Machine Learning meets Kalman Filtering

Andrea Carron and Marco Todescato and Ruggero Carli and Luca Schenato and Gianluigi Pillonetto

***Abstract*— In this work we study the problem of non-parametric estimation for non-linear time-space dynamic stochastic processes and in particular for Gaussian processes (GP). GP methods have been mainly applied to spatial regression and represent the state of the art for machine learning thanks to their universal representing properties. However, their extension to dynamic processes has been elusive so far since standard machine learning tools give rise unscalable algorithms. In this work we propose a systematic and explicit procedure to address this problem by pairing GP regression with Kalman Filtering. In particular, under the specific time-space separability assumption of the kernel which models process and periodic sampling on a (possibly non-uniform) space-grid, we show how to build a finite dimensional discrete-time state-space exact representation for the modeled process. The major finding is that the state at instant k of the associated Kalman Filter represents a sufficient statistic to compute the minimum variance prediction of the process at instant k over any arbitrary finite subset of the space. The proposed strategy is then compared with the standard non-parametric estimation and truncated non-parametric estimation strategies both in terms of estimation performance and computation complexity.**

***Index Terms*— Gaussian regression, machine learning, Kalman filtering, spatio-temporal Gaussian processes.**

I. INTRODUCTION

Gaussian process-based (GP) regression [1] is a Bayesian learning framework where GP are used as nonparametric priors for regressors functions. Nowadays, GP based methods have heavily increased their popularity [2], [3] in disciplines such as statistical inference and machine learning [3]. In the classical machine learning setup the modeled process is considered static. Consequently, classical GP based regression, i.e., Kriging [4], often assumes as input variables just spatial locations. Nevertheless, the method can be extended to learn spatio-temporal processes by treating the time variable as an additional input feature [3]. In dynamical scenarios however, the classical GP based regression paradigm presents practical limitations. These are mainly due to the heavy memory and computational requirements which grow cubical as the number of input data. As second drawback, the classical paradigm is usually based on a batch implementation where the data are processed at once, after they have been collected. Conversely, concerning dynamical learning, Kalman filter [5] offers a computational efficient recursive procedure to learn dynamical processes. However, this approach requires a-priori knowledge of the process to learn.

This work is supported by Progetto di Ateneo CPDA147754/14-New statistical learning approach for multi-agents adaptive estimation and coverage control.

The authors are with the department of Information Engineering of the University of Padova, via Gradenigo 6/B 35100 Padova, Italy. author@dei.unipd.it author = [carronan|todescat|carlirug|schenato|giapi].

In the last decades, much effort has been put to cope with the computational complexity needed to implement GP regression methods. For instance, in [3], [6] sparse approximations are exploited. Differently, in [7], [8], the authors propose a finite memory implementation of the classical approach based on truncated observations. An alternative approach, based on the connection between GPs and Kalman filtering, is presented in [9], [10], [11], whose inception can be traced back to [1]. The works mainly focus on building equivalent state-space representations for gaussian processes. The models are then used to implement a Kalman filter. In [9] the authors present a preliminary result which applies to temporal GP regression models. In the more recent [10], [11], the authors extend the approach to spatio-temporal GPs. These are reformulated as infinite dimensional state space models to which Kalman Filtering can be applied.

This work, inspired by [10], [11], emphasizes the practical implementability of the estimating procedure in terms of computational complexity. Indeed, while in [10], [11], to deal with infinite dimenasional operators, only approximated inference schemes are proposed, e.g., based on eigenfunctions expansions of some operators which govern the stochastic dynamics, in this paper we develop an algorithm which is exact. In particular, it returns the exact minimum variance estimate and is computationally efficient. We confine our analysis to discrete finite-dimensional spaces. Moreover, we restrict to a specific yet sufficiently rich class of space-time separable kernel functions, which, as opposed to [10], [11], do not require stationarity of the space kernel. These assumptions paired with the additional assumptions of periodic sampling on a finite number of space locations, allow to provide a solution which can be exactly implemented without requiring any additional numerical approximation and whose complexity scales cubically in terms of the number of distinct measurements locations and scales only linearly on the number of prediction locations. In fact, the major finding is to show that the prediction locations do not need to appear in state-space representation of the Kalman filter. Differently, a naive implementation of a finite-buffer non-parametric estimator which uses only the most recent measurements, would have a complexity per iteration which grows cubically in terms of the total size of the buffer, which could be much larger than the number of distinct measurement locations, and which does not provide an exact solution.

II. PRELIMINARIES

In this section we briefly recall the required preliminaries on nonparametric estimation, Kalman filtering and spectral factorization.

A. Nonparametric Estimation

In this section we review some fundamental aspects regarding the nonparametric Gaussian regression.

Let $f : \mathscr{A} \mapsto \mathbb{R}$ be a zero-mean Gaussian field with covariance, also called kernel, $K : \mathscr{A} \times \mathscr{A} \mapsto \mathbb{R}$, where $\mathscr{A}$ is a compact set. Assume to have a set of $N \in \mathbb{N}_{>0}$ noisy measurements of the form

$$y_i = f(a_i) + v_i, \tag{1}$$

where v_i is a zero-mean Gaussian noise with variance σ^2, i.e. $v_i \sim \mathscr{N}(0, \sigma^2)$, independent from the unknown function. Given the data set of input locations $\{a_i, y_i\}_{i=1}^N$, it is known [12], [2] the estimate $\widehat{f}$ of f is a linear combination of the kernel sections $K(a_i, \cdot)$, i.e., the kernel sampled in the values corresponding to the available input locations. In particular, for any $a \in \mathscr{A}$, it holds that

$$\widehat{f}(a) \coloneqq \mathbb{E}\left[f(a) | \{a_i, y_i\}_{i=1}^N\right] = \sum_{i=1}^N c_i K(a_i, a). \tag{2}$$

The expansion coefficients c_i are obtained as

$$\begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} = (\bar{K} + \sigma^2 I)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \bar{K} \in \mathbb{R}^{N \times N}, \ [\bar{K}]_{ij} = K(a_i, a_j), \tag{3}$$

where I denotes the identity matrix of suitable dimension and $[\bar{K}]_{ij}$ denotes the ij-th entry of the matrix $\bar{K}$. Finally, the posterior variance of $\widehat{f}(a)$ evaluated at the generic location $a \in \mathscr{A}$ is given by

$$V(a) = \mathrm{Var}\left[f(a) | \{a_i, y_i\}_{i=1}^N\right] = K(a,a) - \begin{bmatrix} K(a_1,a) & \cdots & K(a_N,a) \end{bmatrix} (\bar{K} + \sigma^2 \mathbb{I})^{-1} \begin{bmatrix} K(a_1,a) \\ \vdots \\ K(a_N,a) \end{bmatrix}. \tag{4}$$

Clearly, because of the matrix inversion in both (3) and (4), the method scales as $\mathscr{O}(N^3)$. Moreover, in real-time applications, where a certain number of measurements are collected at each iteration, all the past measurements must be kept in memory. Thus, the method is more suitable for a batch, almost static implementation rather than an iterative time-varying one.

