State-Space Kriging: A data-driven method to forecast nonlinear dynamical systems

A. Daniel Carnerero, Daniel R. Ramirez, Teodoro Alamo

***Abstract*— This paper presents a new method for modelling dynamical systems. The method uses historical data of the outputs to predict the evolution of the system. The proposed method is based on Direct Weight Optimization and the Kriging method. These data-based methods provide predictions as linear combinations of past outputs after solving a quadratic optimization problem. We introduce a novel methodology that we named *state-space Kriging*, which models the time evolution of the weighting parameters using a state-space formalism. In this way, the potential of Kriging, along with classical estimation methods, as the Kalman filter, can be leveraged to forecast the output of a nonlinear dynamical system. The optimization problems involved are easy to solve, and analytical solutions are provided. Some numerical examples and comparisons are provided to demonstrate the effectiveness of our proposal.**

***Index Terms*— Machine Learning, Identification, Estimation, Kalman filtering.**

I. Introduction

The objective of system identification is to find a certain mathematical model of a system that fits with some input-output data measurements, obtained from the real system [1]. The most common way to define fitting would be to minimize the forecasting error with respect to the available data. Letting y_i be the i-th real output and $\tilde{y}_i(\theta)$ the i-th forecasting obtained with the parameters θ, then the parameters θ of the model could be chosen according to

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \left(y_i - \tilde{y}_i(\theta)\right)^T \left(y_i - \tilde{y}_i(\theta)\right).$$

The well known Least-Squares Estimator is obtained if the predictor is chosen as

$$\tilde{y}_i(\theta) = z_i^T \theta^*,$$

being $z_i \in \mathcal{R}^{n_y}$ the time delay embedding of the system, compounded by some past outputs, that is, $z_i = [y_{i-1}, y_{i-2}, ..., y_{i-n_y}]^T$ and $\theta \in \mathcal{R}^{n_y}$. This estimator is known to provide good results in the linear case under simple identifiability conditions. When the dimension of z_i is much larger than the number of samples, Dynamic Mode Decomposition (DMD) techniques can be applied in order to tackle the problems that arise from working with high dimensional data [2].

Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n, 41020 Sevilla, Spain (e-mail: {acarnerero,danirr,talamo}@us.es).

On the other hand, nonlinear system identification is a more complicated task for several reasons, see, for example, [3]. It is also a more interesting field because it allows us to model more complex systems and thus reducing model mismatches. The first challenging task would be to select a model archetype for the system under consideration. As we briefly discuss below, there are many possibilities in the nonlinear system identification literature.

Typically, when no assumption is made on the structure of the system, a universal approximator can be used. Neural networks, and more recently deep neural networks [4], [5] are a popular tool in system identification because they are known for being able to reproduce any nonlinear function. However, the choices on the type of network used, training algorithm and structure are not easy, as a bad set of parameters could lead to wrong results, cause overfitting, etc. [6]. As an alternative, one could rely on Reservoir Computing [7] or Echo State Networks [8] approaches which are easier to train. Hinging hyperplanes are also capable of approximating any function within a bounded error and have also been used in the context of system identification and control applications [9]. Other black-box model approaches for system identification that do not make significant assumptions on the non-linearity of the systems include the NARMAX methods [10], Gaussian Processes [11] and Lipschitz interpolation which has been recently applied in the context of system identification and control [12], [13].

An alternative approach is to assume that the system behaves as a time-varying linear (LTV) system [14], [15]. Related to this would be to consider the system as an interpolation between the different computed linear models, taking into account some rules. This method leads to models based on fuzzy logic [16]. However, these models do not fit correctly in the regions where there is not enough data [17]. Another option would be to rely on Hammerstein-Wiener [18] or Volterra [19] models in order to represent the nonlinearities of the system.

Other techniques based on obtaining predictions as a linear combination of past outputs can be used. This includes Direct Weight Optimization (DWO) [20] and the Kriging method [21], [22]. See also [23] for a similar technique in the context of interval predictions. The objective of these techniques is, given z_k, to obtain a prediction $\tilde{y}_k$ as a combination of past outputs by means of an optimization problem. For that purpose, it is necessary to compute certain weights called λ. These methods have been tested in some applications (i.e. time-series forecasting) with good results [24] and have been proven to be an extension of the least-squares problem [25]. In the field of dynamic systems, they have been used, for

example, to quantify the uncertainty of a surrogate model [26]. Also, kriged models have been used in fault detection applications [27], for noise cancellation [28] and also in the spatial estimation of measurements in sensor networks using kriged distributed Kalman filters [29]. In [30], Kriging models obtained from the generalized least squares estimator and a basis of precomputed regression functions are explored. In the context of designing a NMPC controller, [31] uses a basic kriged black box prediction model, whereas in [32] it is used to learn the actual control law. However, usually the problem is considered strictly static and thus vector λ is computed from scratch at every time instant without taking into account the past, which could be potentially beneficial.

