Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

State-Space Kernelized Closed-Loop Identification of Nonlinear Systems

M.F. Shakib * **R. Tóth** ** **A.Y. Pogromsky** * **A. Pavlov** ***
N. van de Wouw *,****

* *Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: {m.f.shakib; a.pogromsky; n.v.d.wouw}@tue.nl).*
** *Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: r.toth@tue.nl).*
*** *Department of Geoscience and Petroleum, NTNU, Trondheim, Norway (e-mail: alexey.pavlov@ntnu.no).*
**** *Department of Civil, Environmental & Geo- Engineering, University of Minnesota, Minneapolis, U.S.A. (e-mail: nvandewo@umn.edu).*

Abstract: In this paper, we propose a non-parametric state-space identification approach for open-loop and closed-loop discrete-time nonlinear systems with multiple inputs and multiple outputs. Employing a least squares support vector machine (LS-SVM) approach in a reproducing kernel Hilbert space framework, a nonlinear auto-regressive model with exogenous terms is identified to provide a non-parametric estimate of the innovation noise sequence. Subsequently, this estimate is used to obtain a compatible non-parametric estimate of the state sequence in an unknown basis using kernel canonical correlation analysis. Finally, the estimate of the state sequence is used together with the estimated innovation noise sequence to find a non-parametric state-space model, again using a LS-SVM approach. The performance of the approach is analyzed in a simulation study with a nonlinear system operating both in open loop and closed loop. The identification approach can be viewed as a nonlinear counterpart of consistent subspace identification techniques for linear time-invariant systems operating in closed loop.

Keywords: Nonlinear State-Space Identification, NARX modeling, Kernel Canonical Correlation Analysis, LS-SVM.

1. INTRODUCTION

Identification of nonlinear systems is a challenging and active field of research (Schoukens and Tiels (2017); Chiuso and Pillonetto (2019)). A generic model class is the class of nonlinear state-space models. This model class is attractive as it is particularly suitable for parsimonious representation of multiple input multiple output (MIMO) systems. Furthermore, many analysis and controller design tools exist for this class of models, see Khalil (1996).

A discrete-time nonlinear state-space (NL-SS) model is characterized by its state-transition map and its output map. As there is a recursion loop in the evolution of the hidden state variable, identification of these mappings is a challenging task (Marconato et al. (2013)). Many NL-SS identification methods exist that are based on direct identification of these mappings using specific parametric model structures. However, the identification corresponds to a computationally-demanding nonlinear optimization problem with the need of efficient initialization and a model parametrization to-be-provided by the user, see Giri and Bai (2010); Schoukens and Tiels (2017). Alternatively, non-parametric identification techniques exist for the identification of nonlinear input-output models, see Pillonetto et al. (2011). However, for control, NL-SS models are favored over nonlinear input-output models as many design techniques use Lyapunov's second method, which requires a state-space representation of the system. Unfortunately, state-space realization of nonlinear input-output models is a difficult task with many unsolved problems, see Kotta et al. (2015).

If the state sequence of the underlying NL-SS model would be available, for example, by means of additional measurements, the identification of the state-space mappings would become a static problem, which simplifies the identification problem significantly. However, the state sequence is often unavailable, whilst obtaining an accurate estimate of it is non-trivial. For Linear Time-Invariant (LTI) systems, subspace techniques, see Larimore (1990) and Van Overschee and De Moor (2012), can be employed for the estimation of the state sequence. For nonlinear systems, to this extent, Marconato et al. (2013) proposed a method to approximate the state sequence by, in the first step, identifying a linear model, and, in the next step, minimizing a linear cost function yielding an approximation of the state sequence. Alternatively, Verdult et al. (2004) proposed a method based on least-squares support vector machine (LS-SVM), which solves an intersection problem between future and past input/output data using a kernelized version of Canonical Correlation Analysis (CCA), to obtain a state sequence.

The method of Verdult et al. (2004) estimates the state sequence of the underlying NL-SS model. However, in that work, the effect of noise, being inevitably present in practice, was not taken into account, which can result in biased identification results. To this extent, we relate to the consistency property of an estimator, which, loosely speaking, ensures that the true model is recovered if the

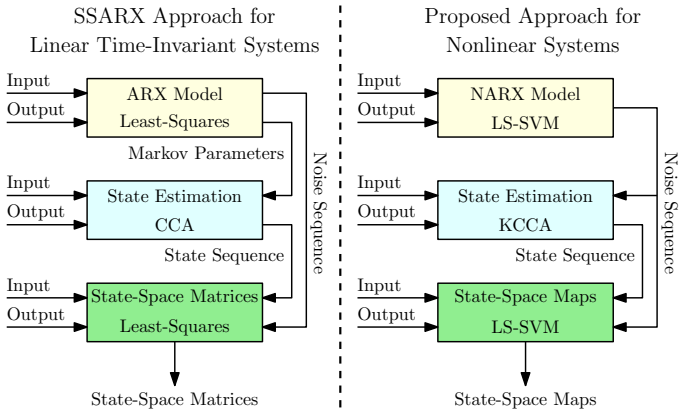

Fig. 1. Existing SSARX approach for LTI system (left). Proposed identification approach for nonlinear systems (right).

