A computationally efficient Kalman filter based estimator for updating look-up tables applied to NO\textsubscript{x} estimation in diesel engines

C. Guardiola\textsuperscript{a}, B. Pla\textsuperscript{a}, D. Blanco-Rodriguez\textsuperscript{a, b}, L. Eriksson\textsuperscript{b}

\textsuperscript{a} CMT Motores Térmicos, Universitat Politècnica de València, Camino de Vera s/n, E-46022 Valencia, Spain
\textsuperscript{b} Vehicular Systems, Department of Electrical Engineering, Linköping University, 5-581 83 Linköping, Sweden

1. Introduction

Legislation on emissions strengthens the importance of electronic controls and embedded software in automotive engines (Ebert & Jones, 2009). Focusing on nitrogen oxides (NO\textsubscript{x}) emissions, selective catalyst reduction system (SCR) or lean NO\textsubscript{x} trap (LNT) requires from reliable information about exhaust gas content, the former for an appropriate urea dosing (Twigg, 2007) and the latter for a proper regeneration (Chen, Wang, Haskara, & Zhu, 2012). Commercial electronic control units (ECU) are mainly programmed with fixed maps and feedback for emissions control is often far from the exhaust, e.g. mass air flow or boost pressure sensor. This argument can be extended to soot or CO\textsubscript{2}, while exhaust oxygen measurement is more usual.

An accurate and fast NO\textsubscript{x} signal would permit its usage for real-time tasks but still some issues need to be solved. NO\textsubscript{x} sensors have been commonly used in test benches, e.g. gas analyzers or fast measurement systems. But until the appearance of zirconia-based (ZrO\textsubscript{2}) potentiometric sensors, there had been no choice of using these in commercial engines, because of cost, size and response limitations (Kato, Nakagaki, & Ina, 1996; Zhuiykov & Miura, 2007). However, even with the sensor there are problems using the raw output signal in real-time functions, in particular the delay from engine to sensor and the response of the sensor; see e.g. Manchur and Checkel (2005). For that purpose, Kalman filter (KF) based methods are explored as having the interesting property of estimating NO\textsubscript{x} in covariance matrix. Nevertheless, when coping with large systems, the computational burden is high, in terms of time and memory, compromising its implementation in commercial electronic control units. However look-up table estimation has a structure, that is here exploited to develop a memory and computationally efficient approximation to the KF, named Simplified Kalman filter (SKF). Convergence and robustness is evaluated in simulation and compared to both a full KF and a minimal steady-state version, that neglects the variance information. SKF is used for the online calibration of an adaptive model for NO\textsubscript{x} estimation in dynamic engine cycles. Prediction results are compared with the ones of the benchmark model and of the other methods. Furthermore, actual online estimation of NO\textsubscript{x} is solved by means of the proposed adaptive structure. Results on dynamic tests with a diesel engine and the computational study demonstrate the feasibility and capabilities of the method for an implementation in engine control units.

Corresponding author. Tel.: +34 963879233.
E-mail addresses: carguaga@mot.upv.es (C. Guardiola), benplamo@mot.upv.es (B. Pla), dablarod@gmail.com, dablarod@mot.upv.es (D. Blanco-Rodriguez), lare@isy.liu.se (L. Eriksson).

Article history:
Received 23 July 2012
Accepted 21 June 2013

Keywords:
NO\textsubscript{x}
Kalman filter
Adaptive model
Look-up tables
Diesel

**Abstract**

NO\textsubscript{x} estimation in diesel engines is an up-to-date problem but still some issues need to be solved. Raw sensor signals are not fast enough for real-time use while control-oriented models suffer from drift and aging. A control-oriented gray box model based on engine maps and calibrated off-line is used as benchmark model for NO\textsubscript{x} estimation. Calibration effort is important and engine data-dependent. This motivates the use of adaptive look-up tables. In addition to, look-up tables are often used in automotive control systems and there is a need for systematic methods that can estimate or update them on-line. For that purpose, Kalman filter (KF) based methods are explored as having the interesting property of tracking estimation error in a covariance matrix. Nevertheless, when coping with large systems, the computational burden is high, in terms of time and memory, compromising its implementation in commercial electronic control units. However look-up table estimation has a structure, that is here exploited to develop a memory and computationally efficient approximation to the KF, named Simplified Kalman filter (SKF). Convergence and robustness is evaluated in simulation and compared to both a full KF and a minimal steady-state version, that neglects the variance information. SKF is used for the online calibration of an adaptive model for NO\textsubscript{x} estimation in dynamic engine cycles. Prediction results are compared with the ones of the benchmark model and of the other methods. Furthermore, actual online estimation of NO\textsubscript{x} is solved by means of the proposed adaptive structure. Results on dynamic tests with a diesel engine and the computational study demonstrate the feasibility and capabilities of the method for an implementation in engine control units.
Without NOx sensors, modeling (or virtual sensing) is possible, see e.g. Schilling, Amstutz, and Guzzella (2008) that develops a real time NOx model using maps, Winkler-Ebner, Hirsch, Del Re, Klinger, and Mistelberger (2010) that designs a virtual NOx sensor for selective catalyst reduction (SCR) control and diagnosis or other examples. Problem is highly non-linear and some examples for coping with this are Takagi–Sugeno fuzzy models (Lughofer, Macian, Guardiola, & Klement, 2011; Takagi & Sugeno, 1985), Hammerstein–Wiener (Falck et al., in press), or neural networks (Yen & Michel, 1991). Nevertheless, in commercial engine ECUs the prevailing approach is to use look-up tables to model nonlinear and operating point dependent behaviors because of the simple programming although it attaches a big calibration effort. These models need to be calibrated but in general will suffer from drift problem. Section 3 presents a gray box model used as benchmark for comparison latter.

A solution for drift correction is adaptive modeling. Main algorithms for online adaptation are recursive least squares (RLS) and Kalman filter (KF). Although both copes with system drift and aging, the latter tracks estimation aging by solving the statistics of estimation error at every iteration (by the covariance matrix). This paper analyzes the computational aspects of look-up table aging, the latter tracks estimation aging by solving the statistics of the estimation error at every iteration (by the covariance matrix). This paper analyzes the computational aspects of look-up table aging, the latter tracks estimation aging by solving the statistics of estimation error at every iteration (by the covariance matrix).

Without NOx sensors, modeling (or virtual sensing) is possible, see e.g. Schilling, Amstutz, and Guzzella (2008) that develops a real time NOx model using maps, Winkler-Ebner, Hirsch, Del Re, Klinger, and Mistelberger (2010) that designs a virtual NOx sensor for selective catalyst reduction (SCR) control and diagnosis or other examples. Problem is highly non-linear and some examples for coping with this are Takagi–Sugeno fuzzy models (Lughofer, Macian, Guardiola, & Klement, 2011; Takagi & Sugeno, 1985), Hammerstein–Wiener (Falck et al., in press), or neural networks (Yen & Michel, 1991). Nevertheless, in commercial engine ECUs the prevailing approach is to use look-up tables to model nonlinear and operating point dependent behaviors because of the simple programming although it attaches a big calibration effort. These models need to be calibrated but in general will suffer from drift problem. Section 3 presents a gray box model used as benchmark for comparison latter.

A solution for drift correction is adaptive modeling. Main algorithms for online adaptation are recursive least squares (RLS) and Kalman filter (KF). Although both copes with system drift and aging, the latter tracks estimation aging by solving the statistics of the estimation error at every iteration (by the covariance matrix). This paper analyzes the computational aspects of look-up table aging, the latter tracks estimation aging by solving the statistics of the estimation error at every iteration (by the covariance matrix). This paper analyzes the computational aspects of look-up table aging, the latter tracks estimation aging by solving the statistics of the estimation error at every iteration (by the covariance matrix). This paper analyzes the computational aspects of look-up table aging, the latter tracks estimation aging by solving the statistics of the estimation error at every iteration (by the covariance matrix).