Remark 1 (Spatio-temporal processes): In the following we consider spatio-temporal processes. Conversely to classical Gaussian processes, where a usually denotes a spatial variable, in spatio-temporal processes a represents both time and space. Hence, without loss of generality, we can write $f(a) = f(x,t)$. Accordingly, the domain $\mathscr{A}$ can be decomposed as $\mathscr{A} \coloneqq \mathscr{X} \times \mathbb{R}_+$, with $\mathscr{X}$ and $\mathbb{R}_+$ denoting the spatial and temporal domain, respectively.

B. Kalman Filtering

In this section we briefly recall some basic notions on Kalman filtering for finite-dimensional discrete-time linear state-space dynamical systems [13].

Consider the following system

$$\begin{aligned} s_{k+1} &= A s_k + w_k, \\ y_k &= C_k s_k + v_k, \end{aligned} \tag{5}$$

where, at each iteration k, $s_k \in \mathbb{R}^n$ is the state vector, $y_k \in \mathbb{R}^m$ is the output vector, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are i.i.d. zero-mean Gaussian random vectors with covariance matrices $Q \geq 0$ and $R > 0$, respectively. $A \in \mathbb{R}^{n \times n}$ is the state matrix and $C_k \in \mathbb{R}^{m \times n}$ is the time-varying output matrix. As commonly done, we assume both the process and measurement noise to be uncorrelated with respect to each other, i.e. $\mathbb{E}\left[w_k^T v_s\right] = 0 \ \forall_{k,s}$. We also assume the initial condition s_0 is drawn from a Gaussian distribution with zero mean and covariance Σ_0, i.e., $s_0 \sim \mathscr{N}(0, \Sigma_0)$.

The Kalman Filter applied to the system (5) is described by the following recursive equations

$$\widehat{s}_{k+1|k} = A \widehat{s}_{k|k} \tag{6a}$$

$$\Sigma_{k+1|k} = A \Sigma_{k|k} A^T + Q \tag{6b}$$

$$\widehat{s}_{k+1|k+1} = \widehat{s}_{k+1|k} + L_{k+1}\left(y_{k+1} - C_k \widehat{s}_{k+1|k}\right) \tag{6c}$$

$$\Sigma_{k+1|k+1} = \Sigma_{k+1|k} - L_{k+1} C_k \Sigma_{k+1|k} \tag{6d}$$

$$L_{k+1} = \Sigma_{k+1|k} C_k^T \left(C_k \Sigma_{k+1|k} C_k^T + R\right)^{-1} \tag{6e}$$

where $\widehat{s}_{k|k}$ and $\Sigma_{k|k}$ represent the filtered estimate of the state and the posterior error covariance, respectively; $\widehat{s}_{k+1|k}$ and $\Sigma_{k+1|k}$ represent the (one step) predicted state estimate and error covariance, respectively; L_{k+1} is the Kalman gain; finally, the filter is initialized assuming $\widehat{s}_{0|-1} = \mathbb{E}[s_0] = 0$ and $\Sigma_{0|-1} = Cov[s_0] = \Sigma_0$.

We recall that, under the assumptions of normal distributed noises and perfect model knowledge, the Kalman filter is optimal, in mean square sense. Then, equations (6) return the minimum mean square error estimate of the state, which corresponds to

$$\widehat{s}_k = \mathbb{E}\left[s_k | y_0, \ldots, y_k\right],$$

that is, the estimate of the state given all the measurements up to the k-th one. Moreover, under the Markovianity (memoryless of the system) property of the state, it holds that

$$\mathbb{E}\left[s_k | y_0, \ldots, y_k\right] = \mathbb{E}\left[s_k | s_{k-1}, y_k\right],$$

that is, the previous state and the last measurement represent the sufficient statistic to compute the optimal estimate of the state at the current time instant.

Finally, it is well known, [14], that if we assume the output matrix constant, i.e. $C_k = C$, under the hypothesis of stabilizability of the pair (A, Q) and detectability of the pair (A, C) the estimation error covariance of the Kalman filter converges to a unique value from any initial condition.

C. Spectral factoriation of random processes

Here, we recall some notions about spectral factorization of random processes and realization theory. In particular we want to show how, specific class of processes admits an equivalent exact state-space representation.

Consider a stationary random process $f(t)$ with covariance $h(\tau)$. Thanks to the Wiener-Khinchin theorem, it is known that the power spectral density (PSD) of the process is equal to the Fourier transform of its covariance h, i.e.,

$$S(\omega) \coloneqq \mathscr{F}[h(\tau)](\omega).$$

Moreover, in the particular case when $S = S_r$ is rational of order $2r$, thanks to spectral factoriation [15], its PSDs can be rewritten as $S_r(\omega) = W(\mathbf{i}\omega)W(-\mathbf{i}\omega)$ with

$$W(\mathbf{i}\omega) = \frac{b_{r-1}(\mathbf{i}\omega)^{r-1} + b_{r-2}(\mathbf{i}\omega)^{r-2} + \cdots + b_0}{(\mathbf{i}\omega)^r + a_{r-1}(\mathbf{i}\omega)^{r-1} + \cdots + a_0}. \quad (7)$$

If necessary, to obtain the form (7) numerator and denominator coefficients of W are expanded and scaled. Finally, from realization thoery, we have that rational functions of the form (7) are in correspondance to the equivalent continuous time state space representation [16] (companion form) given by

$$\begin{cases} \dot{s}_t = Fs_t + Gw_t \\ z_t = Hs_t \end{cases} \quad (8)$$

where $w_t \sim \mathcal{N}(0, I)$, the model matrices are equal to

$$F = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{r-1} \end{bmatrix}, \quad G = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

$$H = \begin{bmatrix} b_0 & b_1 & b_2 & \dots & b_{r-1} \end{bmatrix},$$

and the initial state is $s_0 \sim \mathcal{N}(0, \Sigma_0)$, with Σ_0 computed as solution of the Lyapunov equation $FX + XF^T + GG^T = 0$.

III. MAIN CONTRIBUTION

Here, we present the main contributions of the paper. First, we state the problem at hand. Second, we formally show how to built an exact state space representation for a certain class of GPs. Then, we bridge GP regression and Kalman filtering, providing a clear and systematic methodology to implement the filter. As it will be clear, we first focus on estimating the process of interest over an "observable" finite collection of points. Finally, we show how to extend the estimation over an arbitrary "unobservable" finite collection of locations.