number of data points tends to infinity (Ljung (1987)). Consistent subspace techniques for LTI systems operating in closed loop commonly rely on the identification of a consistent noise model first using an auto-regressive model with exogenous inputs (ARX), see Van der Veen et al. (2013). Then, in the next steps, the identified noise model is used in the consistent estimation of the state sequence and the state-space matrices (up to a similarity transformation). In particular, we highlight the SSARX technique, see Jansson (2003), which uses a three-step approach. In the first step, a consistent ARX model is estimated to obtain a one-step-ahead prediction model of the output, which is an aggregated form of the system and its noise model. In the second step, the state sequence is obtained by performing CCA to infer the state sequence as an intersection of future and past input/output data. The third step entails the estimation of the state-space matrices, which is formulated as a problem that is linear in the parameters and solved using linear least squares.

In this paper, we extend the methodology of SSARX for the identification of LTI systems to the case of nonlinear systems. In the first step, rather than identifying an ARX model, a nonlinear ARX (NARX) predictor model is identified by an LS-SVM approach, which is proven to be consistent, see De Nicolao and Trecate (1999). The prediction error of this NARX model serves as an estimate of the innovation noise sequence, similar as in Mercère et al. (2016) for the LTI case. Next, in the second step, we use the estimated innovation noise sequence as an additional pseudo-input to estimate the state sequence using the kernelized CCA method presented in Verdult et al. (2004). After that, having also an estimate of the state sequence at hand (in an unknown state basis), in the third and final step, we identify the state-transition map and the output map of the NL-SS model non-parametrically, again using a LS-SVM approach. The resulting model is characterized by the state-transition map and output map, both given non-parametrically. An overview of the proposed approach is given in Fig. 1.

Although, we do not give an overall consistency proof, each step of the proposed procedure corresponds to a consistent estimate under the assumption that the true noise and state sequence are provided in the previous steps. Simulation studies, presented in Section 5, show that the proposed identification strategy, both in the open-loop and closed-loop case, outperforms the case where noise is not taken into account in the identification process, i.e., direct application of the method of Verdult et al. (2004). This simulation result clearly demonstrates that using the estimated innovation noise sequence obtained by the identified NARX model, significantly improves the quality of the identified model.

The remainder of this paper is organized as follows. In Section 2, the identification problem is formally introduced. The concept of function approximation, used in the steps of the identification strategy, is described in Section 3. The overall identification approach is presented in Section 4. Simulation examples are given in Section 5. Section 6 presents the conclusions of this paper.

2. PROBLEM FORMULATION & NOTATION

We consider MIMO discrete-time nonlinear systems that can be represented by the following set of first-order difference, i.e., state-space, equations:

$$x_{k+1} = f(x_k, u_k, e_k), \tag{1a}$$
$$y_k = h(x_k) + e_k, \tag{1b}$$

where, at time instance k, the state is denoted by $x_k \in \mathbb{R}^n$, the input by $u_k \in \mathbb{R}^m$ and the output by $y_k \in \mathbb{R}^l$. The functions f and h are called the state-transition map and output map, respectively. The innovation noise sequence $e_k \in \mathbb{R}^l$, is assumed to be drawn from a zero mean normal distribution with finite diagonal covariance matrix Σ_e. The problem we consider is to identify the state dimension n, to identify the functions f and h non-parametrically and to give an estimate of the noise variance Σ_e, based on a data-set $\mathcal{D} = \{u_k, y_k\}_{k=1}^N$ generated by (1), where N is the number of data points. Conditions imposed on the mappings f and h are given in Section 4, where the proposed identification method is presented.

As many systems are unstable by themselves, or only a part of the dynamics of a more complex system are to be identified, we also consider systems that operate in closed loop. However, to avoid the existence of an algebraic loop, i.e., the output not being uniquely determinable, we require the assumption that either the plant or the controller has no feedthrough (or both). Such an assumption is also commonly adopted in the LTI case, see Van der Veen et al. (2013). Here, for the sake of notational simplicity, we assume that the plant has no direct feedthrough, i.e., the mapping h is not a function of u_k in (1b). This implies that the covariance matrix $\mathbb{E}\left\{u_k e_j^\top\right\}$, where $\mathbb{E}$ denotes the expectation w.r.t. the random variables u_k and e_k, is a zero-matrix if $j > k$, but can be non-zero for $j \leq k$ in the closed-loop setting, which typically results in a bias if the noise is not handled appropriately during identification.

In the sequel, the following notation for vectors of shifted sequences of inputs is used: $\bar{u}_k^d := \begin{bmatrix} u_k^\top & \dots & u_{k+d-1}^\top \end{bmatrix}^\top$. Similarly, vectors of shifted outputs and shifted noise are denoted by $\bar{y}_k^d$ and $\bar{e}_k^d$, respectively. Furthermore, iterative evaluations of mapping f w.r.t. x_k are denoted as follows:

$$\begin{aligned} f_k(x_k) &:= f(x_k, u_k, e_k) = x_{k+1} \\ f^d(x_k, \bar{u}_k^d, \bar{e}_k^d) &:= f_{k+d-1} \circ \ldots \circ f_{k+1} \circ f_k(x_k). \end{aligned} \tag{2}$$

Finally, the vector of sequential outputs is defined as:

$$\bar{y}_k^d = h^d(x_k, \bar{u}_k^{d-1}, \bar{e}_k^d) := \begin{bmatrix} h(x_k) + e_k \\ h \circ f^1(x_k, \bar{u}_k^1, \bar{e}_k^1) + e_{k+1} \\ \vdots \\ h \circ f^{d-1}(x_k, \bar{u}_k^{d-1}, \bar{e}_k^{d-1}) + e_{k+d-1} \end{bmatrix}.$$

3. FUNCTION ESTIMATION USING LS-SVM

In this section, we describe an approach to non-parametric function estimation. Such function estimation is used in Section 4.1 to identify a NARX model and in Section 4.3 to identify the mappings f and h of system (1).