2. Experimental setup and engine data

Experimental data is obtained from a twin sequential-parallel turbocharged diesel engine. This engine is a 2.2-liter 4-cylinder common rail. Engine specifications are shown in Table 1. Conventional sensor set is used (mass air flow, boost pressure, etc.) and the following extra-sensors are added: a ZrO2 multilayer based NOx sensor (Kato et al., 1996) giving a twofold measurement of NOx ($y_{NOx}$) and oxygen in the tailpipe and a gas analyzer (HORIBA, 2001) in a long line connecting intake manifold and tailpipe as steady-state calibration standard. These sensors are used for collecting dynamic data and some of them as inputs for the model, removing the need of observing some of the variables.

A rapid prototyping system is connected via ETK to a bypass-allowed ECU, permitting commanding and receiving signals by means of coupling a real time system via CAN. This allows easy access to injection parameters (injection pressure or start of injection SOI), boost and EGR control set points, engine calibration and possibilities to easily test new programs and routines.

The test campaign includes data for steady-state and transient tests for the benchmark model calibration. Steady-state ones cover around 300 operating points, with nominal control inputs. Operating points are fixed and signals stability is checked for storing steady-state points. Transient tests include variations on the actuators and two dynamic cycles: a designed SDMP and NEDC. SDMP performs sharp variations on injected fuel mass and speed and is shown in Fig. 2. This is a good test bench for comparing after look-up table estimation results with the ones given by the model. NEDC is also used as application. For more details about experimental setup see Lughofer et al. (2011). SOI steps are used for NOx sensor characterization; the interested reader can go to Galindo et al. (2011) for further information.

<table>
<thead>
<tr>
<th>Engine technical data. For more information see Galindo et al. (2011).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroke (S)</td>
</tr>
<tr>
<td>Bore (D)</td>
</tr>
<tr>
<td>S/D</td>
</tr>
<tr>
<td>Number of cylinders (z)</td>
</tr>
<tr>
<td>Displacement</td>
</tr>
<tr>
<td>EGR</td>
</tr>
<tr>
<td>Turbocharging system</td>
</tr>
<tr>
<td>Valves by cylinder</td>
</tr>
<tr>
<td>Maximum power</td>
</tr>
<tr>
<td>Compression ratio</td>
</tr>
</tbody>
</table>

![Fig. 1. NOx sensor response to SOI steps. NOx presents a clear delay and filtering with respect to SOI signal. Dynamic response could be fitted with a first order delayed discrete filter. SOI unit is crank angle degree before the top dead center, while NOx is measured in ppm. Delay is in the order of 1 s while response time is about 0.75 s.](image1.png)

![Fig. 2. SDMP test profile with sharp variations in $m_f$, $n$ and NOx](image2.png)
3. A control-oriented model for NOx estimation

Literature about NOx modeling in diesel engines is extensive and some examples have been already referenced in the introduction. Physical models rely on first principles but uses complex structures that require big computing times. In addition, these are not free from drift problem. On the other hand, control-oriented models are often data driven. Here, different possibilities can be found. For example, heuristic approaches can derive maps from complex models (Schilling, Amstutz, Onder, & Guzzella, 2006) or engine data (Lughofer et al., 2011), building a model that is easy to integrate into the engine ECU, but lacking of extrapolation capabilities. This needs some filtering for catching dynamics and if implemented online, drift can be corrected with observers (Payri et al., 2012) or feedback control. Black box models rely on system identification (Karsson, Ekholm, Strandh, Tunestål, & Johansson, 2010) but are often operating point dependent and its adaptation is not an easy task. As an intermediate solution, gray box structures can use fundamental relationships with experimental fittings (Hirsch, Alberer, & del Re, 2008). In the following a gray box NOx model is provided for comparison with the online methods presented hereinafter.

3.1. NOx model used as benchmark

Thermal NOx formation (Lavoie, Heywood, & Keck, 1970) is the dominant mechanism in diesel, benefited from peak temperatures in the cylinder and lean conditions (excess of air). These conditions maximize engine efficiency and reduce soot emissions: the well-known soot-NOx trade-off, see e.g. Neef, Makkee, and Moulijn (1996). Three in-cylinder variables should be necessary for NOx reconstruction: peak temperature, oxygen concentration and residence time, but these cannot be directly measured. In-cylinder pressure may be used as input but signal treatment is still an issue (Guardiola, López, Martin, & García-Sarmiento, 2011). For a fast mean value NOx model, oxygen rate in the cylinder and engine operating conditions (usually speed n and injected fuel mass \(mf\)) seem to be suitable variables for characterizing NOx; see e.g. Ericson, Westerberg, Andersson, and Eggnl (2006), Galindo, Luján, Climent, and Guardiola (2007).

The model is set-point relative and the main equation is

\[ x_m = NOx_{ref}(n, mf) - \dot{C} \beta (\text{comb}, H, T_w, \ldots) \]

where subindex 0 indicates nominal data defined by manufacturer calibration. NOx_{ref} is NOx emissions as function of nominal operating point conditions (\(n\) and \(mf\)), and exponential is built around nominal values of inert gas \(F_{EGR}\), which expresses the real quantity of burned gas fraction, being \(F_{EGR}\) the EGR ratio and \(F_r\) the relative fuel-to-air ratio. EGR flow is solved by a mean value engine model (MVEM) and \(F_r\) is observed from \(mf\), \(Fr\) and oxygen output from ZrO2 sensor at the exhaust: see Appendix A for more details on the sub-model equations. \(F_{EGR,0}\), \(F_{r,0}\), which define the nominal values for \(F_{EGR}\) and \(F_r\), respectively, and \(k\) are mapped with a proper grid using \(n\) and \(mf\) as inputs. The set-point relative structure is suitable for minimizing errors around nominal, especially when real time actions are pursed and allows defining optimal control strategies. Effects of combustion modes, humidity \(H\), engine temperature \(T_w\) are considered with the correction factor \(C\), i.e. \(C = 1\) when no correction. \(T_w\) and \(H\) are directly measured on engine.

The model cannot be used for stand-alone simulation because it not only requires ECU signals but also is suitable for control and diagnosis and for offline NOx diagnosis after tests, i.e. if no NOx sensor is available. Nevertheless, model can be completed with a full air-path MVEM including turbocharger and combustion blocks; a good example of an air-path model can be found in Wahlström and Eriksson (2011).

For comparing with sensor signal, a first order discrete filter \(G(z)\) with a time delay \(r\) is applied to model output

\[ y_m = \frac{1-a}{1-ar} z^{-r} x_m \]

obtaining \(y_m\) where SOI steps are performed for characterizing \(a\) and \(r\). Sensor response time \(a\) can be modeled as constant but delay \(r\) has a non-linear behavior difficult to model. A detailed study on NOx sensor characterization is made in Galindo et al. (2011). In this work, \(r\) is constant as being sufficient for obtaining good results with the SDMP and NEDC for the model and updating methods. Anyway, a robust implementation should include \(r\) variability; the interested reader can go to Trimboli, Di Cairano, Bemporad, and Kolmanovsky (2012) who model delay in \(Fr\) signal from NOx sensor as function of air mass flow, engine operating conditions and speed.