In this paper, a new model for nonlinear system identification and forecasting is proposed. The model is based on a state-space representation of the weights that arise from a modified Kriging method. The obtained model is an LTV approximation of the original nonlinear system. To enhance the approximation, local data is encouraged to be used in the computation of the model at each time instant k. Moreover, the proposed state-space representation allows us to enhance the forecast by means of classic estimation approaches as the Kalman filter [33].

The main contributions of this paper are: developing a new scheme based on a state-space representation of λ that models the evolution of the Kriging weights instead of a black box model, the use of local data to improve the forecasting of nonlinear systems, adding a regularization term to λ to make the system less sensitive to uncertainty and, finally, proposing a way to apply the Kalman filter to the developed strategy in order to improve the predictions in presence of noisy data and/or disturbances.

The paper is organized as follows. In Section II it is shown how Kriging can be used to forecast the output of a nonlinear dynamical system, whereas in Section III the alternative to dynamic Kriging based on a different state-space representation is introduced. Section IV shows how the Kalman filter can be applied to the proposed Kriging state-space formulation. In Section V, the proposed methodology is applied to different numerical examples. The paper ends with a section of conclusions.

II. Dynamic Kriging

Consider an autonomous discrete nonlinear system

$$\begin{aligned} x_{k+1} &= f(x_k) \\ y_k &= h(x_k), \end{aligned} \tag{1}$$

where k is the time instant, $x_k \in \mathcal{R}^{n_x}$ is the state of the system, $y_k \in \mathcal{R}^{n_y}$ is the output of the system, whereas $f(\cdot)$ and $h(\cdot)$ are unknown nonlinear functions such that $f(\cdot) : \mathcal{R}^{n_x} \to \mathcal{R}^{n_x}$ and $h(\cdot) : \mathcal{R}^{n_x} \to \mathcal{R}^{n_y}$.

The objective of this section is to describe how Kriging can be used to obtain an LTV model of the outputs of (1). We will assume that the only available data are the measurable outputs. We define a time delay embedding vector $z_k \in \mathcal{R}^{n_z}$ containing the n_p past outputs of the system[1]. That is, $z_k = [y_{k-1}^T, y_{k-2}^T, ..., y_{k-n_p}^T]^T \in \mathcal{R}^{n_z}$, where $n_z = n_p n_y$. We also denote with z_k^+ as the successor of z_k, i.e. $z_k^+ = [y_k^T, y_{k-1}^T, ..., y_{k-n_p+1}^T]^T \in \mathcal{R}^{n_z}$.

From now on, assume that some historical data of the plant is stored in a database in the form of matrices:

$$\begin{aligned} D &= \left[\begin{array}{cccc} \bar{z}_1 & \bar{z}_2 & \dots & \bar{z}_N \end{array} \right], \\ D^+ &= \left[\begin{array}{cccc} \bar{z}_1^+ & \bar{z}_2^+ & \dots & \bar{z}_N^+ \end{array} \right], \end{aligned}$$

where $N > n_z$ is the number of data points, $\bar{z}$ refers to a sample of z and D^+ is the matrix successor of D. The indexes of the columns of D and D^+ do not refer to the sample time, but to the position in the matrix. Therefore, $\bar{z}_{i+1}$ is not necessarily the successor sample of $\bar{z}_i$. At sample time k, an estimation of the successor of z_k, denoted as $\tilde{z}_{k+1}$, can be obtained by a linear combination of the columns of matrix D^+ using a vector of optimal weights $\lambda_k^* \in \mathcal{R}^N$

$$\tilde{z}_{k+1} = D^+ \lambda_k^*.$$

As in Kriging and DWO methods, this vector of weights is obtained from the following optimization problem

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & \lambda_k^T H_1 \lambda_k \\ s.t. \quad & \left[\begin{array}{c} D \\ \mathbf{1} \end{array} \right] \lambda_k = \left[\begin{array}{c} z_k \\ 1 \end{array} \right], \end{aligned} \tag{2}$$

where $H_1 \in \mathcal{R}^{N \times N}$ is a positive definite weighting matrix and $\mathbf{1}$ a row vector with all its components equal to 1. Forcing the components of λ_k to sum one is equivalent to including a bias term in the estimation process. The simplest choice is making H_1 equal to the identity matrix I_N (other possibilities are explored in the next section). The optimization problem can be rewritten as