The core concept of function estimation is to search for a function inside a Hilbert space $\mathcal{H}$, which is equipped with an inner product $\langle \cdot, \cdot \rangle$ and is complete with respect to the induced norm $\|g\|_{\mathcal{H}} := \sqrt{\langle g, g \rangle}$. Given the data-set $\mathcal{D} = \{z_k, w_k\}_{i=1}^N$, where $w \in \mathbb{R}$ is the set of observed outputs and $z \in \mathcal{Z} \subseteq \mathbb{R}^{n_z}$ the set of (observed) inputs, the search goal is to find the function g that minimizes the cost functional

$$\min_{g \in \mathcal{H}} \left[\frac{\gamma}{2N} \sum_{i=1}^{N} (w_k - g(z_k))^2 + \frac{1}{2} \|g\|_{\mathcal{H}}^2 \right]. \tag{3}$$

In (3), the first term penalizes mismatch in data-fit and the second term penalizes function complexity measured in a Hilbert space norm, which acts as regularization with positive regularization parameter γ. This (continuous) regularization parameter can be viewed as the counterpart of the (discrete) tunable parameter that defines, e.g., the model order in classical parametric system identification approaches. As the data-set $\mathcal{D}$ only contains N samples, the optimization problem (3) is only well-posed if the search is restricted to the reproducing kernel Hilbert space (RKHS) over $\mathcal{Z}$. This is a space of functions $g : \mathcal{Z} \to \mathbb{R}$ that satisfy the following boundedness criterion:

$$\forall z \in \mathcal{Z}, \quad \exists 0 \leq c < \infty : |g(z)| \leq c \|g\|_{\mathcal{H}}, \quad \forall g \in \mathcal{H}. \tag{4}$$

The theorem of Moore-Aronszajn, see Aronszajn (1950), ensures a one-to-one correspondence between RKHS of functions $\mathcal{H}$ over $\mathcal{Z}$ and symmetric positive semidefinite [1] kernel functions $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. It ensures that for every RKHS $\mathcal{H}$ satisfying (4), a unique symmetric reproducing kernel function $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ exists which is positive semidefinite and obeys the reproducing property $g(z) = \langle g(\cdot), K(\cdot, z) \rangle$. The optimization problem in (3) has a closed-form solution by means of the Representer theorem, see Kimeldorf and Wahba (1971). In particular, for the RKHS $\mathcal{H}$, the minimizer $\hat{g}$ of (3) is given by

$$\hat{g}(\cdot) = \sum_{i=1}^{N} \hat{\alpha}_i K_{z_i}(\cdot) \tag{5}$$

with $K_{z_i}(\cdot) := K(\cdot, z_i)$ and $\hat{\alpha} = [\hat{\alpha}_1 \ \dots \ \hat{\alpha}_N] \in \mathbb{R}^N$ being given by

$$\hat{\alpha} = \left(\frac{1}{N} \mathcal{K}_{zz} + \gamma^{-1} I_N \right)^{-1} \frac{1}{N} W,$$

where $W = [w_1 \ \dots \ w_N]^\top$, $I_N \in \mathbb{R}^{N \times N}$ is an identity matrix and $\mathcal{K}_{zz} \in \mathbb{R}^{N \times N}$ is the Gram matrix defined by $\mathcal{K}_{zz}(i, j) := K(z_i, z_j)$. The RKHS optimal function estimator (5) is known in literature as the LS-SVM approach for function estimation.

The quality of the data-fit depends on the selected kernel function $K(\cdot, \cdot)$, the hyper-parameters defining the kernel function and the regularization parameter. A typical choice is the radial basis function (RBF) kernel: $K(z_i, z_j) = \exp\left(- \|z_i - z_j\|_2^2 / \sigma^2 \right)$, where $\sigma > 0$ is tunable hyper-parameter. The selection of the kernel function is rather case specific, where, among others, linear, polynomial, rational, spline and wavelet kernel functions are proposed in literature, see Schölkopf and Smola (2001). The hyper-parameters of the kernel and the regularization parameter γ can be tuned in various ways, here by maximizing the so-called log marginal likelihood function, see Williams and Rasmussen (2006). The marginal likelihood function expresses the likelihood that the mapping $\hat{g}$ maps inputs z to observations w.

Remark 1. If $g : \mathcal{Z} \to \mathbb{R}^{n_g}$ is multidimensional, i.e., $n_g > 1$, then n_g individual functions $g_{(i)}(\cdot)$, are estimated and concatenated to form the n_g-dimensional function $g(\cdot) = \left[g_{(1)}(\cdot) \ \dots \ g_{(n_g)}(\cdot) \right]^\top$.