A. Problem Formulation

Consider a function $f : \mathcal{X} \times \mathbb{R}_+ \to \mathbb{R}$ modeled as a zero-mean Gaussian Process with covariance K. Hereafter, for the sake of notation ease, we use $f_t(x)$ instead of $f(x,t)$. We assume $\mathcal{X}$ to be a finite collection of points, i.e.,

$$\mathcal{X} := \left\{ x_1, \dots, x_N \,|\, x_i \in \mathbb{R}^d \right\}.$$

We assume noisy measurements of the form (1) come from a subset $\mathcal{X}_{\text{meas}} \subseteq \mathcal{X}$ of given locations. We formally define $\mathcal{X}_{\text{meas}}$ as follows.

Definition 2 (Measurements Space): Consider the finite set $\mathcal{X}$. We denote with $\mathcal{X}_{\text{meas}} \subseteq \mathcal{X}$ a finite collection of points containing[1] $M \leq N$ locations from $\mathcal{X}$, i.e.

$$\mathcal{X}_{\text{meas}} := \{ x_1, \dots, x_{\text{M}} \,|\, x_i \in \mathcal{X} \}.$$

[1]Observe that, by following the notation used in Definition 2, we have that the first M points of $\mathcal{X}$ represents the measurements locations. This holds without loss of generality assuming $\mathcal{X}$ has been a priori ordered.

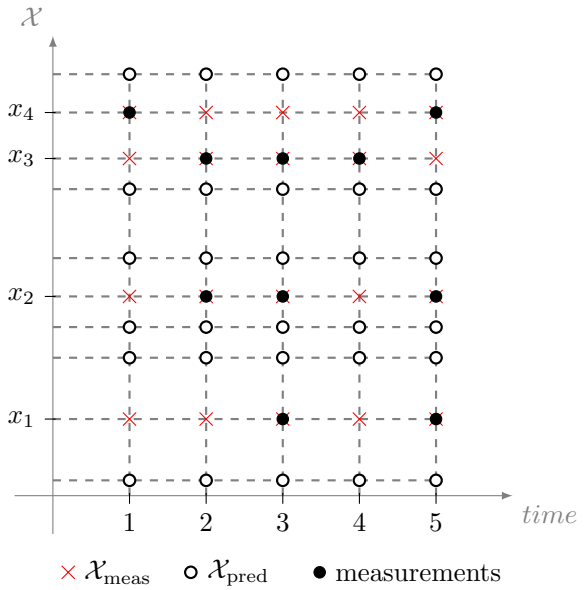


Fig. 1: Spatio-temporal sampling and measurements collection over time: the x-axis represents discrete time instants while the y-axis represents discrete space locations, i.e., $\mathcal{X}$. Red crosses highlight all the possible measurements locations contained in $\mathcal{X}_{\text{meas}}$. Black circles represent the locations $\mathcal{M}(k)$ where measurements are collected. Finally, white circles represent the prediction locations contained in $\mathcal{X}_{\text{pred}}$.

□

Precisely, to consider the most general case, we assume to be able to collect the measurements at discrete time instants $t = kT$, where T denotes the sampling time, only from a time-varying subset of locations, namely $\mathcal{M}(k) \subseteq \mathcal{X}_{\text{meas}}$ ($|\mathcal{M}(k)| = M_k$).

The problem we want to solve is that of estimating f over the entire "partially observable" domain $\mathcal{X}$, exploiting measurements coming from the "observable" set $\mathcal{X}_{\text{meas}}$. The problem could arise in diverse applications, e.g., in weather forecasting where, given a small finite set of weather stations which are able to collect measurements at certain discrete time instants, the goal is to estimate the weather conditions on a larger area.

To state our solution, we restrict the analysis on a specific yet sufficiently rich class of kernel functions.

Assumption 3 (Generating Kernel properties): The kernel function K, covariance of the Gaussian process $f_t(x)$, is separable in time and space and it is stationary in time, namely,

$$K(x, x', t, t') = K_s(x, x')h(\tau), \quad \tau = t' - t.$$

In addition, the power spectral density $S_r(\omega)$ of $h(\tau)$ is a rational function of order $2r$. □

Differently from [10], [11], in Assumption 3 we do not require space stationarity of K_s but only time stationarity. This let us consider a wider class of generating space kernels K_s, e.g., kernel spline.

Our solution consists of two steps: first we show how to estimate the process f_t over $\mathcal{X}_{\text{meas}}$ (Section III-B). Then,

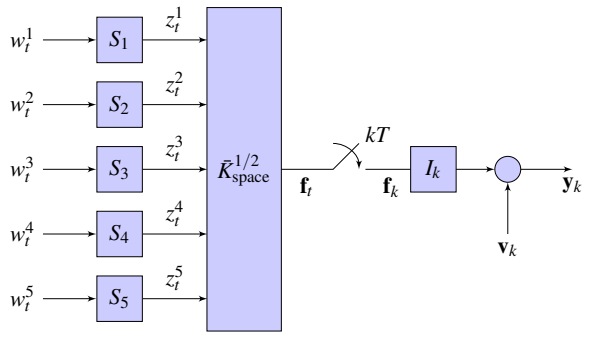


Fig. 2: Process and measurements formation: we assume there are five spatial locations, $x_j \in \mathscr{X}_{\text{meas}}$, each of them described by the state space system S_j of the form (9) driven by the noise w_t^j. The z_t^j's are then coupled trough the space kernel factor $K_{\text{s}}^{1/2}$ to form $\mathbf{f}_t$ which is sampled every T[s]. The matrix I_k "selects" the available locations at kT. Finally the measurements vector $\mathbf{y}_k$ is obtained adding measurements noise $\mathbf{v}_k$, in vector form.

we extend our result to obtain a prediction of the process outside $\mathscr{X}_{\text{meas}}$ (Section III-C). Precisely, we show how our first solution can be exploited to reconstruct f_t on the set $\mathscr{X}_{\text{pred}}$ where

$$\mathscr{X}_{\text{pred}} \coloneqq \mathscr{X} \,/\, \mathscr{X}_{\text{meas}}\,, \qquad (P \coloneqq |\mathscr{X}_{\text{pred}}| = N - M)\,.$$

Figure 1 shows an example of spatio-temporal sampling, as well as the measurements collection process over time.