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & \lambda_k^T H_1 \lambda_k \\ s.t. \quad & C\lambda_k = b\,. \end{aligned} \tag{3}$$

with

$$C = \left[\begin{array}{c} D \\ \mathbf{1} \end{array} \right], \quad b = \left[\begin{array}{c} z_k \\ 1 \end{array} \right].$$

In order to guarantee that any point in $\mathcal{R}^{n_z+1}$ can be expressed as a linear combination of the columns of C, we assume that matrix C is full row rank. This equality constrained quadratic problem has an analytic solution that can be obtained computing the Lagrangian and its derivative (see [34, §10.1.1]):

$$L(\lambda_k, \nu) = \lambda_k^T H_1 \lambda_k + \nu^T (C\lambda_k - b)$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} L(\lambda_k, \nu) = 2H_1\lambda_k + C^T\nu\,,$$

where ν is the dual variable associated with the equality constraint. From the Karush-Kuhn-Tucker (KKT) conditions:

$$2H_1\lambda_k^* + C^T\nu^* = 0\,, \tag{4}$$

which leads to

$$\lambda_k^* = \frac{-H_1^{-1}C^T\nu^*}{2}.$$

[1] Also, nonlinear terms of the outputs could be included.

Pre-multiplying this equality by C, and taking into account that $C\lambda_k^* = b$, we have

$$\nu^* = -2\left(CH_1^{-1}C^T\right)^{-1} b,$$

which applied to equation (4) yields

$$\lambda_k^* = H_1^{-1}C^T\left(CH_1^{-1}C^T\right)^{-1}\begin{bmatrix} z_k \\ 1 \end{bmatrix}.$$

In order to predict z_{k+d}, with $d > 1$, one could use this approach in a recursive way. That is, the i-th ahead prediction $\tilde{z}_{k+i}$ could be used to compute

$$\lambda_{k+i}^* = H_1^{-1}C^T\left(CH_1^{-1}C^T\right)^{-1}\begin{bmatrix} \tilde{z}_{k+i} \\ 1 \end{bmatrix},$$

and thus obtaining $\tilde{z}_{k+i+1} = D^+\lambda_{k+i}^*$. In the next section we propose a modification of this *naive* recursive method. The novel methodology relies on a time-varying state-space modelling of the optimal weighting vector parameter λ_k^*.

III. State-space Kriging

Suppose that the prediction $\tilde{z}_k$ of z_k is obtained from $\tilde{z}_k = D^+\lambda_{k-1}^*$, where the sum of the components of λ_{k-1}^* is assumed to be equal to one. In order to model how the dynamics of the optimal vector of weights λ_k^* depends on λ_{k-1}^*, we add a regularization term to optimization problem (2) that penalizes the difference between λ_k^* and λ_{k-1}^*. In this way, vector λ_k^* not only fulfills the required equality constraints, but also does not depart excessively from λ_{k-1}^*. This will reduce the sensitivity to noise of the identified dynamics. Thus, given $\tilde{z}_k$, λ_k^* is obtained from

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & (\lambda_k - \lambda_{k-1}^*)^T H_2(\lambda_k - \lambda_{k-1}^*) + \lambda_k^T H_1 \lambda_k \\ s.t. \quad & \begin{bmatrix} D \\ \mathbf{1} \end{bmatrix}\lambda_k = \begin{bmatrix} \tilde{z}_k \\ 1 \end{bmatrix}, \end{aligned}$$

where $H_2 \in \mathcal{R}^{N\times N}$ is chosen as $H_2 = \tau I_N$ being $\tau > 0$ a tuning parameter of the proposed methodology that could be selected by cross-validation [1, §16.5]. Because of the assumptions on λ_{k-1}^*, the previous optimization problem can be rewritten as

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & (\lambda_k - \lambda_{k-1}^*)^T H_2(\lambda_k - \lambda_{k-1}^*) + \lambda_k^T H_1 \lambda_k \\ s.t. \quad & \begin{bmatrix} D \\ \mathbf{1} \end{bmatrix}\lambda_k = \begin{bmatrix} D^+ \\ \mathbf{1} \end{bmatrix}\lambda_{k-1}^*. \end{aligned}$$

Thus, we have

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & (\lambda_k - \lambda_{k-1}^*)^T H_2(\lambda_k - \lambda_{k-1}^*) + \lambda_k^T H_1 \lambda_k \\ s.t. \quad & C\lambda_k = C^+\lambda_{k-1}^*, \end{aligned} \tag{5}$$

with

$$C = \begin{bmatrix} D \\ \mathbf{1} \end{bmatrix},\ C^+ = \begin{bmatrix} D^+ \\ \mathbf{1} \end{bmatrix}.$$