4. IDENTIFICATION APPROACH

This section presents the three-step identification approach. First, a NARX model is estimated in Section 4.1 to obtain an estimate of the innovation noise sequence e_k. Next, the state sequence is estimated in Section 4.2. Finally, the mappings f and h are estimated in Section 4.3.

4.1 Consistent noise sequence estimation

A nonlinear auto-regressive model with exogenous terms (NARX) is used to estimate the innovation noise sequence. A NARX model is an input-output model that can be written in the following form:

$$y_k = f_{\text{NARX}}(z_k) + e_k, \tag{6a}$$

$$z_k = \left[y_{k-1}^\top \ \dots \ y_{k-p}^\top \ u_{k-1}^\top \ \dots \ u_{k-p}^\top \right]^\top. \tag{6b}$$

In the NARX model (6a), the output $y_k \in \mathbb{R}^l$ at time k is a function of previous outputs y_{k-i} and inputs u_{k-i} for $i = 1, \dots, p$, where p is the past window length.

Let us rewrite system (1) to a NARX model of the form (6a). First, notice that (1a) can be written in the predictor form by substituting e_k in (1a) with (1b):

$$x_{k+1} = f(x_k, u_k, y_k - h(x_k)) =: \tilde{f}(x_k, u_k, y_k), \tag{7}$$

Next, following the notation of (2), x_{k+p} can be written as $x_{k+p} = \tilde{f}^p(x_k, \bar{u}_k^p, \bar{y}_k^p)$. At this stage, in the LTI case, under the assumption that a stable observer exists, it can be shown that for $p \to \infty$, the effect of x_k in x_{k+p} diminishes completely, see Zhu (1987). In fact, this assumption is not only required to transform a state-space model into an ARX model, but is also a fundamental assumption in *any* subspace identification algorithm, see Van der Veen et al. (2013). In the context of subspace identification, it is reasonable to take a sufficiently large p, such that the influence of state x_k in x_{k+p} is negligibly small, see Jansson (2003).

A similar assumption is required in the nonlinear case as well. To this end, we assume that the effect of x_k on $\tilde{f}^p(x_k, \bar{u}_k^p, \bar{y}_k^p)$ is negligible as follows.

Assumption 1. (Fading memory). The NL-SS model (1) satisfies the following condition:

$$\begin{aligned} &\forall r_x, \forall r_u, \forall r_y, \forall x_0, \forall u_0, \forall y_0, \forall \epsilon, \exists P \text{ s.t.} \\ &\quad \forall x_k, \tilde{x}_k \in \mathcal{B}_{x_0}^{r_x}, u_k \in \mathcal{B}_{u_0}^{r_u}, y_k \in \mathcal{B}_{y_0}^{r_y}, p > P \text{ ensures} \\ &\qquad \left\| \tilde{f}^p(x_k, \bar{u}_k^p, \bar{y}_k^p) - \tilde{f}^p(\tilde{x}_k, \bar{u}_k^p, \bar{y}_k^p) \right\|_2 \leq \epsilon \end{aligned}$$

with the ball $\mathcal{B}_{z_0}^r := \{ z \in \mathbb{R}^{n_r} : \|z - z_0\|_2 < r \}$.

This assumption implies that the state can be written as a function of only the p past values of the input and output, up to an error term $\Delta(\epsilon)$:

$$x_k = F(\bar{u}_{k-p}^p, \bar{y}_{k-p}^p) + \Delta(\epsilon),$$

[1] $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is positive semidefinite, if $\forall n \in \mathbb{N}$, $\sum_{i=1}^n \sum_{j=1}^n c_k K(z_k, z_j) c_j \geq 0, \forall \{z_k, c_k\}_{k=1}^n \in \mathcal{Z} \times \mathbb{R}$.

where $F(\bar{u}^p_{k-p}, \bar{y}^p_{k-p}) := \tilde{f}^p(0, \bar{u}^p_k, \bar{y}^p_k)$. By Assumption 1, the error $\Delta(\epsilon)$ can be bounded as $\|\Delta(\epsilon)\|_2 \leq \epsilon$, which implies that the finite memory approximation

$$x_k \approx F(\bar{u}^p_{k-p}, \bar{y}^p_{k-p}). \tag{8}$$

can be made arbitrarily accurate by taking p sufficiently large. Therefore, in the sequel, we assume that $\Delta(\epsilon) = 0$. Using (8) in (1b) results in $y_k = h(F(\bar{y}^p_{k-p}, \bar{u}^p_{k-p})) + e_k = f_{\text{NARX}}(z_k) + e_k$, which is of the form (6a).

For the estimation of f_{NARX}, the function estimator described in the previous section is employed. After defining window length p, the kernel function $K_{\text{NARX}}(\cdot,\cdot)$ and having optimized its hyper-parameters and the regularization parameter γ, the estimate $\hat{f}_{\text{NARX}}$ is given by (5), where the input data z_k is as defined in (6b) and, as output data, $w_k := y_k$ is used. The estimate of the innovation noise sequence is then given by

$$\hat{e}_k = y_k - \hat{f}_{\text{NARX}}(z_k). \tag{9}$$

Assuming $\hat{e}$ is zero mean, the empirical estimate for the noise covariance matrix $\hat{\Sigma}_e$ is given by $\hat{\Sigma}_e = \frac{1}{N}\sum_{k=1}^{N} \hat{e}_k \hat{e}_k^\top$.

