Nonlinear black box identification of a Selective Catalytic Reduction system

Soma Tayamon ∗ Darine Zambrano ∗ Torbjörn Wigren ∗ Bengt Carlsson ∗

∗ Uppsala University, Department of Information Technology, Division of System and Control, Uppsala, Sweden

Abstract: This paper discusses the identification of linear and non-linear black-box models for describing a diesel engine selective catalytic reduction (SCR) system. SCR aftertreatment systems form an important technology for reducing the NOx produced by diesel engines, and therefore good models are essential for the control of these systems. This paper compares a linear and a non-linear model for identification of the system. The output signals of the SCR were generated from 4 measured input signals, using a simulated 18 state model. The experiments with a recursive prediction error method, RPEM, with only 2 states show that the system can be accurately approximated with a much simpler model. The RPEM estimates 16 unknown parameters while the linear model uses 9 parameters. The results were compared based on the model fit and it was clear from the validation data set that the non-linear model gives better results and captures more of the system dynamics as compared to the linear model. A comparison of the RPEM using the midpoint integration method and the Euler method for discretisation was also made for the models. The results clearly show that the more accurate discretisation algorithm results in a better model fit.

Keywords: Non-linear systems, prediction error methods, recursive algorithms, selective catalytic reduction systems, diesel engines, system identification.

1. INTRODUCTION

Heavy duty vehicles are important in our society. These vehicles are often powered by diesel engines and although fuel efficient, the diesel engines emit a substantial amount of nitrogen oxide (NOx) which is dangerous for both human beings and the environment. NOx includes NO and NO2 and these particles play an important role in the formation of harmful particulate matter, ground-level ozone (smog) and acid rain. The main difficulty in NOx production is the direct correlation with fuel efficiency. A more efficient engine will result in higher NOx emissions. Since NOx reduction is important for environmental reasons while it must be kept cost efficient, the correlation between fuel efficiency and NOx reduction makes engine tuning a difficult task. New efficient methods for reducing these emissions, while keeping the fuel efficiency as good as possible are hence necessary, and therefore aftertreatment systems are becoming more important. Today, selective catalytic reduction, SCR, utilising ammonia as a reduction agent is widely used for this purpose. In the 1970’s the first SCR systems were developed for use in stationary applications such as thermal power plants (Belli et al. (1996)), giving sufficient reduction of the NOx. Later on, the technique was used in vehicle applications but since the volume of the catalytic converter is limited, certain problems were encountered.

Another important complication of the SCR is the emitted unreacted ammonia. This ammonia is released at the catalyst tailpipe and is created either due to low temperature or over-dosing of urea. At the same time too little urea is not sufficient for the reduction of the NOx. To achieve a balance between the injected urea, the tailpipe NOx and NH3, and to be able to use the SCR optimally, good models of the system are essential.

Several authors have dedicated their work to this subject and explained the systems by different complex models that model the behaviour of the SCR systems. See for example the work of Dolanc et al. (2001). In the work of Ericson (2009), the author proposed an approach of using two wall layers and 6 segments to physically describe the catalyst with two states in each segment. From there, the differential equations were derived using the chemical reactions. In the work of Schär et al. (2006), the system is modelled using the same states, namely the temperature inside the catalyst, $T_{in}$, and the ammonia coverage, $\theta_{NH3}$. The ammonia coverage is defined as the stored ammonia inside the catalyst over the ammonia storage capacity. Another approach to model the system is described by Feng (2010) where the system has three states, the NOx concentration, the NH3 concentration and the ammonia coverage in a two-cell structured SCR model. However, the total number of parameters is large, reducing the usefulness of the models for controller design.

It is also important to consider that the catalyst, in the above references, is modelled by means of grey-box identification, meaning that the system is modelled based on the physical properties of the system (Bohlin (1994)). This results in a model with a high accuracy. Nevertheless, the major drawback of the grey-box approach is the time consuming modelling process and the high cost it brings.
about. Therefore it is plausible to address the modelling problem using black-box system identification techniques (Ljung (1997)). Since the process is varying significantly with driving conditions, it is a further advantage if recursive techniques can be applied (Ljung and Söderström (1983)).

