2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

Perspectives of Orthonormal Basis Functions Based Kernels in Bayesian System Identification*

Mohamed Darwish†, Gianluigi Pillonetto‡ and Roland Tóth†

Abstract— **Kernel-based regularization approaches for linear time-invariant system identification have been introduced recently. This class of methods corresponds to a particular regularized least-squares methodology that may achieve a favorable bias/variance trade-off compared with classical Prediction Error Minimization (PEM) methods. However, to fully exploit this property, the kernel function itself needs to be appropriately designed for the identification problem at hand to be able to successfully capture all relevant aspects of the data-generating system. Hence, there is a need for a methodology that can accomplish this design step without affecting the simplicity of these approaches. In this paper, we propose a systematic kernel construction mechanism to capture dynamic system behavior via the use of orthonormal basis functions (OBFs). Two special cases are investigated as an illustration of the construction mechanism, namely Laguerre and Kautz based kernel structures. Monte-Carlo simulations show that OBFs-based kernels with Laguerre basis perform well compared with stable spline/TC kernels, especially for slow systems with dominant poles close to the unit circle. Moreover, the capability of Kautz basis to model resonant systems is also shown.**

I. INTRODUCTION

Identification of Linear Time-Invariant (LTI) systems is a well-established field. The most commonly applied approaches fall into the category of Maximum Likelihood/Prediction Error Minimization ML/PEM methods [1], [2]. In ML/PEM, a parametric model structure is proposed and the model order, i.e., model complexity, is usually tuned via classical tools, e.g., AIC, BIC or Cross-Validation [1], [3]. However, classical methods for model order selection do not always give satisfactory results especially for short and noisy observations [4]. Alternatively, a novel kernel-based regularization approach has been introduced in [5] and further developed in [4], [6]. Although, the underlying concept of such approaches is much older see, e.g., [7], [8], in [5], the impulse response model structure is postulated and its estimation is dealt with as a function estimation problem. It has been shown that this approach corresponds to a particular regularized least-squares method that may achieve a favorable bias/variance trade-off compared to classical ML/PEM [5]. Moreover, it admits a Bayesian interpretation where the unknown impulse response is assumed to be a realization of a zero-mean Gaussian stochastic process with a certain covariance (Bayesian prior or kernel) function and the objective of the estimation problem is to estimate a particular realization of this Gaussian process which is most likely the true impulse response of the system. The assumed kernel function encodes the prior knowledge about the system under study and also restricts the high degree of freedom offered by the nonparametric estimation. The latter is important to define an estimator with a good bias/variance trade-off. Hence, the kernel function should express important qualitative properties of the dynamic system at hand like stability and/or oscillatory behavior. Hence, it becomes a question how to design a suitable kernel structure without knowing the true system. Such a structure is typically known up to some kernel parameters, called hyperparameters, which are needed to be estimated from data [9]. Hyperparameters estimation replaces traditional model order selection with the so-called *model complexity selection*. Such a "selection" in practice can be performed by the empirical Bayes approach, i.e., maximizing the marginal likelihood of these parameters with respect to the observed data [10], [11].

As for the kernel structure design, there are some structures available in the literature, for example, the stable spline (SS) kernels [5], diagonal/correlated (DC) kernels [6] and first order stable spline kernels known as tuned/correlated kernels (TC) [6], etc. The aim of the present work is to introduce a new way to construct kernels based on some classes of *orthonormal basis functions* (OBFs) with attractive properties in both system identification and series expansion representation of LTI systems [12]. These OBFs can be generated by a cascaded network of stable inner transfer functions, i.e., all-pass filters, completely determined, modulo the sign, by their poles [12]. In frequency-domain, OBFs provide basis for the Hilbert space $\mathcal{RH}_{2-}$ (space of strictly proper rational functions over $\mathbb{C}$ with real coefficients which functions are squared integrable on the unit circle and analytic outside of it). Moreover, in the time-domain, their correspondents, i.e., their impulse responses, provide basis for $\ell_2(\mathbb{N})$, i.e., the space of impulse responses of causal LTI systems. This means that, in the context of kernel-based regularization for impulse response estimation, a positive-definite kernel function can be expressed in terms of the impulse response of OBFs. This kernel is associated with a reproducing kernel Hilbert space (RKHS) that corresponds to the Hilbert space of impulse responses spanned by the OBFs associated impulse responses. There have already been few attempts to introduce OBFs-based kernels, e.g., [13]. However, the introduced OBFs-based kernels do not perform

*This research has benefited from the financial support of the Student Mission, Ministry of Higher Education, Government of Egypt and the Netherlands Organization for Scientific Research (NWO, grant no.: 639.021.127).