Note that λ_k^* is determined only by λ_{k-1}^* and matrices C and C^+. Optimization problem (5) can be rewritten as

$$\begin{aligned} \lambda_k^* = \arg\min_{\lambda_k} \quad & \frac{1}{2}\lambda_k^T H \lambda_k + f^T\lambda_k \\ s.t. \quad & C\lambda_k = b, \end{aligned} \tag{6}$$

with $H = 2(H_1 + H_2)$, $f = -2H_2\lambda_{k-1}^*$ and $b = C^+\lambda_{k-1}^*$. Also note that we get rid of the constant term $\lambda_{k-1}^{*T}H_2\lambda_{k-1}^*$ because it will not affect the solution λ_k^*. The Lagrangian of this problem is given by

$$L(\lambda_k, \nu) = \frac{1}{2}\lambda_k^T H\lambda_k + f^T\lambda_k + \nu^T(C\lambda_k - b).$$

Note that ν is the dual variable associated with the equality constraint. The derivative of the Lagrangian is

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_k}L(\lambda_k, \nu) = H\lambda_k + f + C^T\nu.$$

In the optimum the derivative fulfills the KKT conditions [34, §10.1.1]. That is,

$$H\lambda_k^* + f + C^T\nu^* = 0,$$

and thus

$$\lambda_k^* = -H^{-1}f - H^{-1}C^T\nu^*. \tag{7}$$

Pre-multiplying both sides of last equality by C yields

$$b = C\lambda_k^* = -CH^{-1}f - CH^{-1}C^T\nu^*,$$

and thus $\nu^* = (CH^{-1}C^T)^{-1}(-CH^{-1}f - b)$. Substituting this into equation (7) we obtain

$$\lambda_k^* = H^{-1}C^T\left(CH^{-1}C^T\right)^{-1}\left(CH^{-1}f + b\right) - H^{-1}f.$$

Taking into account that $f = -2H_2\lambda_{k-1}^*$ and $b = C^+\lambda_{k-1}^*$, we have

$$\lambda_k^* = A\lambda_{k-1}^*, \tag{8}$$

with

$$A = 2H^{-1}H_2 + H^{-1}C^T(CH^{-1}C^T)^{-1}(C^+ - 2CH^{-1}H_2).$$

Note that this means that λ_k follows linear dynamics. Thus, we have obtained a new model for the outputs of system (1) using historical data of these outputs. This new autonomous system allows us to compute the next values of λ and z. With an abuse of notation, we drop here the * and $\tilde{}$ to write the alternative model of the outputs of (1) as

$$\begin{aligned} \lambda_{k+1} &= A\lambda_k \\ z_k &= D\lambda_k. \end{aligned} \tag{9}$$

Note that the previous model is linear and time-invariant as the matrix A is constant. However, it is possible to weigh the points in the data set with respect to z_k, which would encourage the use of local data and thus provide better results when interpolating nonlinear systems. This can be done by choosing H_1 appropriately. For example, $H_1 \in \mathcal{R}^{N\times N}$ could be a diagonal matrix whose elements are computed with a function $g(z\,,\,D) : \mathcal{R}^{n_z} \times \mathcal{R}^{n_z\times N} \to \mathcal{R}^N$ that measures the dissimilarity (see [25] and chapter 2 of [35]) of the current z_k to the points saved in the data set. In the numerical examples of section V, the squared Euclidean distance is used. That is, for a given z

$$g(z, D) = \begin{bmatrix} (z - \bar{z}_1)^T(z - \bar{z}_1) \\ \vdots \\ (z - \bar{z}_N)^T(z - \bar{z}_N) \end{bmatrix}, \tag{10}$$

and thus

$$H_1 = \text{diag}(g(z, D)) .$$

where, given $u \in \mathcal{R}^N$, diag(u) denotes a diagonal matrix $\mathcal{R}^N \times \mathcal{R}^N$ whose non-zero entries are the components of u. Note that, unlike the previous case, the matrix A will be calculated at each sample time k, leading to an LTV dynamics for λ and z, that is,

$$\begin{aligned} \lambda_{k+1} &= A_k \lambda_k \\ z_k &= D\lambda_k . \end{aligned} \tag{11}$$

IV. Kalman filter for state-space Kriging

Under the assumption that $z_k = D^+\lambda_{k-1}$ and $\mathbf{1}\lambda_{k-1} = 1$, the following nominal LTV state-space model was derived in Section III:

$$\begin{aligned} \lambda_{k+1} &= A_k \lambda_k \\ z_k &= D\lambda_k . \end{aligned}$$