This type of NARX modeling is analyzed in De Nicolao and Trecate (1999). There, it is shown that if the input z_k is uncorrelated with e_k, and if some other mild technical conditions on the mapping f hold, then the estimate (5) is consistent, i.e.,

$$\lim_{N\to\infty} \mathbb{E}\left\{ \left\| f_{\text{NARX}} - \hat{f}_{\text{NARX}} \right\|^2_{\mathcal{H}_{\text{NARX}}} \right\} = 0. \tag{10}$$

As e_k is uncorrelated with z_k (even in the closed-loop case due to the absent of a feedthrough term in $h(x_k)$ in (1b)), the estimate of the innovation noise sequence in (9) is also consistent, i.e., the true innovation noise sequence is recovered if the number of data points tends to infinity. Based on this estimate, we define the extended data-set $\tilde{\mathcal{D}} := \{u_k, y_k, \hat{e}_k\}_{k=p+1}^N$.

The consistent estimation of the innovation noise sequence is key in our three-step approach as it allows to perform the subsequent two steps, namely, estimation of the state sequence and estimation of the state-space mappings, also in a consistent manner. In the LTI case, it is already shown that the crucial step of estimating a noise model is essential to proof consistency for the subsequent steps in the closed-loop case, see Van der Veen et al. (2013).

4.2 State sequence estimation

Having the data-set $\tilde{\mathcal{D}}$ at hand, the goal in this section is to provide a method to estimate the state sequence. Verdult et al. (2004) proposed a method to do this based on input and output data. However, in that work, noise was neglected, which could lead to poor model quality in case noise is present, especially in the closed-loop case. Considering the data-set $\tilde{\mathcal{D}}$, we view the estimated innovation noise sequence as an additional pseudo-input. In that way, we can use the method proposed in Verdult et al. (2004) to find an estimate of the state sequence. Furthermore, a form of consistency can be proven for this state sequence estimate, assuming that the true innovation noise sequence was obtained in the previous step. What follows is a brief recap of the approach of Verdult et al. (2004) adapted to our notation.

Canonical Correlation Analysis (CCA), introduced by Hotelling (1936), is a statistical method to study the linear relations between sets of variables. It is the foundation of the so-called intersection-based subspace algorithms for LTI systems, which rely on the state sequence being a minimal interface between past and future input and output data. In those algorithms, CCA is applied to find the state sequence as a linear combination of the past data, such that it optimally predicts the future data. By introducing a kernel function, linear CCA can be performed in a kernel feature space, hence the name Kernel CCA (KCCA). KCCA transforms the sets of variables *nonlinearly* in order to find their maximal correlation.

The basic idea is to, first, write the state x_k as a function of, on the one hand, past inputs and outputs $\bar{\phi}^d_k := \left[(\bar{y}^d_{k-d})^\top \ (\bar{u}^d_{k-d})^\top \ (\bar{e}^d_{k-d})^\top\right]^\top$ and, on the other hand, future inputs and outputs $\bar{\phi}^d_{k+d}$, i.e., $x_k = \Phi_{pp}\left(\bar{\phi}^d_k\right) = \Phi_{ff}\left(\bar{\phi}^d_{k+d}\right)$ where d is the window length. The existence of mapping $\Phi_{pp}\left(\bar{\phi}^d_k\right)$ is ensured by Assumption 1, see (8). To support the existence of mapping $\Phi_{ff}\left(\bar{\phi}^d_{k+d}\right)$, the notion of strong local observability is adopted.

Definition 2. (Nijmeijer (1982)). System (1) is strongly locally observable at (x_k, u_k, e_k) if

$$\operatorname{rank}\left(\frac{\partial h^n(x_k, \bar{u}^{n-1}_k, \bar{e}^n_k)}{\partial x_k}\right) = n.$$

It is assumed that (1) has a fixed-point (x^0, u^0, e^0, y^0), such that $x^0 = f(x^0, u^0, e^0)$ and $y^0 = h(x^0) + e^0$. Next, the following observability assumption is posed.

Assumption 2. System (1) is strongly locally observable at the equilibrium (x^0, u^0, e^0, y^0).

By formulating a KCCA problem on past data $\bar{\phi}^d_k$ and future data $\bar{\phi}^d_{k+d}$ using a LS-SVM approach, we arrive at the regularized generalized eigenvalue problem (RGEP)

$$\begin{bmatrix} \nu_p \mathcal{K}_{pp} + I & 0 \\ 0 & \nu_f \mathcal{K}_{ff} + I \end{bmatrix} \begin{bmatrix} \eta \\ \kappa \end{bmatrix} = \begin{bmatrix} 0 & \mathcal{K}_{ff} \\ \mathcal{K}_{pp} & 0 \end{bmatrix} \begin{bmatrix} \eta \\ \kappa \end{bmatrix} \Lambda, \tag{11}$$