This paper presents a black-box approach to identify the SCR catalyst aftertreatment process for mobile heavy duty applications. The reduction process shows a non-linear behaviour as discussed in the work of Ericson (2009), Schär et al. (2006) and Feng (2010). Therefore, in this paper, the identification is based on a non-linear black-box model for the process. The objectives of this paper are to represent the SCR system using a black-box model and to obtain this by applying the identification algorithm described by Tayamon and Wigren (2010) to the SCR catalyst aftertreatment process. In particular the gains of introducing a non-linear model is of interest. A further objective is to study the effect of the choice of discretisation scheme in the RPEM algorithm and to qualify any gains of the midpoint integration introduced in Tayamon and Wigren (2010), as compared to the Euler integration in Wigren (2006).

The paper is organized as follows. A brief explanation of the SCR system is given in Section 2, while the identification algorithm is presented in detail in Section 3. The identification of the SCR system is detailed in Section 4, and the results are illustrated and compared in Section 5. Finally conclusions and future work are summarised in Section 6.

2. SCR SYSTEM DESCRIPTION

For automotive applications, the selective catalytic reduction exhaust gas aftertreatment system commonly consists of a honeycomb monolith, where several chemical reactions take place, combined with a dosification system of a reduction agent. Fig. 1 shows the main components of the system: the urea injection system, the catalyst and the sensors for flue flow, temperature and concentrations.

The SCR system consists of two main stages. Firstly, a reduction agent, in this case urea, is injected upstream through a nozzle and mixed with the exhaust flow at the input of the catalyst. Urea is contained in a harmless aqueous solution commercially named AdBlue, which consist of 32.5% of urea. Urea is converted to ammonia through the following chemical reaction:

\[
\text{CO(NH}_2\text{)}_2 + \text{H}_2\text{O} \rightarrow 2\text{NH}_3 + \text{CO}_2
\]

Secondly, the ammonia is partially adsorbed on the surface of the catalyst, where finally the dominant reactions occur in the catalyst, i.e. the ammonia reacts with the NOx emitted by the engine. NOx is composed primarily of NO with lesser amounts of NO2. The resulting compounds of the reaction is nitrogen gas (N2) and water (H2O).

The two main reactions of the SCR are:

- Standard SCR reaction

\[
4\text{NH}_3 + 4\text{NO} + \text{O}_2 \rightarrow 4\text{N}_2 + 6\text{H}_2\text{O}
\]

- Fast SCR reaction

\[
4\text{NH}_3 + 2\text{NO} + 2\text{NO}_2 \rightarrow 4\text{N}_2 + 6\text{H}_2\text{O}
\]

The NO2 reaction, being slower than (2) and (3) and therefore negligible, consist of:

- NO2 SCR reaction

\[
8\text{NH}_3 + 6\text{NO}_2 \rightarrow 7\text{N}_2 + 12\text{H}_2\text{O}
\]

The efficiency of the SCR system is usually evaluated by the amount of NOx reduced within the catalyst, and by the ammonia slip, which represents the unreacted ammonia.

3. THE IDENTIFICATION ALGORITHM

3.1 The model

The identification in this paper was performed using a recursive prediction error method with a restricted black-box parametrisation. The identification model is of output error (OE) type and is described as

\[
x^{(1)} = \begin{pmatrix}
x_1^{(1)} \\
\vdots \\
x_{n-1}^{(1)} \\
x_n^{(1)}
\end{pmatrix} = \begin{pmatrix}
x_2 \\
\vdots \\
x_n \\
f(x, u, \theta)
\end{pmatrix} \equiv f(x, u, \theta)
\]

\[
y = (1 \quad 0 \quad \ldots \quad 0) \cdot x.
\]  

where \( x = (x_1 \quad x_2 \ldots \quad x_{n-1} \quad x_n)^T \) is the state vector and \( \theta \) is the unknown parameter vector. The input vector \( u(t) \) is given by

\[
u(t) = (u_1(t) \ldots u_{n_1}(t) \ldots u_k(t) \ldots u_{n_k}(t))^T,
\]

and the output vector \( y(t) \) is given by

\[
y(t) = (y_1(t) \ldots y_{p}(t))^T.
\]