Model is calibrated using steady-state measurements and dynamic tests minimizing \(\sum_{i=1}^{n_p} [y_{NOx} - y_{m(i)}]^2\) by least squares fitting where \(n_p\) is the number of points included. Model validation has been made with a set of engine cycles, including NEDC and SDMP. Fig. 3 shows results in the SDMP test comparing \(y_{NOx}\) and \(y_m\). The fitting is good and the model is capable of estimating \(y_{NOx}\) with a minimum error. Fig. 18 presented in Section 6.4 shows results of a non-optimized calibration for the NEDC and even though some drift appears, dynamic behavior is cached.

3.2. Motivation for updating look-up tables

Two problems can be underlined when working with control-oriented models. On one hand, the model accuracy is driven by the collection of the appropriate data and calibration of all the parameters. This is a hard and time consuming task. In fact, ECU has a big number of maps and parameters for engine and vehicle management. On the other hand, independently of how well the model has been calibrated there is inevitably a drift between the system and the model as the surrounding conditions changes and the engine ages. Data based models are highly sensitive to the calibration data set and will have problems with aging, manufacturing discrepancies, slowly varying parameters and other non-modeled variables. To show this, an old calibration set for the model (for maps and sensor dynamics) is used for simulating the SDMP test. Results can be seen in Fig. 4, where dynamics are well cached but a clear drift exists. Here is not only aging that affects, but test or ambient conditions.

These problems are similar for other model structures and therefore learning algorithms for look-up tables can be used for calibration and/or online adaptation. In the next, the paper focuses on the online adaptation of look-up tables and develops a
being $F$ the linearized process matrix and $H$ the linearized output matrix. From now, discussion is valid both for linear and non-linear systems and $KF$ will be used referring to Kalman filter based methods, including non-linear Extended version and standard one.

Noises $w_k \in \mathbb{R}^{n_u}$ and $v_k \in \mathbb{R}^{n_h}$ are assumed to be independent and both generated by Gaussian distribution with zero mean and covariance matrices $Q_k$ resp. $R_k$, defined by

$$E[w_k w_k^T] = Q_k$$
$$E[v_k v_k^T] = R_k$$

In applications these are often chosen to be constant, i.e. $Q$ and $R$, and diagonals.

Then, $x_k \in \mathbb{R}^{n_x}$ is the observation of the state vector $x_k$

$$\dot{x}_{k|k-1} = f(\hat{x}_{k-1}, u_k)$$
$$e_k = y_k - h(x_{k|k-1}, u_k)$$

(6a) (6b)

$$\hat{x}_k = \hat{x}_{k|k-1} + K_k e_k$$

(6c)

where $K_k$ Kalman gain is solved by the following iterative equation:

$$P_{k|k-1} = (F_k P_{k-1} F_k^T + Q)$$

$$K_k = P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1}$$

$$P_k = (I - K_k H_k) P_{k|k-1}$$

(7a) (7b) (7c)

where matrix $P_k$ is the covariance matrix of the state estimate error ($\text{Ljung, } 1999$).

$$P_k = E[\{x_k - \hat{x}_k\}^2]$$

(8)

4.2. Look-up tables

A look-up table $T \in \mathbb{R}^{n_{x_{\scriptscriptstyle \text{d}}} \times n_t}$ is defined as a $N$-Dimensional mapping $(T : \mathbb{R}^N \rightarrow \mathbb{R})$ defined by a grid in $N \in \mathbb{Z}^+$ dimensions, where each one has $n_t$ grid points. The mapping further relies on a multivariate interpolation $q(\cdot)$ to calculate the function value from the input using the grid for the interpolation variables and $T$. In automotive systems, the multivariate interpolation schemes are often linear in the dimensions, and this is the case that will be considered here. Without loss of generality the presentation will use 2D tables as being the most frequently occurring dimensions, since working with other dimensions is only a matter of reducing or increasing indexes. Then, for $N=2$,

$$T = [t_{ij}]$$

(9)

where $i = 1, \ldots, n_x$ and $j = 1, \ldots, n_t$ ($r$ stands for row and $c$ for column). The multivariate interpolation function for generating the output $y_k$ from input $u_k = [u_{1t}(k) \ u_{2t}(k)]^T$ can be expressed as

$$y_k = \text{vec}(q(u_k))^T \ \text{vec}(T)$$

where $\text{vec}(\cdot)$ is the vectorization transformation and $q(u_k)$ is the interpolation matrix that both selects the elements to be interpolated and contains the weights. In the 2D case $q(u_k)$ selects the 4 $[2^N]$ in the ND case) active elements $t_{i_j}$, $t_{i+j}$, $t_{i+1j}$, $t_{i_1j+1}$ (with $i,j$ fulfilling $u_{1k} \in [r_i, r_{i+1}]$ and $u_{2k} \in [c_j, c_{j+1}]$) and thus contains the following block with non-zero weights

$$\begin{bmatrix}
q(u_k)_{ij} & q(u_k)_{i+j} & q(u_k)_{i+1j} & q(u_k)_{i_1j+1} \\
\end{bmatrix} = \begin{bmatrix}
(1-\eta_{1k})(1-\eta_{2k}) & (1-\eta_{1k})\eta_{2k} & \eta_{1k}(1-\eta_{2k}) & \eta_{1k}\eta_{2k} \\
\eta_{1k}(1-\eta_{2k}) & (1-\eta_{1k})\eta_{2k} & \eta_{1k}(1-\eta_{2k}) & \eta_{1k}\eta_{2k} \\
\end{bmatrix}$$

(10)

where

$$\eta_{1k} = \frac{u_{1k} - r_i}{r_{i+1} - r_i}, \quad \eta_{2k} = \frac{u_{2k} - c_j}{c_{j+1} - c_j}$$

computationally efficient method for correcting drift, aging and indeed allowing self-calibration.

4. Updating look-up tables

In the literature different approaches for online learning of look-up tables can be found, Peyton Jones and Muske (2009) use a recursive least-squares method with a forgetting factor, Wu (2006) distributes proportionally the existing error on the estimation between the active parameters and Karlsson, Ekholm, Strandh, Johansson, and Tumestal (2008) use the subspace method for identifying NOx and soot emission models in heavy-duty diesel engines. However KF based approaches (see a complete book about KF methods in Simon, 2006) provide a systematic way for updating map parameters and it is also appealing for engine applications as it can also handle aging (see Höckerdal et al., 2011 for a discussion) both for parameters and estimation, i.e. parameters that have not been excited for a longer time are expected to have bigger drift with respect to those have been excited recently. KF manages this issue optimally in the sense of minimizing expected estimation error.

Nevertheless, when observing large maps there are high memory requirements and computational burden. The core of the problem is that a KF implementation for updating a look-up table with 20-by-20 elements gives rise to a covariance matrix of 400-by-400 elements and thereby also a significant computational burden when solving the Riccati equations. Therefore an important discussion of the paper is how the KF can be used or modified to get an efficient updating procedure for look-up tables without an important loss of properties.

4.1. Kalman filter basics

The main equations in the KF Kalman (1960) are recalled here for presenting the nomenclature used after although the reader familiarized with control engineering could skip this subsection. In the setting the data is assumed to be generated by the following discrete time system

$$x_k = f(x_{k-1}, u_k) + w_k$$
$$y_k = h(x_k, u_k) + v_k$$

(3a) (3b)

where $x_k \in \mathbb{R}^{n_x}$ represents the state vector, $u_k \in \mathbb{R}^{n_u}$ the input vector, $y_k \in \mathbb{R}^{n_y}$ the output vector. If $f$ and/or $h$ are non-linear a previous linearization step is required for the filter and then, elements $ij$ of $F_k$ and $H_k$ are obtained

$$F_{k,ij} = \left. \frac{\partial f}{\partial x_i} \right|_{x_k = x_i} \quad H_{k,ij} = \left. \frac{\partial h}{\partial x_i} \right|_{x_k = x_i}$$

(4)

Fig. 4. Model simulation $y_{m}$ and sensor signal $y_{n}$ in the SDMP test with a non-optimized calibration data set. Even though dynamics are well cached, $y_m$ is drifted with respect to $y_n$.
4.3. Modeling for learning, drift or aging in tables

In the setting, T models a nonlinear function and the interesting aspect is to allow the model to adapt to the system to either learn the system and/or follow the aging of the system. This is modeled in the standard way as a random walk process, where the table parameters are collected in a state vector \( x_k \in \mathbb{R}^{n_t \times n_t} \). The full model \( u_k \rightarrow v_k \) is

\[
x_k = x_{k-1} + w_k
\]

\[
y_k = \text{vec}(q(u_k))^T x_k + v_k
\]

(11a)

(11b)

Note that no uncertainties are allowed in the interpolation variable \( u_k \).