In order to address the existence of noise (D and D^+ contain noisy data), disturbances and modelling mismatches (i.e. $D^+\lambda_{k-1}$ is just an estimation of z_k), we modify this nominal model and include disturbance and model errors (w_k) and measurement noise (v_k). That is,

$$\begin{aligned} \lambda_{k+1} &= A_k \lambda_k + w_k \\ z_k &= D\lambda_k + v_k . \end{aligned}$$

In order to enforce the equality $\mathbf{1}\lambda_k = 1$, we consider an extended output defined as

$$z'_k = \begin{bmatrix} D \\ \mathbf{1} \end{bmatrix} \lambda_k + \begin{bmatrix} v_k \\ 0 \end{bmatrix} . \tag{12}$$

Defining $C = \begin{bmatrix} D \\ \mathbf{1} \end{bmatrix}$ and $v'_k = \begin{bmatrix} v_k \\ 0 \end{bmatrix}$ we obtain the extended system:

$$\begin{aligned} \lambda_{k+1} &= A_k \lambda_k + w_k \\ z'_k &= C\lambda_k + v'_k . \end{aligned}$$

In order to improve the estimation of λ_k, correcting it with each new measurement z_k we propose to use the well known Kalman filter [33]. Note that in our case, the state in the Kalman filter is λ_k and the output is z'_k. The notation will be as follows: $\tilde{\lambda}_k$ refers to the predicted value of λ_k whereas the corrected version of $\tilde{\lambda}_k$ will be denoted as $\hat{\lambda}_k$.

We assume that w_k and v_k are uncorrelated white noise signals fulfilling

$$\text{E}(w_k w_k^T) \leq W_k, \quad \text{E}(v_k v_k^T) \leq V_k,$$

that is, their covariance is bounded. From the bound on the covariance of v_k we obtain a bound on the covariance of v'_k:

$$\text{E}(v'_k v'^T_k) = \text{E}\left(\begin{bmatrix} v_k v_k^T & 0 \\ 0 & 0 \end{bmatrix} \right) \leq \begin{bmatrix} V_k & 0 \\ 0 & 0 \end{bmatrix} = V'_k .$$

We denote $\tilde{P}_k$ the bound on the covariance of the (non corrected) estimation error $\lambda_k - \tilde{\lambda}_k$:

$$\text{E}\left((\lambda_k - \tilde{\lambda}_k)(\lambda_k - \tilde{\lambda}_k)^T \right) \leq \tilde{P}_k .$$

We are now in a position to apply Kalman filter [33]. Given an estimation $\tilde{\lambda}_k$, we obtain a corrected version $\hat{\lambda}_k$ from

$$\hat{\lambda}_k = \tilde{\lambda}_k + K_k \left(z'_k - C\tilde{\lambda}_k \right) ,$$

where K_k is the optimal gain, which is calculated as

$$K_k = \tilde{P}_k C^T \left(C\tilde{P}_k C^T + V'_k \right)^{-1} .$$

We notice that the corrected vector $\hat{\lambda}_k$ fulfills the equation $1 = \mathbf{1}\hat{\lambda}_k$ because the last component of the extended output ($\mathbf{1}\lambda_k$) is not affected by noise (see (12)). However, the equality $z_k = D\hat{\lambda}_k$ is not necessarily satisfied because of the noise term v_k. This term prevents the Kalman filter from departing arbitrarily from the initial estimation $\tilde{\lambda}_k$ to satisfy $z_k = D\hat{\lambda}_k$. The size of $z_k - D\hat{\lambda}_k$ will be determined by the relative sizes of the covariance bounds V_k and $\tilde{P}_k$.

Following the formulation of the Kalman filter, the matrix $\tilde{P}_{k+1}$ is computed as

$$\tilde{P}_{k+1} = A_k \hat{P}_k A_k^T + W_k ,$$

where $\hat{P}_k = \tilde{P}_k - K_k C \tilde{P}_k$. Finally, $\tilde{\lambda}_{k+1}$ is obtained from

$$\tilde{\lambda}_{k+1} = A_k \hat{\lambda}_k$$

The covariance matrices describing disturbance, noise, and initial uncertainty on λ_0 could be set as the identity matrix multiplied by scalars that are considered tuning parameters.

V. Numerical Examples

In this section, two examples are provided to show the effectiveness of the proposed strategy. In both examples, H_1 has been obtained using the squared Euclidean distance, see equation (10). Four baselines based on Gaussian Processes (GPs), Nonlinear ARX models (NARX), Reservoir computing (RC) and Dynamic Mode Decomposition (DMD) are provided to compare the results obtained with our proposed approach (SS-K).