B. Kalman Regression on $\mathscr{X}_{\text{meas}}$

To implement the Kalman equations (6), the first step is to build a state space representation for the Gaussian process f_t. In particular, we are interested in reconstructing f_t over the "observable" $\mathscr{X}_{\text{meas}}$. To compactly represent the process over $\mathscr{X}_{\text{meas}}$, it is convenient to define the vector

$$\mathbf{f}_t \coloneqq [f_t(x_1), \ldots, f_t(x_M)]^T\,,$$

The next proposition exploits Assumption 3 and the state-space realization for rational PSD given in (8) to show that the process $\mathbf{f}_t$, admits an equivalent exact continuous-time state-space representation.

Proposition 4 (Equivalent CT-SS representation for $\mathbf{f}_t$): Consider the process $\mathbf{f}_t : \mathscr{X}_{\text{meas}} \times \mathbb{R}_+ \mapsto \mathbb{R}^M$. Assume the generating kernel K satisfies Assumption 3. Let the triplet (F, G, H) be a state-space representation for $S_{\text{r}}(\omega)$ as described in II-C. Then, $\mathbf{f}_t$ admits the following strictly proper state-space representation

$$\begin{cases} S_j : \begin{cases} \dot{s}_t^j &= F s_t^j + G w_t^j \\ z_t^j &= H s_t^j \end{cases} & j \in \{1, \ldots, M\}\,, \\ \mathbf{f}_t = \bar{K}_{\text{s}}^{1/2} \mathbf{z}_t \end{cases} \tag{9}$$

where j is an index cycling through all the input locations of $\mathscr{X}_{\text{meas}}$; where $\mathbf{z}_t \coloneqq \left[z_t^1, \ldots, z_t^M\right]^T$ and $\bar{K}_{\text{s}} \in \mathbb{R}^{M \times M}$ is obtained sampling K_{s} over $\mathscr{X}_{\text{meas}}$; and where w_t^j and s_0^j are defined as for system (8).

Proof: First of all, notice that the process $\mathbf{f}_t$ is a Gaussian process since it is the solution of a linear differential equation driven by Gaussian noise w_t. To conclude the proof we need to show that the covariance of $\mathbf{f}_t$ is indeed $\bar{K} = \bar{K}_{\text{s}} h(\tau)$. As previously shown, the first two equations of model (9) are the state space representation of the rational power spectral density $S(\omega)$ thus $\mathbb{E}\left[z_{t+\tau}^j z_t^j\right] = h(\tau)$. It follows that

$$\mathbb{E}\left[\mathbf{f}_{t+\tau} \mathbf{f}_t\right] = \bar{K}_{\text{s}}^{1/2} \left[I\, h(\tau)\right] \left(\bar{K}_{\text{s}}^{1/2}\right)^T = \bar{K}_{\text{s}} h(\tau)\,.$$

■

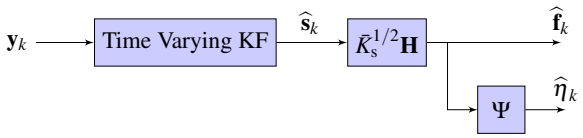


Fig. 3: Block-diagram of the estimation scheme. The block "Time Varying KF" implements Eqs. (6) applied to the system of Proposition 5. All other blocks are static: according to Eq. (12), $\bar{K}_{\text{s}}^{1/2}\mathbf{H}$ is needed to compute the estimate $\widehat{\mathbf{f}}_k$ over $\mathscr{X}_{\text{meas}}$. According to Eq. (16), Ψ is used to compute $\widehat{\eta}_k$ over $\mathscr{X}_{\text{pred}}$.

Observe that the subsystems S_j in (9) are independent one from each other in the sense that one can easily verify that $\mathbb{E}\left[(s_t^i)^T (s_t^j)\right] = 0 \;\; \forall t, \forall i \neq j$. Basically, Proposition 4 states that, for each location $x_j \in \mathscr{X}_{\text{meas}}$, the time evolution of f_t admits a state space representation given by the system S_j in equation (9). Then, these state space representations are "combined" through the sampled spatial kernel $\bar{K}_{\text{s}}$ to build a representation for the overall process $\mathbf{f}_t$. Observe that Proposition 4 gives a continuous-time state-space representation for the process. However, the Kalman equations (6) are defined in discrete time. Thus, in the following we show how to reconstruct an estimate $\widehat{\mathbf{f}}_k$ of $\mathbf{f}_t$ at discrete time instants, say $t = kT$ [2], defined as

$$\widehat{\mathbf{f}}_k \coloneqq \mathbb{E}\left[\mathbf{f}_{kT} \,|\, \{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0, \ldots, k\right]. \tag{10}$$

Finally, we recall from Section III-A that we assume to collect measurements coming from $\mathscr{M}(k)$ at $t = kT$, i.e.,

$$y_k^j = f_k(x_j) + v_k^j, \qquad x_j \in \mathscr{M}(k)\,,\; v_k^j \sim \mathscr{N}(0, \sigma^2)\,, \tag{11}$$

Then, given the state-space model of Proposition 4 and a measurement model (11), we have all the necessary elements to bridge GP regression and Kalman filtering on $\mathscr{X}_{\text{meas}}$.

Proposition 5 (Kalman regression on $\mathscr{X}_{\text{meas}}$): Assume Assumption 3 holds. Moreover, assume to collect periodic measurements of the form (11) at every $t = kT$. Then, the estimate $\widehat{\mathbf{f}}_k$ of $\mathbf{f}_k$ is given by

$$\widehat{\mathbf{f}}_k = \bar{K}_{\text{s}}^{1/2} \mathbf{H} \widehat{\mathbf{s}}_k\,, \tag{12}$$

where $\mathbf{H} \coloneqq blkdiag(H, \ldots, H)$, $\widehat{\mathbf{s}}_k$ is the output of the time-varying Kalman filter (6) applied to the discrete-time system (5) with matrices (A, C_k, Q, R) where $A \coloneqq blkdiag(\bar{F}, \ldots, \bar{F})$, $Q \coloneqq blkdiag(\bar{Q}, \ldots, \bar{Q})$, being $\bar{F}$ and $\bar{Q}$ defined as

$$\bar{F} = e^{FT}\,, \qquad \bar{Q} = \int_0^T \left(e^{F\tau}\right) G G^T \left(e^{F\tau}\right)^T d\tau\,,$$

[2] In the following, for brevity, we might drop the sampling time T from kT and use just k to denote the corresponding discrete time instant.

and where $R := \sigma I$ and $C_k := I_k \bar{K}_{\mathrm{s}}^{1/2}\mathbf{H}$, being $I_k \in \{0,1\}^{M_k \times M}$ the matrix selecting the locations contained in $\mathscr{M}(k)$. The Kalman filter is initialized as $\widehat{\mathbf{s}}_{0|-1} = 0$, $\Sigma_{0|-1} = blkdiag(\Sigma_0,\ldots,\Sigma_0)$, where Σ_0 is solution of the Lyapunov equation $FX + XF^T + GG^T = 0$. □

Proof: The proof directly follows from the discretization of the CT-SS models S_j of Proposition 4. Once the overall system discrete system, with space vector $\mathbf{s}_k := \left[(s_k^1)^T,\ldots,(s_k^M)^T\right]^T$, is rewritten in compact matrix form, the Kalman equations (6) straightforward apply. ■

Figure 2 shows a representation of the process and of the measurements formations. All the bold symbols refer to vector notation and are obtained stacking in vector form the corresponding non-bold symbols. Additionally, the upper part of Figure 3 shows a block diagram of the estimation scheme. Observe that the block "Time Varying KF", which implements the Kalman equations (6), is the only time-varying block. This is due to the time-varying measurements. Consequently, if the C_k matrix is constant, the Kalman gain converges to a constant value which can be computed offline. In this case the filtering correspond to a static matrix multiplication hence, the computational burden (see Section IV) is alleviated.