† Mohamed Darwish and Roland Tóth are with the Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, {m.a.h.darwish,r.toth}@tue.nl

‡ Gianluigi Pillonetto is with the Information Engineering Department, University of Padova, Padova 35131, Italy, giapi@dei.unipd.it

well compared with other kernels, i.e., the TC kernel, as shown in [13, Section V]. Moreover, the question of how many basis functions should be used to generate such kernels has not been answered yet. In this paper, we introduce the formulation of an OBFs-based kernel directly in the time-domain based on the impulse response formulation of these basis functions and introduce a decay term that weights the OBFs. In this way the difficult problem of selecting the number of basis functions to be introduced in the model is totally circumvented. Moreover, since the OBFs are uniquely determined in terms of the generating poles, these poles can be interpreted as hyperparameters. Hence, estimation of the poles can be performed in a Bayesian setting by maximizing the marginal likelihood with respect to the observed data. As an illustration of the construction mechanism, two special cases of OBFs-based kernel structures, i.e., Laguerre and Kautz basis [14], [15], are introduced and compared to other structures, i.e., the TC kernel.

This paper is organized as follows: the problem formulation of the finite impulse response estimation of LTI systems is presented in Section II, whereas Section III gives a brief overview on the regularized approaches to tackle this problem. The formulation of the proposed OBFs-based kernels in the time-domain is introduced in Section IV, followed by an extensive Monte-Carlo study in Section V. Conclusions then end the paper.

Notation: Consider the following definitions: let $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ be the interior of the unit disc in the complex plane, $\mathbb{J} = \{z \in \mathbb{C} \mid |z| = 1\}$ the unit circle and $\mathbb{F} = \{z \in \mathbb{C} \mid |z| > 1\}$ to represent the exterior of $\mathbb{D} \cup \mathbb{J}$, i.e., $\mathbb{F} = \mathbb{C} \setminus (\mathbb{D} \cup \mathbb{J})$. Let $\mathbb{N}$ denotes the natural numbers. If a is a matrix, then $|a|$ denotes the determinant of a and I_N denotes the N-dimensional identity matrix. Finally, let ℓ_1 and ℓ_2 denote the classical spaces of real sequences over $\mathbb{N}$ with summable absolute or squared values, respectively.

II. Problem Formulation

Consider a Single-Input-Single-Output (SISO), finite order, asymptotically stable and LTI discrete-time data generating system described by

$$y(t) = G_0(q)u(t) + v(t), \tag{1}$$

where $t \in \mathbb{Z}$ is the discrete-time, q is the forward time-shift operator, i.e., $qx(t) = x(t+1)$, $y : \mathbb{Z} \to \mathbb{R}$ is the output, $u : \mathbb{Z} \to \mathbb{R}$ is the input of the system, $v(t)$ is a zero-mean quasi-stationary noise process, independent of u, and

$$G_0(q) = \sum_{k=1}^{\infty} g_k^0 q^{-k}, \tag{2}$$

is the transfer operator of the system, where $g_0 = \{g_k^0\}_{k=1}^{\infty}$ is the unknown impulse response of the deterministic part of (1) represented by G_0. Since G_0 is assumed to be asymptotically stable, i.e., the sequence g_k^0 is absolute convergent, then, there exist a $n > 0$ such that $|g_k^0| \approx 0$ for $k > n$, where n is typically large. Hence, it is possible to consider

$$G(q, \theta) = \sum_{k=1}^{n} g_k q^{-k}, \quad \theta = [g_1 \cdots g_n]^\top, \quad \theta \in \mathbb{R}^n, \tag{3}$$

which is the n^{th} order Finite Impulse Response (FIR) model of $G_0(q)$. Given N data points, $\mathcal{D}_N = \{u(t), y(t)\}_{t=1}^{N}$ generated by (1), our goal is to estimate the n-truncated impulse response, i.e., the parameter vector θ, "as well as" possible. The corresponding identification criterion to achieve this objective will be defined later.

III. Regularization based methods

It is well-known that for high order FIR, PEM/ML gives an unbiased estimate at the price of high variance, due to the large number of parameters [1]. The standard way to cope with the variance increase is to introduce regularization, which can be understood from two equivalent point of views: the first is functional approximation in RKHS and the second is a Bayesian point of view. In the following, we will focus on the first perspective, where the second perspective will be occasionally highlighted to give further interpretation of the considered approach.

A. Regularization in RKHS

Let us first recall the definition of a positive definite kernel and its associated RKHS.

