = \theta - \theta_0$ parameter change with at least probability $1-\beta$ can be determined. If the distribution of the test statistic satisfies the theory, then the per sample noncentrality parameter is determined using (12) as

$$\delta_s = (\theta^* - \theta_0)^T I_s(\theta_0)(\theta^* - \theta_0) \qquad (19)$$

The least sample size N is then determined such that

$$\chi^2(\nu, N\delta_s) \geq 1 - \beta \qquad (20)$$

where $\chi^2(\nu, N\delta_s)$ is the χ^2 distribution function on ν degrees of freedom with noncentrality parameter $N\delta_s$.

If the test statistic deviates significantly from the theory, then fault detection is based on the typical linear rate δ_s of growth per sample of the score statistic, such as in Figure 5. A threshold is set for fault detection and the number N of samples needed to exceed the threshold is calculated as the threshold divided by the per sample noncentrality parameter δ_s. For example, in Figure 5 the rate of growth of the test statistic is 0.25 per sample. If the detection threshold is set at 500, then the number of samples required is $500/0.25 = 2000$ samples.

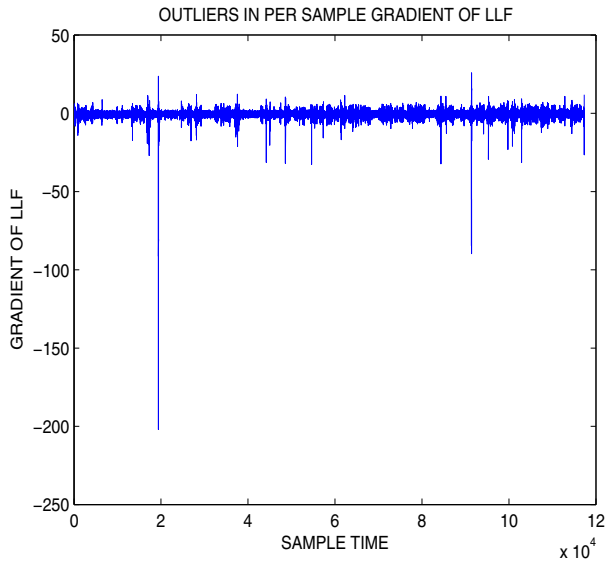

Fig. 6. Outliers in Per Sample Gradient of LLf for Fail data with fpw ($fuel\_pw$) gain parameter $\gamma = 0.7$.

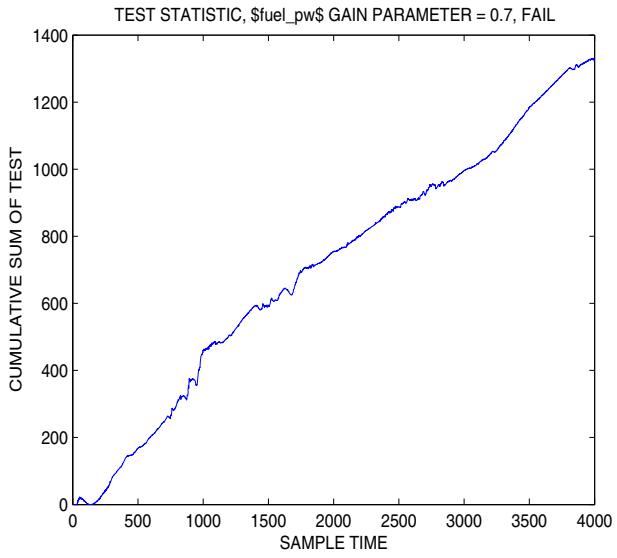

Fig. 7. Magnified Score Test Statistic using Fail data with fpw ($fuel\_pw$) gain parameter $\gamma = 0.7$.

X. Impact of Model Variation on Monitoring and Fault Detection

In this study, the basic concept of monitoring and fault detection has been to obtain a dynamic model of the engine and to use the model to remove all of the predictable variation induced by the dynamics within the error of estimating the model from data. This presumes that the model is time invariant or any time variation is known. It also presumes that an experiment can be done to obtain data resulting in a model that is estimated substantially more accurately than the faults to be determined.

If such a model is available, then the maximum likelihood fault detection methods discussed in this report can be applied to obtain near optimal detection of any faults that may occur. There are, however, several potential problems that can occur that prevent reliable detection or isolation of a fault. If the model has regions of the state space that produce outliers, then this may cause anomalous computation in the monitoring and fault detection algorithms. This was seen to be the case for the fpw clog fault. Unlike the situation for linear time-invariant systems where all regions of the state space are identified with uniform accuracy, in nonlinear systems the model identification may be poor in regions of the state space containing a sparsity of trajectories leading to a poor model in such regions.