For convenience, the non-zero elements in \( \text{vec}(q(u_k)) \) are denoted \( q_{n_k}^o \in \mathbb{R}^{1 \times d} \) (\( o \) stands for observable) and the corresponding elements in the state vector, \( Q^o \in (\mathbb{R})^d \). Then, following expression is obtained for the output:

\[
y_k = q_{l+k}^o x_k^o + v_k
\]

(12a)

\[
q_{n_k}^o = [(1-\eta_1)k(1-\eta_2)k] \eta_1(1-\eta_2)k \eta_1k(1-k) \eta_2k
\]

(12b)

where \( Q \) and \( R \) are

\[
Q = \sigma_{n_k}^2 I_{n_r \times n_r}
\]

(13a)

\[
R = \sigma_v^2
\]

(13b)

and \( I_{n_r \times n_r} \) is the identity square matrix. If there is engineering application knowledge available, \( \sigma_{n_k}^2 \) might be selected individually for each table element building a vector \( \Sigma^o = [\sigma_{n_k}^o \ldots \sigma_{n_r}^o, \ldots, \sigma_{n_r}^o] \) that contains individual variances in such way \( Q = \Sigma^o I_{n_r \times n_r} \). In automotive, this is useful for considering order of magnitude of the parameters, i.e. absolute error will not be the same for low emissions at lower loads than highest peaks at higher ones, and on the other hand, if driving pattern exists, noise could be mapped over the table grid. This is linked with the foreseen probability that engine is running at a certain operating condition, i.e. elements aging in more usual grid areas will be slower and the opposite.

4.4. Kalman filter based method for updating look-up tables, KF

The EKF can now be used to observe \( x_k \), when measurements of \( y_k \) are gathered. At every iteration, \( K_i \) is calculated based on (7), where \( F = I_{n_r \times n_r} \) is constant and \( H_k = \text{vec}(q(u_k))^T \). Although only the active elements \( x_k^o \) are updated at every \( k \), all \( P_k \) elements enter in the equation, which leads to huge calculations and big required memory resources. This makes difficult the implementation in commercial ECUs.

There are several publications on computational aspects of Kalman filters, for example some have studied the filter optimization when different nonlinear functions are handled, e.g. [Charalampidis and Pavlidi 2011], [Jorgensen and Homsen, Madsen, and Kristensen 2007] and [Singer and Sea 1971]; the latter make an interesting study on the total number of operations required for the updating phase. Chandrasekar, Kim, and Bernstein (2007) present an interesting methodology when system order is extremely large by using the finite-horizon optimization technique for obtaining reduced-order systems. But key observation here is that the look-up table estimation has structural properties that can be exploited to reduce the computational and memory requirements significantly in a simple way.

To understand the problem, covariance matrix in the KF \( P \) is studied. For instance, a 2D look-up table with a size \( n_t \times n_t \) needs system \( F \) and covariance matrices \( P \) of \((n_t, n_t) \times (n_t, n_t)\). However, only the active elements are influenced during the update and by defining a local observable system (which corresponds to the active elements), one receives a system which is \( 4 \times 4 \) against \( (n_t, n_t) \times (n_t, n_t) \). That means that non-active elements do not affect the updating until they become active (zero values for the \( K \) related elements). Furthermore, \( P_k \) is a symmetrical and positive-semidefinite matrix, which allows further simplifications in the KF calculations.

4.5. Local observable system and steady-state approach, SSKF

The Local Observable System is defined and analyzed on a computationally efficient approximation for the KF, from Guardiola et al. (2013), is described. Following with the 2D look-up table application, at every iteration a maximum number of 4 elements can be updated and then, if no dynamics are accounted, the general \( n_t \times n_t \) system (11) can be written as \( 4 \times 4 \)

\[
x_k = I \times x_{k-1} + w_k
\]

\[
y_k = q_{l+k}^o x_k^o + v_k
\]

(14a)

(14b)

The simplest 2D map is a table of only 4 elements, and where the one existing area is always active, i.e. system (14) is exactly (11). For the general case, (14) must be rewritten at every \( k \) as \( k \) and matrices change. Supposing that system (3) is linear time-invariant (LTI): \( \sigma_{l+k}^2 \), \( \sigma_v^2 \), \( F \) and \( H \) are constant, then KF is steady-state (Simon, 2001). Of course system (14) is not time invariant, because \( q_{n_k}^o \) (and then \( H \)) depends on \( u_{l+k} \) and \( u_{l+k} \), which are the inputs. But, if inputs are considered stationary during a certain time, sequence \( \{K_i\}_{i=1}^{4} \) converges to a steady-state gain \( K^S \). For the 2D case

\[
K^S = \left[ \begin{array}{ccc}
   k_l(\eta_1, \sigma_{l+k}^2 / \sigma_v^2) & k_l(\eta_2, \sigma_{l+k}^2 / \sigma_v^2) \\
   k_l(\eta_1, \sigma_{l+k}^2 / \sigma_v^2) & k_l(\eta_2, \sigma_{l+k}^2 / \sigma_v^2) \\
   k_l(\eta_1, \sigma_{l+k}^2 / \sigma_v^2) & k_l(\eta_2, \sigma_{l+k}^2 / \sigma_v^2) \\
   k_l(\eta_1, \sigma_{l+k}^2 / \sigma_v^2) & k_l(\eta_2, \sigma_{l+k}^2 / \sigma_v^2)
\end{array} \right]
\]

(15)

where \( k_l \) is computed as follows:

\[
k_l(\eta, \sigma_{l+k}^2 / \sigma_v^2) = \frac{0.5(1-\eta)(1+s)}{0.5(1+s)(1-2\eta+2\eta^2) + \sigma_f^2 / \sigma_v^2}
\]

(16a)

\[
s = \sqrt{1 + \frac{4}{(1-2\eta+2\eta^2) \sigma_v^2}}
\]

(16b)

This approach allows mapping \( K \) and this only depends on inputs \( \eta_1 \) and \( \eta_2 \) calculated in (10) and noise trade-off \( \sigma_f^2 / \sigma_v^2 \). Nevertheless, as parameter aging is not considered, this makes the algorithm quite fast and light but robustness must be assessed in the applications. Algorithm pseudo-code is in Appendix B.