A. Sunspot Number

Forecasting the sunspot number is considered quite difficult as the time series is nonstationary and because the nature of its dynamics is unknown. Monthly observations of the historical evolution of the number of sunspots since 1749 will be used in this example. We will use the first 2500 samples as matrices D and D^+ with a regressor $z_k = [y_{k-1}, \ldots, y_{k-40}]$. The next 150 samples will be used as a test set. The observations are assumed to be noise-free and thus the approach of section III is used. The prediction will be done from time instant $k = 0$ exclusively. That is, the prediction will be k-step ahead, with $k = 1, \ldots, 150$. Note that the forecasting horizon is quite long (it comprises more than a solar cycle), making the forecasting task even harder. Here, in the proposed approach, $\tau = 2500$, NARX and GPs are computed using Matlab functions ("nlarx" and "fitrgp") and the RC implementation considers a reservoir size of 300, a leakage rate of 0.9 and a spectral radius of 0.4.

Figure 1 and table I show the forecasting results. Note that only the proposed approach is shown in the figure for the sake of clarity. It can be seen that the proposed approach works better than the aforementioned baselines, obtaining smaller errors and standard deviations in general.

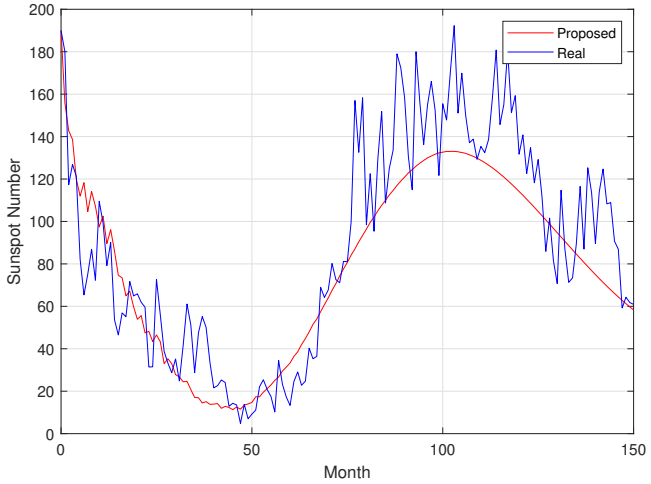

Fig. 1. Forecasting the Sunspot Number 150 steps ahead.

	SS-K	GP	NARX	DMD	RC
MSE	608.26	999.77	991.11	5795.6	6048.1
Std	22.646	35.147	26.490	72.897	62.552

TABLE I
MSE FOR THE SUNSPOT NUMBER EXAMPLE.

B. Rössler Attractor

Consider now the system described by the following set of differential equations

$$\begin{aligned} \dot{o} &= -p - l \\ \dot{p} &= o + ap \\ \dot{l} &= b + l(o - c)\,, \end{aligned}$$

also known as the Rössler attractor. The set of parameters considered in this example is $a = 0.2$, $b = 0.2$ and $c = 5.7$ which are known to correspond to a chaotic behaviour. In order to obtain samples of the continuous system, we integrate numerically the equations with a fixed sample time of 0.1 seconds a total simulation time of 20 seconds starting from random initial points in the space, making a total of 1000 samples in the matrices D and D^+ (comprising 5 trajectories of 200 samples each one). The regressor considered here is $z_k = [o_{k-1}, p_{k-1}, l_{k-1}]$.

Our objective is to show the effectiveness of the proposed strategy to model nonlinearity with measurement noise. For that reason, we will consider that the measurements obtained from the Rössler attractor are noisy. This noise will follow a normal distribution with zero mean and unit variance $\mathcal{N}(0, 1)$ and is completely uncorrelated (that is, the noise of each state is also uncorrelated to that of the other states). Thus, the scheme of section IV is used here. Also, note that here the forecasting is done 1-step ahead instead of k-steps ahead from the previous sample, in order to be able to apply the methodology of the Kalman filter. That is, at each time instant k the value of the output is sampled and the forecasting at $k-1$ is corrected with this new measurement. After that, the prediction for $k+1$ is done with all the information available at k (which includes the corrected measurements). 100 trajectories of 15 seconds obtained from random initial conditions are considered. Here, $\tau = 5$, $W_k = 1.65 \cdot 10^{-5}$ and the variance of v_k is assumed to be known. On the other hand, GPs are computed using a radial basis kernel with a regularization term equal to the true variance of the process. The RC implementation considers a reservoir size of 50, a leakage rate of 0.2 and a spectral radius of 0.3. Finally, we added a Kalman filtering layer to the linear system obtained with the DMD in this section (K-DMD), where the variance of the noise is also assumed to be known.