To conclude this section, we present an exhaustive example to help the reader's intuition in the building process of the presented estimation procedure.

Example 6: Consider the exponential time kernel $h(\tau)$

$$h(\tau) = \lambda e^{-\sigma_t|\tau|}$$

satisfying Assumption 3 since its PSD S_r is equal to

$$S_r(\omega) = \frac{\sqrt{2\lambda\sigma_t}}{(\sigma_t + \mathbf{i}\omega)}\frac{\sqrt{2\lambda\sigma_t}}{(\sigma_t - \mathbf{i}\omega)} \tag{13}$$

which is rational of order 2. Now, consider a zero-mean Gaussian process $f_t(x)$ with covariance

$$K(x,x',\tau) = K_{\mathrm{s}}(x,x')h(\tau) = e^{-\sigma_x(x_1-x_2)^2}\lambda e^{-\sigma_t|\tau|} \tag{14}$$

that is, a Gaussian spatial kernel and a exponential time kernel. Thanks to Proposition 4, since K satisfies Assumption 3, $\mathbf{f}_t$ admits a state space representation. In particular, given S_r as in (13) with

$$W(\mathbf{i}\omega) = \frac{\sqrt{2\lambda\sigma_t}}{(\sigma_t + \mathbf{i}\omega)},$$

it is easy to see the state-space model matrices are equal to

$$F = -\sigma_t, \qquad H = \sqrt{2\lambda\sigma_t}, \qquad G = 1,$$

while the matrix $\bar{K}_{\mathrm{s}}^{1/2}$ is computed as the Cholesky factorization of the sampled kernel $\bar{K}_{\mathrm{s}}$.
Finally, as stated in Proposition 5, the discrete time state-space representation for $\mathbf{f}_k$ is given by

$$\bar{F} = e^{-\sigma_t T}, \qquad H = \sqrt{2\lambda\sigma_t}, \qquad \bar{Q} = \int_0^T e^{-2\sigma_t\tau}d\tau.$$

To conclude the example we show one case when $h(\cdot)$ does not satisfy Assumption 3. Indeed, considering the squared exponential (Gaussian) kernel defined as

$$h(\tau) = \lambda e^{-\sigma_t^2\tau^2}.$$

it can be seen its power spectral density is not rational,

$$S(\omega) = \sqrt{\pi}\frac{\lambda}{\sigma_t}e^{-\left(\frac{\omega}{2\sigma_t}\right)^2}. \tag{15}$$

It is worth noticing that, in cases when the PSD S of h is not rational it is always possible to build a rational PSD $\widehat{S}_{\mathrm{r}}$ which approximate the true one. Different approximating methods can be used, e.g., Taylor series expansion or Pade approximation. This leads to an approximate state-space model for $\mathbf{f}_t$. In Section V we present some simulations and, as it will be explained, to approximate S we use a different approach. Indeed, the $\widehat{S}_{\mathrm{r}}$ is computed as the solution of a suitable non-linear weighted least-squares problem. □

C. Kalman Regression on $\mathscr{X}_{\mathrm{pred}}$

Here we extend the result of Proposition 5 to build an estimate of the process $f_k T$ over the "prediction" space $\mathscr{X}_{\mathrm{pred}}$ as defined at the end of Section III-A.

To this end, let

$$\eta_k := f_k(\mathscr{X}_{\mathrm{pred}}),$$

be the vector representing the process f_{kT} sampled over $\mathscr{X}_{\mathrm{pred}}$. Finally, we introduce the following symbols

$$\begin{aligned}
\widehat{\eta}_k &:= \mathbb{E}\left[\eta_k | \{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0,\ldots,k\right],\\
\Gamma &= Cov(\eta_k, \mathbf{f}_k) = \bar{K}_{\mathrm{s}}(\mathscr{X}_{\mathrm{pred}}, \mathscr{X}_{\mathrm{meas}}),\\
V_\eta &= Var(\eta_k) = \bar{K}_{\mathrm{s}}(\mathscr{X}_{\mathrm{pred}}, \mathscr{X}_{\mathrm{pred}}),
\end{aligned}$$

where $\bar{K}_{\mathrm{s}}(\cdot,\cdot)$ denotes the kernel K_{s} evaluated in all the locations contained in its arguments.

Proposition 7 (Kalman Regression on $\mathscr{X}_{\mathrm{pred}}$): Consider the process $f_t : \mathscr{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ generated by the kernel K satisfying Assumption 3. Then, the estimate $\widehat{\eta}_k$ of η_k is given by

$$\widehat{\eta}_k = \Psi\widehat{\mathbf{f}}_k, \tag{16}$$

where $\Psi := \Gamma\bar{K}_{\mathrm{s}}^{-1}$. The posterior variance is given by

$$Var\left(\eta_k | \{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0,\ldots,k\right) = V_\eta - \Gamma\left(\bar{K}_{\mathrm{s}}^{-1} - \bar{K}_{\mathrm{s}}^{-1}\left(\bar{K}_{\mathrm{s}}^{-1} + R^{-1}\right)^{-1}\bar{K}_{\mathrm{s}}^{-1}\right)\Gamma^T.$$