Here, the superscript \( n_i \) denotes differentiation \( n_i \) times. See Wigren et al. (2010) for more details. As is shown in Theorem 1 in Wigren (2006) the model described in (5) can be used to model also ODEs with general right-hand sides, locally in the state space. The model in (5) concentrates the non-linearity to one component of the equation, minimising the risk for over-parametrisation. The right-hand side function, \( f(x, u, \theta) \), is chosen to be a polynomial

\[
f(x_1, \ldots, x_n, u_1, \ldots, u_{(n_1)}, \ldots, u_{(n_k)}, \theta) = \sum_{i_1=0}^{I_{x_1}} \cdots \sum_{i_n=0}^{I_{x_n}} \cdot \sum_{i_{n_1}=0}^{I_{u_1}} \cdots \sum_{i_{n_k}=0}^{I_{u_{n_k}}}
\]
This polynomial can be written as a regressor vector generated from the state vector and the input vector multiplied by an unknown parameter vector as follows
\[
f(x, u, \theta) = \varphi^T (x(t, \theta), u(t)) \theta. \quad (8)
\]
The main advantage of this model is the wide range of applications that it covers for identification. Therefore it could be applied to identify new catalysts, from which new controllers for the systems can be designed.

### 3.2 The RPEM

The algorithm used for identification in this paper is fully described by Tayamon and Wigren (2010) and is based on the prediction error:
\[
\varepsilon(t, \theta) = y_m(t, \theta) - y(t, \theta),
\]
using the minimisation criterion
\[
V(\theta) = \frac{1}{2} E [\varepsilon^T(t, \theta) \Lambda^{-1}(t, \theta) \varepsilon(t, \theta) + \det(\Lambda(t, \theta))], \quad (9)
\]
where \(y_m(t)\) is the measured output and where \(\Lambda(t, \theta)\) is the unknown covariance matrix. To describe the RPEM, the model in (5) is discretised using the midpoint integration algorithm. To improve the numerical properties of the algorithm, a scaling factor \(\alpha\) is applied to the sampling time such that \(T_s = \alpha T\). For a detailed decimation of the effect of this scaling on the results of the identification, see Tayamon and Wigren (2010) and Wigren (2005).

\[
\begin{pmatrix}
  x_1(t + T) \\
  \vdots \\
  x_{n-1}(t + T) \\
  x_n(t + T)
\end{pmatrix} = \begin{pmatrix}
  x_1(t) \\
  \vdots \\
  x_{n-1}(t) \\
  x_n(t)
\end{pmatrix} + T_s \begin{pmatrix}
  x_2(t) + \frac{T_s}{2} x_3(t) \\
  \vdots \\
  x_{n-1}(t) + \frac{T_s}{2} \int f (x(t), u(t), \theta) \\
  f (x(t) + \frac{T_s}{2} \int f (x(t), u(t), \theta))
\end{pmatrix}
\]
\[
y(t) = (1 \ 0 \ \ldots \ 0) x(t), \quad (10)
\]
where \(f(\cdot)\) is defined by (5). Finally, the RPEM is formulated using the Gauss Newton minimisation algorithm of Ljung and Söderström (1983), as
\[
\Lambda(t) = \Lambda(t - T) + \frac{\mu(t)}{T} (y(t) \psi^T(t) - \Lambda(t - T))
\]
\[
R(t) = R(t - T) + \frac{\mu(t)}{T} (\psi(t) \Lambda^{-1}(t) \psi^T(t) - R(t - T))
\]
\[
\dot{\theta}(t) = \left[ \theta(t - T) + \frac{\mu(t)}{T} \Lambda^{-1}(t) \psi(t) \varepsilon(t) \right]_{\Delta \theta}. \quad (11)
\]
where \(\mu(t)\) is the gain sequence and \(R(t)\) is the estimate of the Hessian. \(\Delta \theta\) is the model defining the allowed values of the estimated parameters. To ensure stability, the model set is restricting the linearised model from becoming unstable. The updating is stopped if the parameter values are outside the model set. The resulting algorithm parallels equation (21) of Wigren (2006).