where Λ is the diagonal matrix containing the eigenvalues, ν_p and ν_f are regularization parameters (to be optimized) and $\mathcal{K}_{pp}$ and $\mathcal{K}_{ff}$ are Gram matrices whose elements are evaluations of the selected kernel function $K_{pp}(\cdot,\cdot)$ and $K_{ff}(\cdot,\cdot)$ on the past data $\bar{\phi}^d_k$ and future data $\bar{\phi}^d_{k+d}$, respectively. The state dimension n can be estimated by the number of dominant eigenvalues values in the RGEP (11). Solving this RGEP results in the canonical vectors η and κ. Subsequently, the estimate for the state sequence obtained from future data is $\hat{x} = \kappa_{1:n}^\top K_{ff}$, where $\kappa_{1:n} = [\kappa_1 \ \dots \ \kappa_n]$ and n is the state dimension. Similarly, the estimate for the state obtained from past data is $\check{x} = \eta_{1:n}^\top K_{pp}$. The sequences $\hat{x}$ and $\check{x}$ are estimates of x in an unknown nonlinearly transformed state basis. With the state estimate at hand, the data-set $\tilde{\mathcal{D}}$ is extended with the estimated $\hat{x}_k$ and denoted by $\hat{\mathcal{D}} = \{u_k, y_k, \hat{e}_k, \hat{x}_k\}_{k=p+d+1}^{N-d}$. The regularization parameters ν_p, ν_f, and the hyper-parameters of the kernel functions $K_{pp}(\cdot,\cdot)$ and $K_{ff}(\cdot,\cdot)$ can again be tuned in various ways. In general, this is a nonlinear optimization problem that can be solved, for example, by a grid search.

Consistency of the estimation of the state sequence via this KCCA approach is claimed by Fukumizu et al. (2007) and is understood in the sense that the so-called regularized $\mathcal{F}$-correlation is consistently estimated if the number of data points tends to infinity. Hereto, it is required that the mappings $\Phi_{pp}(\cdot)$ and $\Phi_{ff}(\cdot)$ belong to the RKHS $\mathcal{H}_{pp}$ and $\mathcal{H}_{ff}$, defined through the kernel functions $K_{pp}(\cdot,\cdot)$ and

$K_{ff}(\cdot,\cdot)$, respectively. The consistency claim only holds if the true innovation noise sequence is recovered in the previous step.

4.3 State transition and output mappings estimation

Once the estimate $\hat{x}$ for the state is available, the recursive nature of (1) vanishes and the identification of the state transition map f and the output map h of (1) becomes a static problem. Hereto, again, the function estimator introduced in Section 3 is employed to identify the mappings f and h non-parametrically based on the data-set $\hat{\mathcal{D}}$.

For the identification of f, the input $z_k := \begin{bmatrix}\hat{x}_k^\top & u_k^\top & \hat{e}_k^\top\end{bmatrix}^\top$ and output $w_k := \hat{x}_{k+1}$ are selected. Then, after defining the kernel function $K_f(\cdot,\cdot)$ and tuning its hyper-parameters and the regularization parameter, the identified $\hat{f}$ is given by (5). The mapping h is identified using input $z_k := \hat{x}_k$ and corrected output $w_k := y_k - \hat{e}_k$. Again, after defining the kernel function $K_h(\cdot,\cdot)$ and tuning its hyper-parameters and the regularization parameter, the identified output map $\hat{h}$ is given in (5).

The identified NL-SS model is then given by

$$x_{k+1} = \hat{f}(x_k, u_k, e_k), \quad y_k = \hat{h}(x_k) + e_k, \qquad (12)$$

where the mappings $\hat{f}$ and $\hat{h}$ are non-parametric. This non-parametric model is characterized by the data-set $\hat{\mathcal{D}}$, the kernel functions K_f and K_p, their associated hyper-parameters and the regularization parameters used to estimate $\hat{f}$ and $\hat{h}$.

The identification of the mappings f and h is a special form of the identification of the NARX model in Section 4.1. Therefore, a similar consistency claim as in (10) can be formulated. Again, for consistency here, it is crucial that in the previous steps the true innovation noise sequence and the true state sequence are recovered.

5. ILLUSTRATIVE EXAMPLE

In this section, we assess the performance of the proposed identification approach in a simulation study, both in the open- and closed-loop case. The system under study is inspired by the so-called *logistic map*:

$$x_{k+1} = \frac{1}{2}x_k(1 - x_k) + u_k + e_k, \quad y_k = x_k + e_k. \qquad (13)$$

In the open-loop case, the input u is selected as a zero-mean normal distribution with variance $\sigma_u^2 = 0.01$. The closed-loop case considers the simple feedback law $u_k = r_k - y_k$, where r_k is the reference trajectory, taken from a zero-mean normal distribution with variance $\sigma_r^2 = 0.01$. The innovation noise sequence e_k is drawn from a zero-mean normal distribution with variance σ_e^2, which is chosen to ensure the prescribed Signal-to-Noise Ratio [2] (SNR) of $\{1, 10, 20\}$ dB. For each SNR, a training and validation data-set containing $N = 1000$ samples starting from a zero initial condition is generated.