□

Proof: Since K satisfies Assumption 3 then, we have that $K(x,x',\tau) = K_{\mathrm{s}}(x,x')h(\tau)$, $\tau = (k-j)T$, and, without loss of generality, we can assume $h(0) = 1$. Then, it holds that $Var(\mathbf{f}_k) = \bar{K}_{\mathrm{s}}$. In the following we drop the sampling time T from the notation; moreover, we assume to be at time instant k, while $j < k$ represents a generic previous time instant.
Now, let be $\varphi_k := [\mathbf{f}_1^T,\ldots,\mathbf{f}_{k-1}^T]^T$. Moreover, for brevity instead of $h(k-j)$ we use the simpler h_{k-j}. Hence, it can be seen that $Cov(\mathbf{f}_j, \eta_k) = h_{k-j}\Gamma^T$.
We first study $p(\varphi_k, \eta_k | \mathbf{f}_k)$. For the conditional variance, we have that

$$\begin{aligned}
Var(\varphi_k, \eta_k | \mathbf{f}_k) = Var([\varphi_k^T \ \eta_k^T]^T) - \\
Cov([\varphi_k^T \ \eta_k^T]^T, \mathbf{f}_k) Var(\mathbf{f}_k)^{-1} Cov(\mathbf{f}_k^T, [\varphi_k^T \ \eta_k^T]^T)
\end{aligned} \tag{17}$$

where

$$Var([\varphi_k^T \;\; \eta_k^T]^T) = \begin{bmatrix} \bar{K}_{\mathrm{s}} & h_1\bar{K}_{\mathrm{s}} & h_2\bar{K}_{\mathrm{s}} & \cdots & h_{k-1}\Gamma^T \\ \bar{K}_{\mathrm{s}}^T h_1^T & \bar{K}_{\mathrm{s}} & h_1\bar{K}_{\mathrm{s}} & \cdots & h_{k-2}\Gamma^T \\ \vdots & & \ddots & & \vdots \\ & & & \bar{K}_{\mathrm{s}} & h_1\Gamma^T \\ \Gamma h_{k-1}^T & \Gamma h_{k-2}^T & \cdots & & V_\eta \end{bmatrix},$$

$$Cov(\mathbf{f}_k^T, [\varphi_k^T \;\; \eta_k^T]^T) = \begin{bmatrix} \bar{K}_{\mathrm{s}} h_{k-1} & \cdots & \bar{K}_{\mathrm{s}} h_1 & \Gamma^T \end{bmatrix}.$$

It is easy to see that the second term of the right hand side of (17) has the following structure

$$\begin{bmatrix} & & & h_{k-1}\Gamma^T \\ & * & & h_{k-2}\Gamma^T \\ & & & \vdots \\ & & & h_1\Gamma^T \\ \Gamma h_{k-1}^T & \Gamma h_{k-2}^T & \cdots & * \end{bmatrix}.$$

Hence, by subtracting it to the first term, i.e., $Var([\varphi_k^T \eta_k^T]^T)$, the last column and the last row cancel out (except for the diagonal block). This means that φ_k and η_k are conditionally independent given $\mathbf{f}_k$. Thus we have that

$$p(\varphi_k, \eta_k | \mathbf{f}_k) = p(\varphi_k|\mathbf{f}_k) p(\eta_k|\mathbf{f}_k). \tag{18}$$

Thank to this we can write

$$\begin{aligned} p(\eta_k|\varphi_k, \mathbf{f}_k) &\overset{Bayes}{\propto} p(\varphi_k, \eta_k|\mathbf{f}_k) p(\mathbf{f}_k) \\ &\overset{(18)}{=} p(\eta_k|\mathbf{f}_k) p(\varphi_k|\mathbf{f}_k) p(\mathbf{f}_k) \\ &\propto p(\eta_k|\mathbf{f}_k) p(\mathbf{f}_k) \propto p(\eta_k|\mathbf{f}_k), \end{aligned}$$

so η_k is independent from all the past $\mathbf{f}_j$ contained in φ_k. Then, we have that

$$\begin{aligned} &\mathbb{E}\left[\eta_k \;|\{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0, \ldots, k\right] = \\ &\quad = \mathbb{E}\left[\mathbb{E}[\eta_k|\varphi_k, \mathbf{f}_k] \,|\{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0, \ldots, k\right] \\ &\quad = \mathbb{E}\left[\mathbb{E}[\eta_k|\mathbf{f}_k] \,|\{x_j, y_\ell^j\}, x_j \in \mathscr{M}(\ell), \ell = 0, \ldots, k\right] \\ &\quad = \Gamma \bar{K}_{\mathrm{s}}^{-1} \widehat{\mathbf{f}}_k = \Psi \widehat{\mathbf{f}}_k, \end{aligned}$$

where the first equality holds because we are conditioning on a larger σ-algebra; the second holds thanks to (18) and the third comes from the definition (10) of $\widehat{\mathbf{f}}_k$. Finally, for the posterior variance we refer the reader [17], Appendix A - Lemma 1 combined with the conditional independence stated in Eq. (18). ■

The combination of Proposition 5 and Proposition 7 state that the output of the Kalman filter captures all the necessary information, contained in the measurements, to estimate the entire process. Indeed, $\widehat{\mathbf{f}}_k$ is a sufficient statistic to reconstruct f_{kT} over the entire domain $\mathscr{X}$. Figure 3 shows a block diagram of the overall estimation scheme.

IV. COMPUTATIONAL COMPLEXITY

Before presenting compelling numerical tests, we discuss some computational aspects.

As already mentioned, the computational burden per iteration for the standard nonparametric approach grows cubical with the total number of collected measurements. Interestingly, thanks to its recursive implementation, the proposed Kalman's computational complexity scales as $\mathscr{O}(rMM_k + rM_k^3)$, where M_k is the number of collected measurements per iteration and r is the order a single state space model S_j in (9). The first term in the cost is due to the matrix vector multiplication to compute the state update of Eq. (6c); while the second to the computation of the Kalman gain as in Eq. (6e). Additionally, it is worth noticing that to compute $\widehat{\eta}_k$, the matrix $\Psi \in \mathbb{R}^{P \times M}$ is fixed thus, in a real-time implementation, can be pre-computed offline. Therefore, to reconstruct the entire process only a matrix-vector multiplication, which costs $\mathscr{O}(MP)$, is needed. Thanks to this, our Kalman regression procedure is characterized by an overall computational complexity per iteration which scales as

$$\mathscr{O}(rMM_k + rM_k^3 + MP). \tag{19}$$

Conversely, the nonparametric approach needs to invert a matrix of increasing size at every iteration plus a matrix-vector multiplication to compute the estimate for the overall process, leading to a complexity of order

$$\mathscr{O}\left(\left(\sum_{\ell=1}^{k} M_\ell\right)^3 + P\sum_{\ell=1}^{k} M_\ell\right). \tag{20}$$

Thus, in a real-time implementation, the computational cost per iteration for the Kalman scales linearly with the model complexity r. Conversely the cost for the classical nonparametric implementation grows cubically with the total number of collected measurements.