Due to physical boundaries, the output NO\textsubscript{x} can never become negative. Therefore, to improve the performance of the algorithm and keeping the physical properties, \(y(t) = (x_1(t))_{\text{sat}}\), is introduced, which is a saturation of the state.
\[
y(t) = x_1(t) \quad x_1(t) \geq 0
\]
\[
y(t) = 0 \quad x_1(t) < 0
\]
The derivative \(\frac{dy(t)}{dx}\) is affected in the same way:
\[
\frac{dy(t)}{dx} = (1 \ 0) \quad x_1(t) \geq 1
\]
\[
\frac{dy(t)}{dx} = (0 \ 0) \quad x_1(t) < 0.
\]

### 4. IDENTIFICATION SETUP

#### 4.1 Scaling

The order of magnitude of the input and output signals differ largely from each other. To obtain a data set which is easy to identify they are normalised by scaling using the standard deviation (SD) and zero mean signals of comparable amplitude are obtained. The values used for preprocessing of the data is shown in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>(y)</th>
<th>(u_1)</th>
<th>(u_2)</th>
<th>(u_3)</th>
<th>(u_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.325 \times 10^{-3}</td>
<td>0.619 \times 10^{-3}</td>
<td>8.418</td>
<td>533.97</td>
<td>3.596</td>
</tr>
<tr>
<td>SD</td>
<td>0.2667 \times 10^{-3}</td>
<td>0.482 \times 10^{-3}</td>
<td>11.212</td>
<td>48.130</td>
<td>1.338</td>
</tr>
</tbody>
</table>

#### 4.2 Signal selection

For identification of the model, a simulator based on the SCR model described by Ericson (2009) has been used to generate the output while the input signals were all measured on a real diesel engine. As seen in Fig. 1, the simulator uses the following measured input signals:
\[
\begin{align*}
  u_{1,r} &= c_{\text{NO}_x, \text{in}} \\
  u_{2,r} &= c_{\text{NH}_3, \text{in}} \\
  u_{3,r} &= T_{\text{in}} \\
  u_{4,r} &= n_{ef}
\end{align*}
\]
where, \(c_{\text{NO}_x}\) and \(c_{\text{NH}_3}\) corresponds to the concentration of NO\textsubscript{x} in [fraction] and the injected urea NH\textsubscript{3} in [g/min]
output signals: 

\[
y_1 = cNOx_{\text{out}} \\
y_2 = cNH_3_{\text{out}}.
\]

In real life application, there is only one sensor measuring both outputs. The results obtained from this sensor is then used to calculate the corresponding output NOx and ammonia, which can cause cross coupling. The behaviour of four input signals and the main output signal for a World Harmonised Transient Cycle (Steven (2001)), WHTC, can be seen in Fig. 2.

As previously mentioned, the system consists of three input signals coming from the engine, and a fourth signal being the injected urea. The third input signal \( u_3(t) \) is modified and redefined as an exponential of the temperature:

\[
u_3^*(t) \equiv e^{-\frac{u_3(t)}{a}}.
\]

The choice of this input signal is based on the improved results it brings about and is introduced based on the information about the original physical model from Schär et al. (2006). Here, the constant \( a \) is chosen based on the values described in Schär et al. (2006) to be equal to \( 10^4 \).