4.6. About system observability

The system (11) is not fully observable in one iteration. But local observability can be ensured if parameters or states in the local system (14) can converge with a given data set. As system is not LTI, the ordinary observability rank condition (Ogata, 2001) is not directly applicable. \( q_{n_k}^o \) depends on the input data, and it is evident that if a enough level of excitation is given, then the system could be observed, whatever the method chosen. This minimum level of excitation could be proved if 4 elements have been excited. A sufficient observability matrix for local observability may be built with the first 4 independent observations not necessarily consecutive in instants \( i_1, i_2, i_3 \) and \( i_4 \) of the elements of the involved area:

\[
O = \begin{bmatrix}
   q_{l+k}^o \\
   q_{l+k}^o \\
   q_{l+k}^o \\
   q_{l+k}^o
\end{bmatrix}
\]

(17)

and if the rank of this matrix is 4, then system (14) could be observable. But here the problem is linked with the noise tuning and
indeed full rank $O$ does not lead to system full convergence (full observability does not lead to full convergence as method and model structure affect). Majority of elements are also included in other neighbor areas and because of this the minimum condition of observability of a given state depends on the number of independent measurements that affect this state, and as other elements affect this, the condition stated in Höckerdal et al. (2011) gives a general condition.

4.7. The simplified Kalman filter SKF

Despite that only the active elements affect $K_k$, all $P_k$ elements are predicted and updated at every $k$. $P_k$ can be reorganized in such way that variances related with observable $P_k^o$ and unobservable $P_k^u$ elements are split.

$$P_k = \begin{bmatrix} P_k^o & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & P_k^u \end{bmatrix}$$

$(18)$

$K_k$, $q_k$ and $Q$ can also be split in the same way. Iterating one step ahead (7c) is written

$$P_{k+1} = \begin{bmatrix} 1 & 0 \\ 0 & 1_{4\times 4} - K_k^o q_k^T \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & P_k^o = Q^o \end{bmatrix} \begin{bmatrix} P_k^o + Q^o \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ (1_{4\times 4} - K_k^o q_k^T) (P_k^o + Q^o) \end{bmatrix}$$

$(19)$

where all observable matrices are $4 \times 4$, while the unobservable ones are $(n,n,-4) \times (n,n,-4)$. $(19)$ shows how only diagonal elements of $P_k^u$ are affected by adding the diagonal matrix $Q^o$. However, both diagonal and non-diagonal elements of $P_k^o$ are affected as in a $4 \times 4$ KF. Crossed relationships between observable and unobservable system, showed as dots in the matrix, are kept constant. This allows simplifying the complete system to the $4 \times 4$ active state-space system, whose resolution is highly computationally efficient.

Current work develops an approximation of the KF that requires both less memory and computations in the iterations and is named SKF. SKF builds and solves the local observable system $(14)$ at every $k$. Some simplifications are assumed: for the non-active elements, whose covariance matrix is $P_k^u$, non-diagonal elements are neglected building a vector $P^u = \text{diag}(P_k^u)$. This is justified for two reasons. First, when the system is running, then leaves one area and later returns to it, related variances reflect the time that the system has been out of this area. This is a desired property since the aging is captured by the variance increase due to $Q^o$ which indeed here is defined as a diagonal matrix. Older crossed correlations are maintained over the time and whether or not they are valid is not sure. Second, simulation results show that the full EKF and the SKF have similar performances when running on synthetic data and real data from the engine application. Then, at every iteration, the local system $(14)$ is solved with the only additional operation of $P_k^o = Q^o$ (converted to a vectorial calculation). SKF is a suboptimal filter but with a similar behavior than KF, as will be demonstrated in Sections 5 and Appendix A.

Pseudo-code for the SKF implementation is given in Algorithm 1. Note that code is to show the algorithm where some variables are not strictly needed in the computer programming, e.g. $P_k$ always that $P_k$ exists, but are kept for the sake of clarity. Here, noise variance $\sigma^2_{w_k}$ is the same for all elements, allowing to pre-define diagonal matrix $Q^o = \sigma^2_{w} 1_{4\times 4}$ for active system and vector $Q^u = \sigma^2_{w} \text{diag}(1_{n,n-4})$ for the non-active one.

Algorithm 1. Pseudocode for the simplified Kalman filter SKF.

```
input : \mathbf{x}_{k-1}, u_k, P_{k-1}, R, Q^o, Q^u
output: \hat{\mathbf{x}}_k, P_k
1 while Algorithm is running do
  2 if Active area changes then
  3 \hspace{1em} \mathbf{x}_{k-1}^o is stored in the correct positions of \mathbf{x}_{k-1}
  4 \hspace{1em} diag(P_{k-1}^o) and P_{k-1}^u are stored in a global variance vector $P_{k-1}$ according to their
  5 \hspace{1em} correct positions
  6 \hspace{1em} Redefinition of $P_{k-1}^o$, $P_{k-1}^u$, $\mathbf{x}_{k-1}^o$
  7 end
  8 Computation of $\eta_{1,k}$ and $\eta_{2,k}$ as in (10)
  9 Definition of $q_k^o(\eta_{1,k}, \eta_{2,k})$ as in (12b)
  10 $P_{k|k-1}^o = P_{k-1}^o + Q^o$
  11 $K_k^o = P_{k|k-1}^o q_k^T (q_k P_{k|k-1}^o q_k^T + R)^{-1}$
  12 $x_{k|k-1}^o = x_{k-1}^o + K_k^o (y_k - q_k^o x_{k-1}^o)$
  13 $P_{k|k-1}^o = (1_{4\times 4} - K_k^o q_k^T) (P_{k|k-1}^o + Q^o)$
  14 $P_k^u = P_{k|k-1} + Q^u$
  15 end
  16 Updating of $x_k$ and $P_k$ considering $x_{k|k-1}^o$, diag($P_{k|k-1}^o$) and $P_{k|k-1}^o$
```

In the following section the memory allocation and computational complexity of the three methods, full EKF (here denoted KF), the SKF developed here and described about, as well as the SSKF is described.
4.8. Analysis of memory and computations

A computation time study is made programming code in Matlab software on a laptop computer Intel Core 2DUO T9300 2.5 GHz with Windows VISTA 64 bits. Fig. 5 shows the relative computation time used by the three methods for updating 2D look-up tables of different sizes. Only square $n \times n$ tables have been considered with $n$ ranging from 2 to 21, since it is the total number of elements in vec($T$) that influences memory allocation and computations. The Y-axis shows the relative computation time with respect to the average time that the SSKF takes for performing 1000 iterations, i.e., a relative time of 1 is that method needs exactly the same time that SSKF. SKF needs around 1.15 times the SSKF calculation time, whatever the number of parameters. This slight difference is explained by the manipulation of variances and the need of redefining active vector and matrices $P^k$ and $x^k$ when active area changes (see lines 4 and 5 of Algorithm 1 and compare with Algorithm 2). However, required time for KF rapidly grows when number of parameters increases, i.e. for 256 parameters ($16 \times 16$), KF needs around 7.8 times the one for SSKF. An exponential function is fitted by least squares method to the KF time as function of number of parameters $n_p$, whose result is the relative time KF with respect to the one of SSKF.

Table 2 gives some numbers when comparing required memory resources and computation times for the algorithms. SKF permits reducing the requirements in terms of memory resources and the general system $(n_i,n_o) \times (n_i,n_o)$ is reduced to a simple $4 \times 4$. These results show the benefits in memory and computational when using the SSKF and SKF compared to the full KF.

5. Simulation of the updating algorithms

With these promising results the convergence and robustness of approximations need also to be studied and this is first performed using simulations. The objective of this section is comparing the abilities and performance of the algorithms against different synthetic cases where no dynamics are accounted (real system and model are static); later, Section 6 will prove the algorithms under real conditions in a diesel engine considering dynamics (sensor output will be used there as reference) and completing the study on methods capabilities.