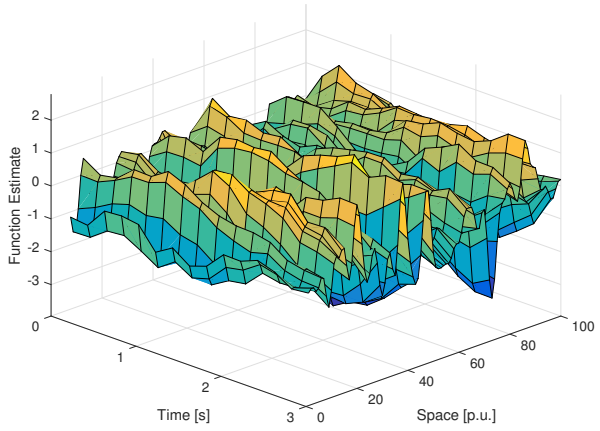
In the next section, we compare the proposed Kalman regression scheme with a modified version of the classical nonparametric implementation [7] based on a finite memory approach, which we refer to as *truncated nonparametric*. That is instead of storing in memory all the collected measurements up to the current iteration, only the measurements collected during the last q time instants are stored and processed.This is commonly done in practice in order to keep memory and computational requirements fixed. From (20), it is easy to see that the truncated nonparametric scales as

$$\mathscr{O}\left(\left(\sum_{\ell=k-q}^{k} M_\ell\right)^3 + P\sum_{\ell=k-q}^{k} M_\ell\right). \tag{21}$$

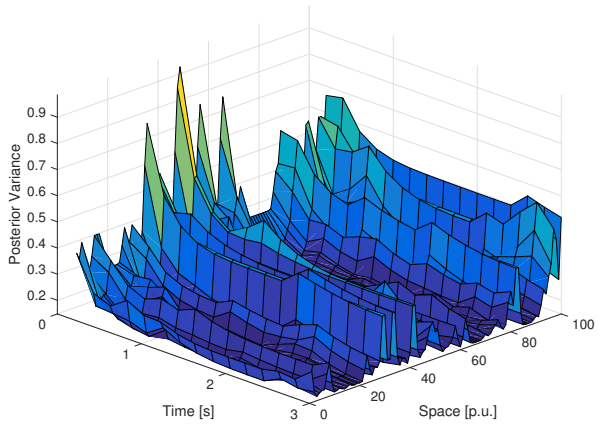
V. SIMULATIONS

In this section we present some simulations to show the effectiveness of the proposed Kalman regression.

The hardware test-bed consists of a 2,7 GHz Intel Core i5 processor with 16GB RAM running MATLAB® 2015. To plot the spatio-temporal evolution of the modeled function, we work on a 1D space. More specifically, $\mathscr{X}$ consists of a line of length 100 [p.u.], uniformly sampled every 1 [p.u.]

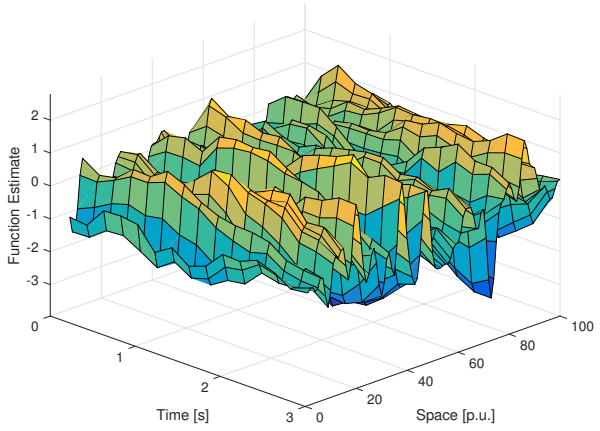


(a) Nonparametric estimate.

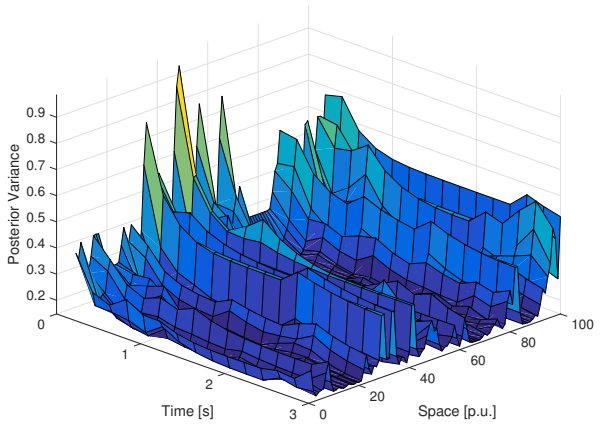


(b) Posterior variance.

Fig. 4: 3D representation of the optimal output obtained using the nonparametric estimation procedure.



(a) Kalman regression on $\mathscr{X}$.



(b) Posterior variance.

Fig. 5: 3D representation of the output of the proposed Kalman regression procedure. State-space model of order $r = 6$.

($|\mathscr{X}| = N = 100$). The sampling time is fixed and equal to 0.2 [s]. $\mathscr{X}_{\text{meas}}$ consists of $M = 80$ randomly selected locations. Finally, $\sigma = 1$ [p.u.]. To test the effectiveness of the proposed approach even on processes whose kernel does not satisfy Assumption 3, the selected process is drawn by a spatio-temporal Gaussian kernel K with

$$K_{\text{s}}(x,x') = e^{-0.2\|x-x'\|^2}, \quad h(\tau) = e^{-\|\tau\|^2/2}.$$

As mentioned at the end of Exampled 6, to approximate the non rational PSD $S(\omega)$ we compute $\widehat{S}_{\text{r}}(\omega)$ as the solution of a parametric non-linear weighted least-squares problem. More specifically, for a given order r we have that

$$\widehat{S}_{\text{r}}(\omega) = \underset{\{a_i\}_{i=0}^{r},\{b_i\}_{i=0}^{r-1}}{\operatorname{argmin}} \int_0^{\infty} \|S_{\text{r}}(w) - S(w)\|_{S(\omega)}\, dw.$$

where $\{a_i\}_{i=0}^{r}$ and $\{b_i\}_{i=0}^{r-1}$ are the coefficients of the spectral factor $W(\mathbf{i}\omega)$ of $S_{\text{r}}(\omega)$.

A. Estimation performance

First we compare Kalman with respect to the classical nonparametric method. At each time step k, we collect noisy measurements of the form (11) from M_k randomly selected locations within $\mathscr{X}_{\text{meas}}$, where M_k is randomly drawn from the uniform distribution over the set $[60, 80]$.

Figures 4 and 5 show the estimates and the corresponding posterior variance obtained using the nonparametric method and Kalman, respectively. The nonparametric approach, at every iteration k, uses all the measurements collected up to k. Observe how the output are almost exactly the same. The difference is due to the fact that Kalman is built on $\widehat{S}_{\text{r}}(\omega)$, with $r = 6$, instead of the true $S(\omega)$. Finally, notice that since the measurements locations change at every iteration, the posterior variance oscillates.