The results are shown in figure 2, and table II. Figure 2 shows the real evolution of the state along with the noisy measurement of the output and the prediction obtained with the proposed approach for a representative trajectory. Table II summarizes the numerical results of the experiment. It can be seen that both the proposed strategy and the DMD with a Kalman filter achieve the best results both in MSE and standard deviation. Taking into account the ratio between trajectories in the validation and training set, the results for both strategies are quite remarkable. Note however, that the proposed strategy achieves the best overall results in both numerical examples, as the DMD open loop predictions in the previous examples were clearly worse.

On the other hand, computational times of the different baselines for the Rössler atractor are provided in table III. Although the proposed approach seems to be most the costly method, it remains in the order of 60 milliseconds, what can be considered fast enough for time series/dynamic systems forecasting. Also, it should be noted that most of the computation time is due to the Kalman filter step. The computation times would drop to 12.477 milliseconds without the Kalman filter.

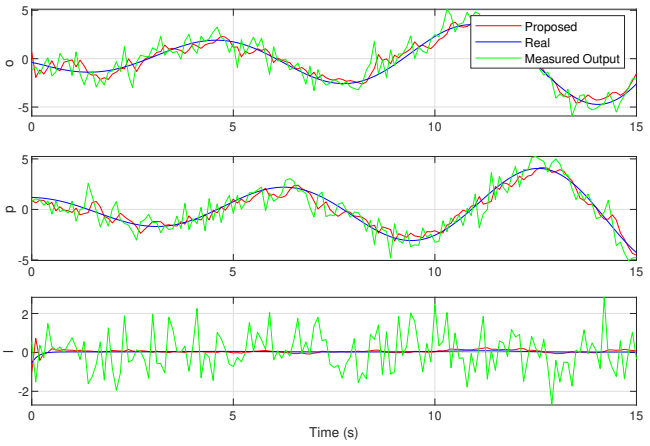

Fig. 2. Forecasting the noisy Rössler attractor with Kalman filtering.

	SS-K	GP	NARX	K-DMD	RC
MSE	0.275	0.638	0.635	0.276	0.334
Std	0.524	0.799	0.797	0.526	0.578

TABLE II
MSE FOR THE NOISY RÖSSLER ATTRACTOR.

SS-K	GP	NARX	K-DMD	RC
62.909	2.591	14.872	0.036	0.868

TABLE III
AVERAGE ONLINE COMPUTATIONAL TIME IN MILLISECONDS (RÖSSLER ATTRACTOR).

VI. CONCLUSIONS

This work presents a new data-based methodology to approximate nonlinear systems with an LTV model. The approach is based on a modified Kriging problem in which a regularization term is added. From the solution of the Kriging problem, a state-space representation of the nonlinear dynamics is obtained. In this new model, the Kriging weights form the state and the output of the system is also obtained from those weights. This structure allows for the application of the Kalman filter to improve the forecasting in noisy measurements. The results in the numerical examples show that the proposed strategy can be compared favourably with some baseline approaches. Future works will consider the combination of the proposed strategy with NN and DMD methods in order to enhance the predictions.

ACKNOWLEDGMENTS

This work was supported by the Agencia Estatal de Investigación (AEI)-Spain under Grant PID2019-106212RB-C41/AEI/10.13039/501100011033 and also by Junta de Andalucía and FEDER funds under grant P20_00546.