For automotive applications, engine speed $n$ and fuel mass injection quantity $m_f$ usually define the engine operating point and the look-up table scheduling points selected in the simulation are inspired by these quantities. Hence $r$ represents a grid for $u_1(k)$ (speed in rpm) and $c$ for $u_2(k)$ (injected mass fuel in mg/str):

$$r = \{500, 1500, 2500, 3500\}$$

and the true map $T_r$ is defined with the surface

$$T_r = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

where $x_i = \text{vec}(T_r)$ is the expanded state vector objective of the estimation process. No initial process knowledge is considered: $x_0$, $x_0^c$, $P_0$ are initialized with zeros. $\sigma^2_v = \sigma^2_w = 1$ for all cases, unless it should be pointed.

System measurements are given and algorithms performance is evaluated. As synthetic signals are used, this allows defining error metrics. Simulations include two worst-cases and a favorable one:

- Input with random variation: a random shot of values following an uniform probability distribution. This is an ideal situation for learning as all areas are excited and parameters aging is low (excitation is high). Measurements are perfect and constant $y_k = 1 \forall k$.
- Linear variation: varying $u_2$ keeping constant $u_1$. This variation tests a degenerated case where observability is critical along the $u_1$-dimension. Measurements are perfect and constant $y_k = 1 \forall k$.
- Measurement noise rejection: studying the effects of noise transmission between output and observation when measurement vector is noisy.

5.1. Simulation 1: input with random variation

Sequence $\{u_1(k)\}^{1400}_{k=1}$ following an uniform distribution over grid defined in (20) is used for exciting the system. A complete identification is possible as grid is completely covered. This situation is not so far of the reality for diesel engines, e.g. urban cycles with a lot of speeding/braking actions cause quasi-random variations in partial-low load areas of the engine map (see Section 6 for seeing how dynamics are considered). The selected grid and covered points are shown in Fig. 6.

Here, variance information is not as relevant as in other cases, because of the stochastic nature of inputs. The mean value of all states is plotted for at each time step in Fig. 7. All three methods perform well and rate of convergence can be tuned varying $\sigma^2_v$ and $\sigma^2_w$.

A quick view on the effect of the hypothesis for the SKF is shown for one element $x_6$ in top plot of Fig. 8, which shows variance $P_{66}$ of $x_6$ for KF and SKF (not applicable for SSKF). This value indicates the observability of $x_6$: when it increases monotonically, $x_6$ is not observable. In some parts, an offset between $P_{66}$ for KF and SKF appears because of the neglected covariances in SKF, but in the end, this offset is absorbed and the two methods behave similar. The bottom plot of Fig. 8 shows $x_6$ evolution for the
three methods. $x_6$ is updated when the element becomes observable. Note that $x_6$ is already active in the first iteration and $K_1$ is non-null. Then, SKF updates $x_6$ in $k=1$ while $K_1$ for KF and SKF is null. Anyway, this is not an advantage of this method: KF and SKF can behave similar if $P_0$ is non-null, but here for the simulations $P_0 = [0]$. When there exists no previous knowledge of the system to be learnt, it should advisable to initialize $P_0$ with a certain value for speeding up updating during first iterations.

5.2. Simulation 2: input with linear variation

$T_i$ is identified with a sequence $(u_{i,k})^{200}$ where $u_i$ varies monotonically from 0 to 60 (from $k=0$ to 100) and coming back from 60 to 0 (from $k=101$ to 200), while $u_i=2000 \forall k$. This condition is highly restrictive as the excitation level is not high: at each $k$ only small variations of input $u_2$ are applied and due to this full convergence condition is not fulfilled in the first running of one area, and interactions between areas and the way back are necessary to ensure the convergence of both KF and SKF. Three matrix areas are excited (see (22)): area 1: elements 5,6,9,10; area 2: 6,7,10,11; area 3: 7,8,11,12. Moreover, due to inputs nature, row 5–8 evolution is equivalent to 9–12.

$$T_k = \begin{bmatrix} x_{5,k} & x_{6,k} & x_{7,k} & x_{8,k} \\ x_{9,k} & x_{10,k} & x_{11,k} & x_{12,k} \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

(22)

Fig. 9 shows the evolution of 4 states $x_5$ to $x_8$. First thing to note is the similar performance of KF and SKF, and the instabilities of SSKF. For instance, $x_5$ is in the area 1, and during the first run it is not capable of converging, but in one round KF and SKF methods converge, and SSKF seems to do it, at least in $k=200$. The other states have slightly different performances, because the first area has been already been covered; e.g. $x_6$ is closer to the convergence as being a member of area 1 and area 2, but until the way round it does not get the full convergence for KF and SKF methods. Similar behaviors are observed for $x_7$ with these methods. In addition, $x_8$ as member of the area 3, has the advantage that areas 1 and 2 have been covered first, and the information is already available for getting the full convergence for KF and SKF in $k=100$. The full observability (in the sense of convergence) of these 8 elements is reached only when $O$ in (17) accounts for the three areas when using KF and SKF. Nevertheless, SSKF estimation does not converge for $x_5$, $x_7$, and $x_8$ due to no state error information is tracked, and the system and convergence becomes more dependent on data.
In order to understand how KF and SKF behave is informative to study also the variance and covariance evolution (remember that this is not present for the steady state filter, SSKF). Fig. 10 shows the variance evolution for row elements, while Fig. 11 shows the evolution of a few covariances between states (corresponding to non-diagonal elements of $P$). As for the random case in 8, variance evolution for KF and SKF is similar, although some deviation exists in elements 6 and 7. This is explained for covariances or non diagonal $P_{ij}$ elements that SKF only accounted for active areas. This is shown in Fig. 11 where some crossed covariances are plotted. When the pair of elements are not active, covariances of these are not tracked, equivalent to reset them to zero in the global $P$. This is a particular difference with KF, which always tracks covariance, although area would not be active, increasing computational requirements without a justified improvement in the estimation. However, when an area is active, the SKF covariance evolves as the KF and in the limit is equivalent to the KF. The difference is larger initially but after the observations have converged the difference is negligible, as shown in Fig. 9. This pattern may be slightly modified linking the covariance values to elements pairs and not to the involved areas (excepting boundaries, all pair of elements are shared in two different areas), although final results are similar.

5.3. Simulation 3: measurement noise rejection

A uniform distributed noise with zero mean and maximum amplitude of 0.2 is applied to the measurement given in simulation 2; $y_k$ is shown in Fig. 12. A robust method must filter this noise and converges to the true values.

The resulting estimations are shown in Fig. 13, and the evolutions are quite similar to the ones of Fig. 9 although with a slight noise transmission. KF and SKF performs well, being able to also filter the noise, but SSKF is not capable of converging again having a similar response as in simulation 2. Variance and covariances are exactly the same of Figs. 10 and 11 as state-space system and input sequence does not change.

5.4. Conclusions from simulations

SKF method is demonstrated to have similar accuracy as KF, at least for the numerical cases proved in this section. Furthermore, KF and SKF solutions converge for all cases, including the simulations 2 and 3, which are specifically restrictive because of the low
level of excitation. KF, as shown in Fig. 5 and Table 2, is computationally heavy. SSKF is the fastest and lightest method as variance is not tracked and KSS formulation is derived analytically. SKF only needs around 1.15 times the SSKF time for calculation, and the memory resources required are similar, except for the track of variances. SSKF behaves well in situations where variance tracking is not critical, e.g. simulation 1, but is less robust when data is structured and the level of excitation is low, e.g. simulations 2 and 3. KF and SKF are also capable of filtering the noise, a minimum condition for a correct updating, and SSKF also filters the noise, but its stability depends on data. In comparison, the SKF method, derived from the KF, is demonstrated to be an efficient and accurate method for updating look-up tables, at least for the simulation conditions presented. In the next, methods are tested in a real application.

6. Online NOx estimation

The attention is turned to the problem of online NOx estimation. The approach consists of modeling NOx using an adaptive model based on a static 2D look-up table whose parameters are estimated with the updating methods. The state-space model (11) is slightly modified for coping with sensor dynamics and inputs u are treated as well.