To assess the influence of using the estimated innovation noise sequence (obtained by identifying a NARX model), three models are identified on each data-set, namely $\mathcal{M}_0, \mathcal{M}_{\hat{e}}$ and $\mathcal{M}_e$. Model $\mathcal{M}_0$ corresponds to the case where noise is not handled in the identification process, implying that only the second step, state estimation by KCCA, and the third step, identification of the state-space maps, are performed. Model $\mathcal{M}_{\hat{e}}$ corresponds to the case where the three-step identification approach is performed, thus the estimated innovation noise sequence $\hat{e}$ is used to estimate the state sequence and to identify the state-space mappings. Model $\mathcal{M}_e$ corresponds to the case where the NARX model returns the true innovation noise sequence, which would be the case for $N \to \infty$ under the consistency claim (10). Thus the true innovation noise sequence e is used in the estimation of the state sequence and the identification of the mappings f and h of model $\mathcal{M}_e$. Performance of each identified model $\mathcal{M}_i$, for $i = \{0, \hat{e}, e\}$, is assessed using the so-called Best Fit Rate [3] (BFR).

Table 1. Open-loop identification results.

Data-Set	Training			Validation		
SNR [dB]	1	10	20	1	10	20
BFR [%] $\mathcal{M}_0$	34.8	64.6	83.5	32.5	62.3	84.8
BFR [%] $\mathcal{M}_{\hat{e}}$	84.4	84.8	98.2	83.9	85.6	97.8
BFR [%] $\mathcal{M}_e$	99.3	99.7	99.5	99.2	99.7	99.5

Regarding implementation, the NARX model and the mappings f and h are identified using the Gaussian Process toolbox, see Rasmussen and Nickisch (2010). The KCCA problem (11) is solved by the KMBOX-toolbox, see Van Vaerenbergh (2010). The hyper-parameters of the KCCA problem are specified as $\nu_p = \nu_f = 3000$ for all cases. A polynomial kernel $k(z_i, z_j) = (z_i^\top z_j + c)^\ell$ with order $\ell = 2$ and constant $c = 1$ is used in all identification steps. The window length $p = 2$ for the estimation of the NARX model is selected and the window length $d = 2$ for the estimation of the state sequence is selected.

Table 1 presents the results of the open-loop case, whereas Table 2 presents the results of the closed-loop case. It can be observed that model M_0 performs the worst, in both the open- and closed-loop case. This is expected, as during identification of this model, no information on the innovation noise sequence is used. The identification approach presented in this paper, yielding the model $M_{\hat{e}}$, shows that estimating the innovation noise sequence using a NARX model indeed improves the model fit quality significantly. However, the quality of the estimated innovation noise sequence determines the quality of the estimated state sequence and, subsequently, the quality of the identified mappings f and h. Therefore, when assuming that the NARX model returns the true innovation noise sequence, as is done in the identification of model M_e, it can be seen that an almost perfect fit is ensured for any SNR, validating the second and third step of the identification approach. Obtaining a good estimate of the innovation noise sequence is a matter of collecting a sufficiently large data-set. It can also be observed that the BFRs in the closed-loop case are generally better than the BFRs in the open-loop case. This is a result of the feedback in the closed-loop case forcing the variance of u_k to become larger than the variance of u_k in the open-loop case.

For the open-loop case with an SNR of 1, a window of the true innovation noise sequence e and the estimated $\hat{e}$ are depicted in the top plot of Figure 2. The bottom plot shows the true response y and the simulated response $\hat{y}$ of model $\mathcal{M}_{\hat{e}}$. For the sake of comparison, also an LTI state-space model is identified on the data-set $\mathcal{D}$, which produces the output $\hat{y}_{\text{LTI}}$, also depicted in the same plot. Clearly, it can be concluded that the nonlinear nature of (13) cannot be captured by an LTI model, which is also reflected in the BFR being only 22.2% for the LTI model.

[2] SNR [dB] $:= 10 \cdot \log_{10}\left(\frac{(y^{\text{det}} - \mu^{\text{det}})(y^{\text{det}} - \mu^{\text{det}})^\top}{ee^\top}\right)$ where y_k^{det} is the output of system (1) for $e_k = 0 \,\forall\, k$ and $\mu^{\text{det}} := \text{mean}(y^{\text{det}})$.

[3] $\text{BFR}(\theta) := 100\% \cdot \max\left(1 - \frac{\|y_k - \hat{y}_k\|_2}{\|y_k - \text{mean}(y_k)\|_2}, 0\right)$.

Table 2. Closed-loop identification results.

Data-Set	Training			Validation		
SNR [dB]	1	10	20	1	10	20
BFR [%] $\mathcal{M}_0$	28.1	69.6	88.1	27.7	69.4	88.6
BFR [%] $\mathcal{M}_{\hat{e}}$	94.1	96.6	98.4	93.4	96.9	98.4
BFR [%] $\mathcal{M}_e$	99.8	99.8	99.8	99.8	99.8	99.8