The strategy taken in this study is to identify such regions or outliers and to ignore faults in such regions that are highly influenced by such outliers. In particular, the score test statistic in the presence of a fault was seen to have an approximately linear trend. Large outliers in the efficient score computation were seen to be strongly related to significant deviations in the linear growth of the score test statistic and particularly to the gradient of the LLF. Thus the monitoring of the efficient score for outliers may provide a means of avoiding modeling errors characteristic of various regions of the state space. It was also noted in the study that there are much larger errors in the prediction of afr at peaks with rapid rates of increase. This may be a related problem.

Thus with the use of outlier detection methods, it may be possible to do reliable detection even though there are regions of the model or state space that have relatively large errors. In this case, a larger number of samples will be required for reliable detection since it will be necessary at times to discard sections of the data containing outliers or large modeling errors. The extent to which this is a practical issue will depend on the particular details of the modeling and detection characteristics.

XI. Summary and Conclusions

A unique feature of the CVA method implemented in the ADAPTx software is the maximum likelihood estimation of the optimal state space model even in the presence of unknown feedback. The above methods were applied to a 5.3L V8 engine to determine appropriate nonlinear basis functions, engine delays, and reduce the model state order from 16 states to 7 states and more than halve the number of estimated model parameters.

The monitoring involves a baseline nonlinear CVA model of the engine and uses the computation of the score statistic that has a number of very attractive

features. In this approach, no optimization is required, but only the first and second partial derivatives of the log likelihood function (LLF) have to be computed. When a fault occurs, the score statistic grows linearly at a characteristic rate determined by the size of the change in the fault parameter.

The method is near optimal in detecting simultaneous faults, i.e. two or more faults occurring simultaneously. A fault can be detected in minimum time. The score test statistic is reasonably linear when observed over many thousands of samples, but there are some large departures from linearity in regions of the state space where there are modeling errors. To isolate errors due to mismodeling, it was found that the gradient of the LLF, computed as part of the monitoring, could be used to flag outliers. These points could be modified in the computation and result in more reliable detection for such faults as injector clogging.

XII. Acknowledgment

We would like to thank Dr. Man-Feng Chang of General Motors R&D for many useful discussions and his support of the research project.

References

[1] Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," 2nd International Symposium on Information Theory, Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.

[2] Akaike, H. (1976). "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," System Identification: Advances and Case Studies, R.K. Mehra and D.G. Lainiotis, eds., New York: Academic Press, pp. 27-96.

[3] Basseville, M., and I.V. Nikiforov (1993), Detection of Abrupt Changes: Theory and Application, Englewood Cliffs, N.J.: Prentice Hall. Web download at http://www.irisa.fr/sisthem/kniga/

[4] Bauer, D. (2005) "Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs", accepted for publication, J. Time Series Analysis.

[5] Conner, J.S., D.E. Seborg, and W.E. Larimore (2004), "A Theoretical Analysis of the DeltaAIC Statistic for Optimal Detection of Small Changes", Proc. 2004 American Control Conference, June 30 - July 2, Boston MA.

[6] Cox, D.R. (2006), Principles of Statistical Inference, Cambridge: Cambridge University Press.

[7] Cox, D.R., and D.V. Hinkley (1974), Theoretical Statistics, New York: Chapman and Hall.

[8] Deistler, M., K. Peternell and W. Scherrer (1995), "Consistency and Relative Efficiency of Subspace Methods," Automatica, Vol. 31, pp. 1865-1875.

[9] Eserin, P.K.N., (1999), "Applications of Canonical Variate Analysis to the Dynamic Modeling and Control of Drum Level in Industrial Boilers," Proc. 1999 American Control Conference, held in San Diego, CA, June 2-4, 1999.

[10] Furnival, G.M., R.C. Wilson, Jr. (1974), "Regression by Leaps and Bounds," Technometrics, Vol. 16, pp. 499-511.

[11] Hotelling, H. (1936). "Relations Between Two Sets of Variates", Biometrika, Vol. 28, pp. 321-377.

[12] Hurvich, C.M. and C.L. Tsai (1991), "Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models," Biometrika Vol. 78, pp. 499-510.

[13] Juricek, B.C., D.E. Seborg, and W.E. Larimore (2004), "Fault Detection Using Canonical Variate Analysis," Ind. Eng. Chem. Res., Vol. 43, pp. 458-474.