Model parameters are estimated online while the engine is running, or re-calibrated if system ages once that an initial calibration is given without any special calibration procedure or test rig, beyond the sensor measurements. The first time that the engine is running, parameters evolve, and when the engine switches off, the stored parameters can be used for predicting engine is running, or re-calibrated if system ages once that an initial calibration is given without any special calibration procedure or test rig, beyond the sensor measurements. The numerical values used are

\[ y_M(k) = \frac{1-a}{1-a z^{-1}} T_{NOx}(k-t) \]  

and the global system (11) must be augmented with one extra-dimension (if a first order filter is considered for sensor dynamics)

\[ x_k^w = \left[ \frac{l_{sa} a}{(1-\theta) l_0} \right] x_k^{-1} + w_k \]  

with

\[ y_k^w = H^w x_k^w + v_k \]

where \( H^w = [0 \cdots 1] \)

being \( x_k^w \in \mathbb{R}^{n_x+1} \) and \( y_k^w \in \mathbb{R} \). Index \( w \) stands for wide.

The local observable system (14) is

\[ x_k^w = F_k x_{k-1}^w + w_k \]

\[ y_k^w = H^w x_k^w + v_k \]

\[ F_k = \begin{bmatrix} I_n & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \]

\[ H^w = [0 \ 0 \ 0 \ 1] \]

\[ x_k^w = \begin{bmatrix} x_k^1 \\ x_k^2 \\ y_k \\ y_k \end{bmatrix} \]

where \( H^w \) is constant, because \( q_k \) and filtering parameter \( a \) are included now in the time varying process matrix \( F_k \). \( Q^w \) process noise matrix is also augmented and includes an extra-noise term \( \sigma^2_f \) for the new state \( y_M \)

\[ Q^w = \begin{bmatrix} Q & 0 \\ 0 & \sigma^2_f \end{bmatrix} \]

Even though the new local system has one more dimension \((5 \times 5)\), SKF hypothesis are still applicable as proved in Appendix C.

6.6. Comparison of the updating algorithms in the SDMP

The SDMP cycle is used for comparing the algorithms with real engine data. \( T_{NOx} \) is the null matrix and is updated with the cycle. Filter is calibrated by trial-and-error (see Payri et al., 2012 for a calibration method based in a Monte Carlo approach for the KF). The numerical values used are

\[ \sigma^2_i = 25^2; \quad \sigma^2_m = 15^2; \quad \sigma_f = 50^2 \]

\[ a = 0.96; \quad \tau = 0.75 \text{ s} \]

\[ c = [0 : 5 : 30 \ 34 : 4 : 50 \ 55 : 5 : 80] \]

\[ r = [750 \ 780 \ 1000 \ 1250 \ 1500 \ 1750 \ 2000 : 500 \ 4500] \]

while the sampling frequency is 50 Hz. Measurement \( y_{NOx} \) is given by the NOx sensor. A test subset of \( t = 400 \text{ s} \) is used for updating the adaptive map and then dynamic model with table \( T_{NOx,400} \) is used for predicting NOx in the whole cycle. KF, SKF and SSKF methods are used for updating the model. Results can be seen in Fig. 14. All three methods behave well and are capable of predicting NOx emissions. Dynamics assumptions are also good enough.
for having a good fitting in the SDMP. Note that indeed SSKF results are also acceptable as covariance tracking is not relevant in this cycle as in slow varying tests. These conditions are similar to the ones exposed in Section 5.1. Fitting could be optimized tuning independently the filter for each method.

It is worth comparing the convergence of the methods and for this, sample subsets of the SDMP are used in a sequence for testing the absolute mean error when the updating time is varied monotonically around the cycle

\[ e_m(k) = \frac{\sum_{i=1}^{n_0} (y_{NO}^i(t) - y_{NO}(t)_{t = k4T})}{k4T} \]  

(35)

where \( \Delta T \) is the sample time and \( t \) is the time that table has been updated. Results are shown in Fig. 15. Focusing the attention in the lines generated with the filter calibration Cal (34), \( e_m \) tends to be lower as \( t \) grows at least during first iterations when system knowledge is poor. The horizontal line shows the benchmark model error, which is constant as model calibration is fixed. It is clear how all three methods, including SSKF, converge to an error similar to that of the benchmark model, but with the advantage that drift and aging is accounted by the online versions but not by the model. KF and SKF behavior is quite similar as expected. On the other hand, Fig. 15 also shows that KF and SKF are faster than SSKF and this is due to the filter tuning. Anyway, there exists a trade-off between convergence speed and estimation robustness, e.g. KF and SKF have a significant oscillation with respect to SSKF around the benchmark model line. This is a matter of the noise tuning and bottom plot is built with a lower gain observer with the calibration set

\[ \sigma_f^2 = 25^2; \quad \sigma_w^2 = 7^2; \quad \sigma_f^2 = 50^2 \]

(36)

for KF and SKF, keeping the ones of SSKF. Results are shown in lines generated by Cal (36) in Fig. 15. Here convergence speed is lower but noise transmission and overfitting is avoided when \( t \) is large. This goes in the direction of robustness and the optimized calibration data set must solve this trade-off, considering uncertainties in the sensor behavior knowledge or in the model quality and data-set quality, among others.

With respect to the global observability, variances of 4 table parameters are compared for both KF and SKF using calibration (36). Fig. 16 shows the results. Top plot shows an element that is never observed during first 200 s and due to this, variance increases monotonically and \( x_{126} \) is still null in \( t = 200 \). Second and third elements represent two elements that are active for some instants. This is clearly seen when variance decreases, while when elements are not active, variance increases again monotonically. Note that in a production system application, that will run for the life time of a vehicle, the elements of the estimation error covariance matrix need to be limited so that they do not grow too much and cause numerical problems, see e.g. Höckerdal et al. (2011) that proposes a saturation for avoiding too large variances. The bottom plot shows the variance of \( y_{133} \) that is fairly constant and is a proof of the system global observability.

KF and SKF lines are pretty similar and differences can be found only when zooming in the figure. This is another proof of that SKF behaves quite similar to KF but with a much lower computational burden involved. SSKF behavior is not bad and prove that the method can be useful when calibration data set fulfills some conditions in order to ensure robustness. Anyway, for the nature of the application here, SKF is the best solution for online updating of maps.

### 6.3. Online estimation

The state-space model for learning maps is useful not only for calibrating the map but also for real time observation of NO\(_x\) signal. Fig. 17 shows an interesting result in line with this discussion for the SDMP and keeping calibration (36). \( \hat{y}_{SKF} \) represents the observation of

![Fig. 14. NO\(_x\) prediction for the three updating methods when maps are updated during 400 s, i.e. a subset of 20,000 measurements of \( y_{NO} \), and then used for the prediction. Sensor measurement \( y_{NO} \) is provided for comparison.](image)

![Fig. 15. Total mean absolute error \( e_m \) for the different methods and different ranges used for identification. SSKF dashed line is calculated used filter calibration (34). Lines left to the SSKF one (dashed): Using calibration (34). During first iterations, error is high, as far as table has no initial knowledge, but once that time is around 400 s, error is nearly minimum, although noise transmission is evident. Lines right to the SSKF one: Using calibration (36), noise transmission is reduced but convergence speed as well. Final filter tuning must be a trade-off between robustness and convergence speed.](image)

![Fig. 16. Variance analysis in the SDMP updating. From top to bottom, variances of elements 126, 126, 133 and 217 are plotted. Black line is KF and gray line is SKF.](image)
drift. Indeed, $NOx_0$ could be used for initializing $TNO$, but here the initial map is the null matrix, which is a worst-case condition. NEDC cycle is running and map is updated. Calibration (36) is used. Fig. 18 shows results comparing $y_{NOx}$, $y_m$ (which exhibits drift) and $y_{SKF}$, which here represents $NOx$ prediction by using the updated map after 1 run of the cycle. Adaptive map gets a good $NOx$ estimation which is slightly better than $NOx$ benchmark model one.