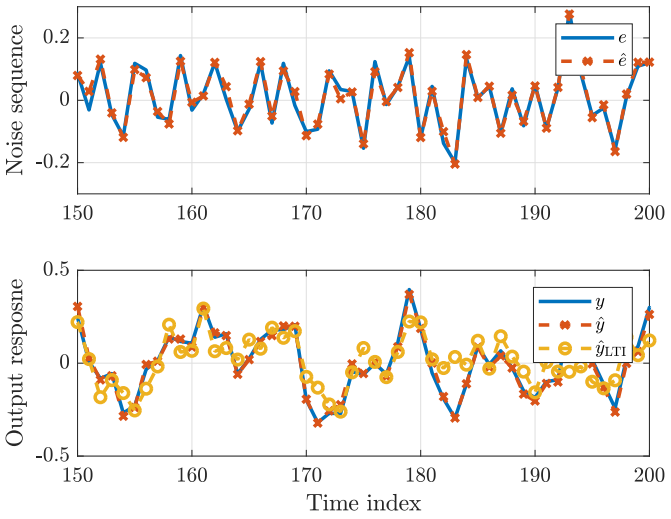

Fig. 2. Top: True innovation noise sequence e and estimated innovation noise sequence $\hat{e}$. Bottom: True response y, response $\hat{y}$ of nonlinear model $\mathcal{M}_{\hat{e}}$ and response $\hat{y}_{\text{LTI}}$ of the identified LTI model.

6. CONCLUSIONS

This paper presents a three-step approach to identification of non-parametric nonlinear state-space models for discrete-time nonlinear systems operating both in open and closed loop. In the first step, a NARX model is identified using a LS-SVM approach, which yields an estimate for the noise sequence. In the next step, the noise sequence is used to estimate the state sequence using KCCA. The final step estimates the state-space mappings using, again, a LS-SVM approach. Although, we do not give an overall consistency proof, the identification approach relies on consistent estimations in each step. Proving overall consistency is considered as a part of future work. In simulation studies, the identification approach obtains accurate predictions on both training and validation data, both in the open-loop and closed-loop case. The proposed approach can be viewed as the non-parametric counterpart for nonlinear systems of the SSARX approach for LTI systems.

REFERENCES

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3), 337–404.

Chiuso, A. and Pillonetto, G. (2019). System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 2, 281–304.

De Nicolao, G. and Trecate, G.F. (1999). Consistent identification of NARX models via regularization networks. *IEEE Transactions on Automatic Control*, 44(11), 2045–2049.

Fukumizu, K., Bach, F.R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb), 361–383.

Giri, F. and Bai, E.W. (2010). *Block-oriented nonlinear system identification*, volume 1. Springer.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.

Jansson, M. (2003). Subspace identification and ARX modeling. *IFAC Proceedings Volumes*, 36(16), 1585–1590.

Khalil, H.K. (1996). Nonlinear systems. *Prentice-Hall, New Jersey*, 2(5), 5–1.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1), 82–95.

Kotta, Ü., Schlacher, K., and Tõnso, M. (2015). Relaxing realizability conditions for discrete-time nonlinear systems. *Automatica*, 58, 67–71.

Larimore, W.E. (1990). Canonical variate analysis in identification, filtering, and adaptive control. In *29th IEEE Conference on Decision and Control*, 596–604. IEEE.

Ljung, L. (1987). *System identification: theory for the user*. Prentice-hall.

Marconato, A., Sjöberg, J., Suykens, J.A., and Schoukens, J. (2013). Improved initialization for nonlinear state-space modeling. *IEEE Transactions on Instrumentation and Measurement*, 63(4), 972–980.

Mercère, G., Markovsky, I., and Ramos, J.A. (2016). Innovation-based subspace identification in open- and closed-loop. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2951–2956. IEEE.

Nijmeijer, H. (1982). Observability of autonomous discrete time non-linear systems: a geometric approach. *International journal of control*, 36(5), 867–874.

Pillonetto, G., Quang, M.H., and Chiuso, A. (2011). A new kernel-based approach for nonlinear system identification. *IEEE Transactions on Automatic Control*, 56(12), 2825–2840.

Rasmussen, C.E. and Nickisch, H. (2010). Gaussian processes for machine learning (gpml) toolbox. *Journal of machine learning research*, 11(Nov), 3011–3015.

Schölkopf, B. and Smola, A.J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schoukens, M. and Tiels, K. (2017). Identification of block-oriented nonlinear systems starting from linear approximations: A survey. *Automatica*, 85, 272–292.

Van der Veen, G., van Wingerden, J.W., Bergamasco, M., Lovera, M., and Verhaegen, M. (2013). Closed-loop subspace identification methods: an overview. *IET Control Theory & Applications*, 7(10), 1339–1358.

Van Overschee, P. and De Moor, B. (2012). *Subspace identification for linear systems: theory-implementation-applications*. Springer Science & Business Media.

Van Vaerenbergh, S. (2010). *Kernel methods for nonlinear identification, equalization and separation of signals*. Ph.D. thesis, University of Cantabria. Software available at `https://github.com/steven2358/kmbox`.

Verdult, V., Suykens, J.A., Boets, J., Goethals, I., and De Moor, B. (2004). Least squares support vector machines for kernel CCA in nonlinear state-space identification. In *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems, Leuven, Belgium*.

Williams, C.K. and Rasmussen, C.E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Zhu, Y. (1987). *Black-box identification of MIMO transfer functions : asymptotic properties of prediction error models*. EUT report. E, Fac. of Electrical Engineering. Technische Universiteit Eindhoven.