[14] Larimore, W.E. (1983a). "Predictive Inference, Sufficiency, Entropy, and an Asymptotic Likelihood Principle", Biometrika, Vol. 70, pp. 175-81.

[15] Larimore, W.E. (1983b). "System Identification, Reduced-Order Filtering and Modeling Via Canonical Variate Analysis", Proc. 1983 American Control Conference, H.S. Rao and T. Dorato, Eds., pp. 445-51. New York: IEEE.

[16] Larimore, W.E. (1989), "System Identification and Filtering of Nonlinear Controlled Markov Processes By Canonical Variate Analysis, " Final Report for Air Force Office of Scientific Res., Computational Engineering, Inc, 1989. Summarized in Larimore (1992b).

[17] Larimore, W.E. (1990a), "Canonical Variate Analysis for System Identification, Filtering, and Adaptive Control, " Proc. 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, December, Vol. 1, pp. 635-9.

[18] Larimore, W.E. (1990b), "Order-Recursive Factorization of the Pseudoinverse of a Covariance Matrix", IEEE Trans. of Automatic Control, Vol. 35, pp. 1299-1303.

[19] Larimore, W.E. (1992a), ADAPT$_X$ Automated System Identification Software Users Manual, Adaptics, Inc, 40 Fairchild Drive, Reading, MA 01867.

[20] Larimore, W.E. (1992b), "Identification and Filtering of Nonlinear Systems Using Canonical Variate Analysis," Nonlinear Modeling and Forecasting, Eds. M. Casdagli and S. Eubank, pp. 283-303. Reading, MA: Addison-Wesley.

[21] Larimore, W.E. (1997a), "Optimal Reduced Rank Modeling, Prediction, Monitoring, and Control Using Canonical Variate Analysis," IFAC Internat. Symp. on Advanced Control of Chemical Processes, Banff, Canada, June 9-11, 1997.

[22] Larimore, W.E. (1997b), "System Identification of Feedback and 'Causality' Structure using Canonical Variate Analysis," Preprints 11th IFAC Symposium on system Identification, held July 8-11, 1997, Fukuoka, Japan, Vol. 3, pp. 1101-6.

[23] Larimore, W.E. (1999), "Automated Multivariable System Identification and Industrial Applications," Proc. 1999 American Control Conference, June 24, 1999, San Diego, CA, pp. 1148-1162.

[24] Larimore, W.E. (2002). Reply to 'Comment on 'Order-recursive factorization of the pseudoinverse of a covariance matrix' '. IEEE Trans. Automat. Contr., 47, pp. 1953-7.

[25] Larimore, W.E. (2003). "Inferring Multivariable Delay and Seasonal Structure for Subspace Modeling". Preprints 13th IFAC Symp. on System Identification, August 27-29, 2003, Rotterdam, Netherlands.

[26] Larimore, W.E. (2004), "Large Sample Efficiency for ADAPTx Subspace System Identification with Unknown Feedback", IFAC Symposium on Dynamics and Control of Process Systems, held July 5-7, Boston, MA.

[27] Larimore, W.E. (2005), "Maximum Likelihood Subspace Identification for Linear, Nonlinear, and Closed-loop Systems," Proc. American Control Conference, held June 6-8, 2005, Portland, OR.

[28] Larimore, W.E. (2006), "Selecting the Past and Future for Subspace Identification of Nonlinear Systems with Feedback and Additive Noise," Proc. American Control Conference, held June 14-16, 2006, Minneapolis, MN.

[29] Rao, C.R. (1966). Generalized inverse for matrices and its application in mathematical statistics. In Festschrift Volume for J. Neyman, Research Papers in Statistics edited by F.N. David. New York: Wiley. pp. 263-299.

[30] Tong, H. (1990), Non-linear Time Series: a Dynamical Systems Approach, Oxford: Oxford University Press.

[31] Van Overschee, P., and B. De Moor (1994), "A Unifying Theorem for Three Subspace System Identification Algorithms," American Control Conference, pp. 1645-1649, June 29-July 1, 1994, Baltimore, MD.

[32] Verhaegen, M. (1994), "Identification of the Deterministic Part of MIMO State Space Models given in Innovations Form from Input-output Data," Automatica, Vol. 30, pp. 61-74.

[33] Wang, Y., D.E. Seborg, and W.E. Larimore (1997), "Process Monitoring Using CVA and PCA", Proc. IFAC ADCHEM Sympos., pp. 523-528, Banff, Canada.