REFERENCES

[1] L. Ljung, *System Identification: Theory for the User*. Pearson Education, 1998.
[2] D. F. Gomez, F. D. Lagor, P. B. Kirk, A. H. Lind, A. R. Jones, and D. A. Paley, "Data-driven estimation of the unsteady flowfield near an actuated airfoil," *Journal of Guidance, Control, and Dynamics*, vol. 42, no. 10, pp. 2279–2287, 2019.
[3] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented road map," *IEEE Control Systems Magazine*, vol. 39, no. 6, pp. 28–99, 2019.
[4] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200209, 2021.
[5] S. H. Rudy, J. N. Kutz, and S. L. Brunton, "Deep learning of dynamics and signal-noise decomposition with time-stepping constraints," *Journal of Computational Physics*, vol. 396, pp. 483–506, 2019.
[6] S. Lawrence and C. L. Giles, "Overfitting and neural networks: conjugate gradient and backpropagation," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks.*, vol. 1. IEEE, 2000, pp. 114–119.
[7] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, "Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach," *Physical review letters*, vol. 120, no. 2, p. 024102, 2018.
[8] D. Goswami, A. Wolek, and D. A. Paley, "Data-driven estimation using an echo-state neural network equipped with an ensemble Kalman filter," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 2549–2554.
[9] D. Ramírez, E. Camacho, and M. Arahal, "Implementation of min–max MPC using hinging hyperplanes. application to a heat exchanger," *Control Engineering Practice*, vol. 12, no. 9, pp. 1197–1205, 2004.
[10] S. A. Billings, *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013.
[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.
[12] J.-P. Calliess, "Lipschitz optimisation for lipschitz interpolation," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 3141–3146.
[13] J. M. Manzano, D. Limon, D. M. de la Peña, and J.-P. Calliess, "Robust learning-based MPC for nonlinear constrained systems," *Automatica*, vol. 117, p. 108948, 2020.
[14] K. Liu, "Identification of linear time-varying systems," *Journal of Sound and Vibration*, vol. 206, no. 4, pp. 487–505, 1997.
[15] P. L. Dos Santos, T. P. A. Perdicoulis, C. Novara, J. A. Ramos, and D. E. Rivera, *Linear parameter-varying system identification: New developments and trends*. World Scientific, 2011, vol. 14.
[16] K. Zeng, N.-Y. Zhang, and W.-L. Xu, "A comparative study on sufficient conditions for Takagi-Sugeno fuzzy systems as universal approximators," *IEEE Transactions on fuzzy systems*, vol. 8, no. 6, pp. 773–780, 2000.
[17] P. Hušek, "System identification using monotonic fuzzy models," in *Recent Developments and the New Direction in Soft-Computing Foundations and Applications*. Springer, 2020, pp. 229–242.
[18] A. Wills, T. B. Schön, L. Ljung, and B. Ninness, "Identification of Hammerstein-Wiener models," *Automatica*, vol. 49, no. 1, pp. 70–81, 2013.
[19] J. Gruber, D. Ramirez, D. Limon, and T. Alamo, "Computationally efficient nonlinear min–max model predictive control based on Volterra series models—application to a pilot plant," *Journal of Process Control*, vol. 23, no. 4, pp. 543–560, 2013.
[20] J. Roll, A. Nazin, and L. Ljung, "Nonlinear system identification via direct weight optimization," *Automatica*, vol. 41, no. 3, pp. 475–490, 2005.
[21] N. Cressie, "Kriging nonstationary data," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 625–634, 1986.
[22] J. P. Kleijnen, "Kriging metamodeling in simulation: A review," *European journal of operational research*, vol. 192, no. 3, pp. 707–716, 2009.
[23] J. M. Bravo, T. Alamo, M. Vasallo, and M. E. Gegúndez, "A general framework for predictors based on bounding techniques and local approximation," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3430–3435, 2017.
[24] G. Alfonso, A. D. Carnerero, D. R. Ramirez, and T. Alamo, "Stock forecasting using local data," *IEEE Access*, vol. 9, pp. 9334–9344, 2020.
[25] A. D. Carnerero, D. R. Ramirez, and T. Alamo, "Probabilistic interval predictor based on dissimilarity functions," *arXiv preprint arXiv:2010.15530*, 2020.
[26] B. Bhattacharyya, E. Jacquelin, and D. Brizard, "Uncertainty quantification of nonlinear stochastic dynamic problem using a Kriging-NARX surrogate model," in *3rd International Conference on Uncertainty Quantification in Computational Sciences and Engineering, ECCOMAS*, 2019, pp. 34–46.
[27] A. Shokry, M. H. Ardakani, G. Escudero, M. Graells, and A. Espuña, "Dynamic kriging based fault detection and diagnosis approach for nonlinear noisy dynamic processes," *Computers & Chemical Engineering*, vol. 106, pp. 758–776, 2017.
[28] J.-P. Costa, L. Pronzato, and E. Thierry, "Nonlinear prediction by kriging, with application to noise cancellation," *Signal Processing*, vol. 80, no. 4, pp. 553–566, 2000.
[29] J. Cortés, "Distributed Kriged Kalman filter for spatial estimation," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2816–2827, 2009.
[30] A. F. Hernandez and M. G. Gallivan, "An exploratory study of discrete time state-space models using kriging," in *2008 American Control Conference*. IEEE, 2008, pp. 3993–3998.
[31] J. Marzat and H. Piet-Lahanier, "Design of nonlinear MPC by Kriging-based optimization," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 1490–1495, 2012.
[32] J. R. Salvador, D. R. Ramirez, T. Alamo, and D. M. de la Peña, "Offset free data driven control: application to a process control trainer," *IET Control Theory & Applications*, vol. 13, no. 18, pp. 3096–3106, 2019.
[33] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 03 1960. [Online]. Available: https://doi.org/10.1115/1.3662552
[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[35] S. T. Wierzchoń and M. Kłopotek, *Modern algorithms of cluster analysis*. Springer, 2018.