B. Computational performance

Here, we want to compare the proposed Kalman regression with the truncated nonparametric implementation described in Section IV. In the following we put $M = N$ ($\mathscr{X}_{\text{meas}} \equiv \mathscr{X}$), so $P = 0$ ($\mathscr{X}_{\text{pred}} = \emptyset$). Moreover we assume $M_k = M$, which is equivalent to collect measurements from all the locations. Thanks to this, the computational complexities per iteration (see Section IV) reduce to $\mathscr{O}(rM^3)$ for Kalman and to $\mathscr{O}(q^3M^3)$ for the truncated nonparametric, respectively. Therefore, r and q represent a measure for the complexity of the corresponding approach.
We compare the methods in term of CPU time per iteration

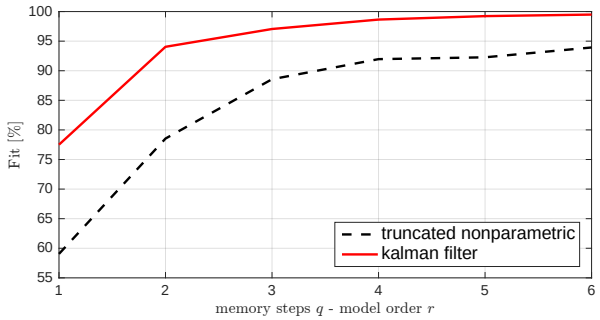


Fig. 6: Plot of the fit defined in (22). Kalman is plotted as function as function of the order r of the rational model used to approximate $S(\omega)$. The truncated nonparametric is plotted as function of the memory steps q.

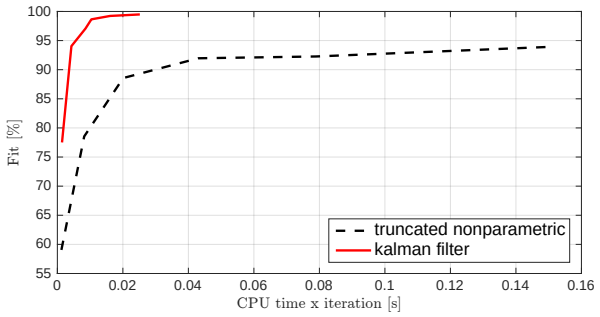


Fig. 7: Plot of the fit defined in (22) versus the CPU time per iteration required.

and in terms of estimation fit computed as

$$\text{Fit } [\%] = \left(1 - \frac{\|\widehat{\mathbf{f}}_* - \widehat{\mathbf{f}}_{\text{np}}\|}{\|\widehat{\mathbf{f}}_{\text{np}}\|}\right) 100\,, \tag{22}$$

where $\widehat{\mathbf{f}}_*$ denotes the estimate obtained either using Kalman or the truncated nonparametric; while $\widehat{\mathbf{f}}_{\text{np}}$ denotes the classical nonparametric estimate using all the available measurements. For the truncated nonparametric, Figure 6 shows the fit as function of the memory steps q. For Kalman, the fit is plotted as function of the model order r. It can be seen that, for the same level of complexity, Kalman in general achieves a better fit. We stress the fact that the performance in terms of fit for the truncated nonparametric highly depends on the ratio between the process and the measurements noise. Indeed, for high process noise, the information contained in the measurements collected during the last few iterations already contains all the necessary information to reconstruct the process. Thus, the fit curve would increase more rapidly. Conversely, Kalman is optimal hence it does not depend on the ratio. As final comparison, Figure 7 reports the fit versus the CPU time. The main fact to be highlighted is that to achieve a higher level of estimation accuracy, Kalman requires even less CPU time than the truncated nonparametric. It is important to underline that the CPU time deeply depend on M as well as on r and q. Indeed, for greater values of M Kalman could overwhelm the truncated approach even more. Conversely, if smaller value for M are chosen, the truncated nonparametric might be more efficient.

VI. CONCLUSIONS AND FUTURE WORKS

In this work we focused of the efficient estimation of time-varying spatio-temporal processes by combining GP nonparametric regression and Kalman filtering. We developed a computationally efficient and exact procedure for estimating the underlying process on a finite number of measurement and prediction locations via a finite dimensional state-space representation. The results are based on a specific separability assumption on the generating kernel for the modeled process, and we showed that the major computational bottleneck is given only by the number of distinct measurement locations, and not by the prediction locations. Future avenues of research regard the relaxation of the assumption on the generating kernel and the extension of the proposed approach to prediction over any desired point in time and space under non-periodic sampling.

REFERENCES

[1] A. O'Hagan and J. Kingman, "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–42, 1978.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American mathematical society*, vol. 39, pp. 1–49, 2001.

[3] C. K. Williams and C. E. Rasmussen, "Gaussian processes for machine learning," *the MIT Press*, vol. 2, no. 3, p. 4, 2006.

[4] N. Cressie, "The origins of kriging," *Mathematical geology*, vol. 22, no. 3, pp. 239–252, 1990.

[5] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[6] S. Oh, Y. Xu, and J. Choi, "Explorative navigation of mobile sensor networks using sparse gaussian processes," in *Decision and Control (CDC), 2010 49th IEEE Conference on*, Dec 2010, pp. 3851–3856.

[7] Y. Xu, J. Choi, and S. Oh, "Mobile sensor network navigation using gaussian processes with truncated observations," *Robotics, IEEE Transactions on*, vol. 27, no. 6, pp. 1118–1131, 2011.

[8] Y. Xu, J. Choi, S. Dass, and T. Maiti, "Sequential bayesian prediction and adaptive sampling algorithms for mobile sensor networks," *Automatic Control, IEEE Transactions on*, vol. 57, no. 8, pp. 2078–2084, 2012.

[9] J. Hartikainen and S. Särkkä, "Kalman filtering and smoothing solutions to temporal gaussian process regression models," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 379–384.

[10] S. Särkkä and J. Hartikainen, "Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 993–1001.

[11] S. Särkkä, A. Solin, and J. Hartikainen, "Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 51–61, 2013.

[12] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*. Washington, D.C.: Winston/Wiley, 1977.

[13] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[14] P. S. Maybeck, *Stochastic models, estimation and control. Volume I.*, A. Press, Ed., 1979.

[15] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT press Cambridge, MA, 1949, vol. 2.

[16] S. G. Mohinder and P. A. Angus, "Kalman filtering: theory and practice using matlab," *John Wileys and Sons*, 2001.

[17] M. Neve, G. D. Nicolao, and L. Marchesi, "Nonparametric identification of population models via gaussian processes," *Automatica*, vol. 43, no. 7, pp. 1134 – 1144, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109807001057