The final input signals used for the algorithm are therefore \( u_1(t), u_2(t), u_3^*(t) \) and \( u_4(t) \) of (13). It was noted, using the linear model as a reference, that the results were improved by including the derivatives of \( u_1(t), u_2(t) \) and \( u_4(t) \) in the model described by (5) and (6). The derivatives are therefore included in the models defined below.

**Remark:** As mentioned in the work of Wigren (2005), using the derivatives of the input signals has consequences for the choice of these signals. Note that, in this case, this choice has been made based on performance of the linear model when including the different derivatives. It might be possible to make further improvements by considering dedicated input signal design in later work.

The real system described by the simulator has two output signals \( y_1(t) \) and \( y_2(t) \), (14). Since the algorithm is of MISO type, \( y_1(t) \), being the output NOx, is selected to be the only output of the chosen model. A second order model was chosen and the input signals were of the first degree. The selected model can be summarised as follows:

\[
\dot{x}_1 = x_2 \\
\dot{x}_2 = f \left( x_1, x_2, u_1, u_2, u_3, u_4, u_1^{(1)}, u_2^{(1)}, u_4^{(1)} \right).
\]

### 4.3 Model simplification

The model described by (16) gives rise to a total number of 512 unknown parameters, where 9 of them correspond to linear signals and the remaining parameters describe the polynomial coefficients of the non-linearities in the states and the inputs. To be able to achieve satisfying identification results and reducing the risk of over-parametrisation using the algorithm, some of these parameters were excluded. During the identification process, each parameter was tentatively removed to reduce the total number of parameters. When the model fit was improved, the parameter was excluded and if not, it was included. This way, the number of parameters to identify are reduced resulting in a simpler model and consequently less computational time is needed. The final result is a system containing 16 parameters. A further advantage is that the convergence properties of the algorithm is improved and the risk of instability is reduced. For the selected non-linear model, the regressor vector of the identification algorithm is:

\[
\varphi = \begin{pmatrix} u_3^{(1)} & u_2^{(1)} & u_1^{(1)} & u_4 & u_3 & u_2 & u_1 & x_1 & x_2 & x_1 \\
 u_3^*u_4^{(1)} & u_2u_4^{(1)} & u_2u_3u_4 & u_1u_4^{(1)}u_2^{(1)} \\
u_1u_3^*u_1^{(1)} & x_2u_1 & x_1u_2 \end{pmatrix}.
\]

The reason for eliminating the bias term is that the estimation of this term dominates the identification process, giving poor results. For comparison, the system was also identified using only the linear terms in the model i.e.

\[
\varphi = \begin{pmatrix} u_3^{(1)} & u_2^{(1)} & u_1^{(1)} & u_4 & u_3 & u_2 & u_1 & x_2 & x_1 \end{pmatrix}.
\]

### 4.4 Initialisation

Because of the RPEM’s sensitivity to initial parameters and local minima, the initialisation step is a very important part of the identification. Therefore, to obtain optimal and reasonable initial parameters, the initialisation algorithm proposed by Brus et al. (2008) has been used. The algorithm is a Kalman filter based method that estimates the initial parameters for a given model based on the model described in (5). To tune the initialisation algorithm, the following values were used: \( P(0) = 10 \), which is the scale factor of the covariance matrix, i.e. \( P(0) = P(0)I \), further:

\[
R_1 = \begin{pmatrix} R_{1,x} & 0 \\
0 & R_{1,\theta} \end{pmatrix}.
\]
is the covariance matrix of the process noise. In the algorithm, these values were chosen as $R_{1,x} = 50$ and $R_{1,\theta} = 0$. The process noise of the parameter vector $\theta$ was chosen as zero (time invariant). Finally the covariance matrix of the measurement noise $R_{2}$ was set to 3. Then, the RPEM was initialised using the parameter vector obtained by the initialisation algorithm.