SKF could be used for online adaptation and/or calibration of complex models, where a number of maps and parameters must be updated. Anyway, a deep study is required for ensuring observability, convergence and robustness properties and for getting a computationally efficient learning structure. A big number of parameters and maps should be updated and problem is non-convex. Of course that computational issue is critical and there SKF can be an effective solution. Here, SKF has been used for updating single maps with succeed.

7. Conclusions

$NOx$ emissions are estimated by using static maps function of a diesel engine operating conditions and a delayed low-pass filter for including dynamics. The solution is not new and several other authors have published about this, but here an online adaptive algorithm is presented for maps calibration and updating. Kalman filtering is used for updating because of its capability for tracking system and parameters aging. Computational issues involved when of updating look-up tables online with the Kalman filter based methods are then addressed. A simplified version of the Kalman filter SKF with similar accuracy as the standard KF, but that requires a much lower memory resources and calculation time, is developed. The local observable system serves as inspiration for analyzing the variance matrix performance and is used for comparison. SKF is based on the KF but neglects covariances of locally unobservable states. The methods results under simulation and real data remark two key points that make that the SKF is a suitable option for online estimation:

- Calculation and memory resources: The KF is a heavy method with similar accuracy than SKF, and the required computational time is rapidly growing when the system complexity increases, as well as the required memory. SSKF is the fastest method and it achieves good results with random-like data and with calculation times similar to the ones of SKF.
- Robustness: KF and SKF perform quite well and do not have robustness problems, while the SSKF exhibited oscillatory behavior when input data are structured.

The methods and especially the SKF are validated using a real world nonlinear model and sportive driving mountain profile SDMP for $NOx$ estimation, being an application of a high interest. The results are compared with the ones of a gray box $NOx$ model, obtaining similar results, which also shows the calibration capabilities of the method. Another contribution is considering sensor dynamics in the original state-space system. The obtained results as well as the simplicity of the formulation could be the key for implementing it in HIW structures and opens the possibility of its use in commercial ECUs. Future work must be in using adaptive strategies for calibrating full engine models and cope with sensor non-linearities, especially the delay.

Appendix A. Submodel equations for the $NOx$ model

For estimating EGR rate $\dot{e}_{EGR}$, a mean value EGR flow model with sensor signals of boost pressure $p_b$ and mass air flow $m_a$ as inputs is used. Intake manifold mass flow $m_i$ is obtained from
volumetric efficiency \( \eta_v \) (mapped with the steady-state tests), assuming constant intake temperature \( T_i \) (the error is pretty low compared with other error sources) and using \( p_2 \) signal

\[
Q_d = \frac{V_d n_t}{2 	imes 60} \text{ (m}^3/\text{s)}
\]

\[
\rho_i = \frac{p_2}{RT_i} \text{ (kg/m}^3\text{)}
\]

\[
m_i = Q_d \rho_i \text{ (kg/s)}
\]

where \( Q_d \) is the volume flow at intake and \( \rho_i \) is the density at intake and all variables are introduced in the international standard system.

Considering mass accumulation effects in the node of air-\( EGR \)-intake, and \( m_w \), \( r_{EGR} \) can be calculated

\[
r_{EGR} = 1 - \frac{m_a}{m_i}
\]

(A.2)

For model calibration, measurements of oxygen at intake and exhaust from the gas analyzer are used

\[
r_{EGR} = \frac{Q_{Exhaust}}{Q_{Intake}} - 0.209
\]

(A.3)

If these measurements would be available online, an observer could be built for having a more reliable estimation. Furthermore, there exist fast \( EGR \) sensors prototypes but because of temperature and pressure limitations, measurements are not reliable; there an observer could be proposed using \( EGR \) model presented here as input and sensor as output.

Actual relative fuel-to-air ratio \( F_r \) can be observed by using the model

\[
F_r = 14.5 \frac{m_i}{m_a}
\]

(A.4)

and using the oxygen output of the \( ZrO_2 \) sensor for drift correction. Oxygen measurement is steady-state reliable but presents filtering and delay; \( F_r \) model is calculated from \( m_a \) and \( m_i \), which are fast signals but present drift. See Payri et al. (2012) for a complete procedure for estimating \( F_r \).

Appendix B. Algorithm for SSKF

Algorithm 2. Pseudocode for the steady-state Kalman filter for updating look-up tables SSKF

input : \( \hat{x}_{k-1}, m_k \)
output: \( \hat{x}_k \)

1 while Algorithm is running do
2 \hspace{1em} if Active area changes then
3 \hspace{2em} \( x^v_{k-1} \) is stored in the correct positions of \( x_{k-1} \)
4 \hspace{2em} Redefinition of new \( x^v_{k-1} \):
5 end
6 Computation of \( \eta_1,k \) and \( \bar{\eta}_1,k \) as in (10)
7 Definition of \( q_k(\eta_{1,k}, \bar{\eta}_{1,k}) \) as in (12b)
8 Computing \( K^SS \) as in (5)
9 \( x^v_k = x^v_{k-1} + K^SS (y_k - q_k^v x^v_{k-1}) \)
10 end
11 Updating of \( x_k \) considering \( x^v_k \)

Appendix C. Variance tracking for the dynamic system

The system (29) is the dynamic local observable system. For this system, (18) is still valid. The only modification is that the new observable part \( P_{\omega}^{w} \) has \( n \) extra dimensions corresponding to the \( n \):th order discrete filter. In the paper, a first order discrete filter is used, and then \( P_{\omega}^{w} \) is 5 × 5.

\[
P^w_k = \begin{bmatrix} P_{\omega}^{w}_k & P_{\omega}^{w_0} \\ P_{\omega}^{w_0 T} & P_{\omega}^{w_0} \end{bmatrix}
\]

(C.1)

\[
P_{\omega}^{w_0} = \begin{bmatrix} P_{\omega}^{w_0} \\ \rho_{\omega}^{w_0} \end{bmatrix}
\]

(C.2)

where \( P_{\omega}^{w_0} \) is the scalar variance coupled to sensor dynamics and \( \rho_{\omega}^{w_0} \) is the vector 1 × 4 with the covariances between sensor estimation and observable parameters. \( P_{\omega}^{w_0} \) is a positive-semidefinite matrix, so is \( P_{\omega}^{w} \) and \( K_k \) is reordered to contain the unobservable part \( K_k^1 \) = 0 and the observable part \( K_k^2 \) and the latter in the parameters related \( K_k^2 \) and dynamics related part \( K_k^3 \)

\[
K_k^w = \begin{bmatrix} K_k^2 \\ K_k^3 \end{bmatrix}
\]

(C.3)

and (19) is now

\[
P_{k+1} = \begin{bmatrix} P^u_k + Q^u & P^w_k \\ \bar{P}^u_k P^w_k \left( P^w_k \bar{P}^u_k + Q^w \right) \end{bmatrix}
\]

(C.4)

\[
P_{k+1} = \begin{bmatrix} P^u_k + Q^u \\ \bar{P}^u_k P^w_k \left( P^w_k \bar{P}^u_k + Q^w \right) \end{bmatrix}
\]

(C.5)

which shows that SKF partitioning is also applicable for the augmented dynamic system. \( P_k \) is a positive-semidefinite matrix so is \( P^u_k \) and \( P^w_k \).

References