For the initialisation of the RPEM, some tuning and initial values other than the parameter vector need to be set. The starting value of the Hessian was chosen as a scale factor times the identity matrix, $R(0) = 100I$, the initial noise variance $\Lambda(0) = 0.1I$ and the initial state vector $x(0) = (y(0), 0)$. An important part of the algorithm is the sampling time scale factor and it was chosen as $\alpha = 1.5$, cf. Wigren (2005). For further improvements of the results, the algorithm was initialised once more with the previously obtained parameters until no more improvement was achieved and the parameters had converged.

5. EXPERIMENTAL RESULTS

5.1 Validation

Since the WHTC data is the only available data at our disposal, to validate the results of the algorithm, the data set was divided into two separate parts. To be able to use the algorithm to capture all NOx dynamics 75% of the data was used for identification and the remaining 25% was used for validation. To measure the accuracy, the percentage fit $\gamma$ was calculated comparing the estimated and real output using:

$$\gamma = \left(1 - \frac{||y - y_{es}||}{||y - \bar{y}||}\right) \cdot 100 \quad (19)$$

where $y_{es}$ is the estimated output and $y$ is the output NOx from the simulator, and $\bar{y}$ is the mean value of the signal $y$. Eventually, the model (17) was obtained. Fig. 3 illustrates the results obtained by the second order linear model on the identification data, and the corresponding results on the validation data is seen in Fig. 4, with the percentage fit of 64.52% and 41.15%, respectively. It is clear that the linear model is not capable of fully describing the system dynamics, especially for the validation data set. When using the non-linear terms in (17), the results were slightly improved, as shown in Fig. 5 and the model fit for the identification data set was increased to 67.31%. Obviously, the non-linear model is better at catching the system dynamics at the beginning of the data set. The major improvements are however seen in the validation data set, were the model fit is improved from 41.68% to 61.54%, see Fig. 6. These results show that the non-linear model is better at describing the areas where the ammonia is mostly active, see Fig. 2.

5.2 Algorithm comparison

To analyse the effect of a more accurate discretisation algorithm suggested by Tayamon and Wigren (2010) compared to the Euler based RPEM of Wigren (2006), the two identification algorithms were applied on the same data set using the same initial parameters. The results were then compared, and are shown in Table 2. It is also important to mention that the identification and validation was made on the same data set in this experiment.

Table 2. Comparison between the Euler and the midpoint integration algorithm.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of parameters (Model)</th>
<th>Percentage fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euler</td>
<td>9 (Linear model)</td>
<td>40.68%</td>
</tr>
<tr>
<td>Midpoint</td>
<td>9 (Linear model)</td>
<td>41.68%</td>
</tr>
<tr>
<td>Euler</td>
<td>16 (Non-linear model)</td>
<td>55.64%</td>
</tr>
<tr>
<td>Midpoint</td>
<td>16 (Non-linear model)</td>
<td>61.54%</td>
</tr>
</tbody>
</table>

It is clear that the more accurate integration algorithm improves the results somewhat.

6. CONCLUSIONS

In this paper, a non-linear black box model has been used to recursively estimate an SCR process and the results were compared to a linear model. The non-linear model was shown to be more accurate than the linear model.
The main reason for this choice of identification approach was to simplify the modelling process and reduce the complexity that the first principle based model brings about. Even though the system has four input signals, it can be described using only 16 parameters. The identification method was an RPEM based on a restricted black-box state space model and the model was discretised using the midpoint integration algorithm. The results were compared to the Euler based algorithm, showing an improvement of the results using a more accurate discretisation algorithm.

Interesting topics for further research would be to design non-linear controllers using the identified model and to evaluate these. Also, the current model could be improved using user defined input signals.

ACKNOWLEDGEMENTS

This work is supported by the Swedish Energy Agency (project 32299-1) which is gratefully acknowledged. The authors would like to thank Scania AB for the data provided for identification.

REFERENCES