Definition 1: (Positive definite kernel) [16]. Let $(\mathcal{X}, d)$ be a metric space with d being its metric and $\mathcal{X} \subset \mathbb{R}^{\cdot}$. A real-valued function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel if it is continuous, symmetric and satisfies $\sum_{i,j=1}^{m} a_i a_j K(x_i, x_j) \geq 0$ for any finite set of points $\{x_1, \cdots, x_m\} \subset \mathcal{X}$ and $\{a_1, \cdots, a_m\} \subset \mathbb{R}$. □

Definition 2: (Reproducing kernel). Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of real-valued functions on $\mathcal{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A real-valued function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel for $\mathcal{H}$ if and only if

1) $\forall x \in \mathcal{X}, K_x = K(x, \cdot) \in \mathcal{H}$, is the kernel section centered at x.
2) The reproducing property holds, i.e., $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}, \quad \forall x \in \mathcal{X}, \forall f \in \mathcal{H}$. □

A Hilbert space of real-valued functions which possesses a reproducing kernel is called a RKHS and defined as the closure of span $\{K_x := K(x, \cdot) : x \in \mathcal{X}\}$, i.e., the functions in $\mathcal{H}$ can be written as

$$\mathcal{H} = \left\{ f : \mathcal{X} \to \mathbb{R} \mid f(X) = \sum_{k=1}^{n} a_i K_{x_i}(x), n \in \mathbb{N}, \right. \\ \left. x_i \in \mathcal{X}, a_i \in \mathbb{R}, \|f\|_{\mathcal{H}} < +\infty \right\},$$

where $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ is the norm in $\mathcal{H}$ induced by the inner product of $\mathcal{H}$ which is defined as

$$\langle g, h \rangle_{\mathcal{H}} = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i b_j K(x_i, x_j),$$

for

$$g = \sum_{i=1}^{m} a_i K_{x_i}, \quad h = \sum_{j=1}^{m} b_j K_{x_j}.$$

Moreover, due to the Moore-Aronszajn Theorem [17], there is a one-to-one correspondence between RKHS $\mathcal{H}$ and its reproducing kernel K, i.e., to every positive definite kernel K, there corresponds a unique RKHS $\mathcal{H}$ with K as its reproducing kernel and vice versa. Hence, in the sequel, we shall

denote the RKHS associated with the kernel function K as $\mathcal{H}_K$ and its inner product as $\langle \cdot, \cdot \rangle_K$ with the associated norm $\| \cdot \|_K$. Now, we are in a position to illustrate the problem of estimating the impulse response as a function estimation problem in RKHS. Equivalently, (1) can be rewritten as

$$y(t) = (g_0 * u)(t) + v(t), \tag{4}$$

where, $(g_0 * u)(t)$ denotes the time convolution between the impulse response g_0 and the input u. One approach of estimating the impulse response g_0 from noisy measurements is to regard it as an element of a $\mathcal{H}_K$ associated with a real causal kernel K, i.e., $K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, and then minimizing a regularized functional with respect to $\mathcal{H}_K$ [5], [4]

$$\min_{g_0 \in \mathcal{H}_K} \sum_{t=1}^{N} (y(t) - (g_0 * u)(t))^2 + \lambda \|g_0\|_K^2 \tag{5}$$

where $\lambda \geq 0$ is the regularization parameter and $\| \cdot \|_K^2$ is the induced norm of $\mathcal{H}_K$. It is worth to mention that the cost function in (5) consists of two terms. The first term is the quadratic loss accounting for the adherence to the observations. The second term, i.e., $\|g_0\|_K^2$, control the model complexity render the problem well-posed by including information about g_0, e.g., smoothness and/or stability. By considering the Output-Error (OE) noise model with (3), i.e., the n-truncated impulse response, (5) becomes equivalent to the following regularized least-squares problem [18]

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta} \|Y_N - \Phi\theta\|_2^2 + \lambda \theta^\top \mathbf{K}(\beta)^{-1}\theta && \text{(6a)} \\ &= (\mathbf{K}(\beta)\Phi_N^\top \Phi_N + \lambda I_N)^{-1}\mathbf{K}(\beta)\Phi_N^\top Y_N, && \text{(6b)} \end{aligned}$$

where $\| \cdot \|_2$ denotes the Euclidean norm, $\mathbf{K}(\beta)$ is an $n \times n$ kernel matrix, which is defined as $[\mathbf{K}]_{ij} = K(i, j)$, the parameter vector β contains the hyperparameters that should be tuned, $Y_N = [y(1) \cdots y(N)]^\top$, $\Phi = \left[\phi^\top(1) \dots \phi^\top(N)\right]^\top$ and $\phi(i) = [u(i-1) \cdots u(i-n)]$.

There is an important issue regarding this estimation approach: the choice of the kernel function K. The design of the structure of K is concerned with choosing a parameterised form of K with some hyperparameters β which can express a wide variety of impulse responses, but at the same time restricts the high degree of freedom by encoding expected dynamical properties like stability, oscillatory behaviour, etc. Furthermore, it is important that the associated restrictions are sensitive to the choice of β, i.e., β can be efficiently used to decrease the RKHS associated with K towards a set capturing the dynamical properties of g_0, but at the same time β is low dimensional. The resulting hyperparameters β can be tuned by using an empirical Bayes approach in terms of marginal likelihood maximization [19]. More specifically, the considered approach admits a Bayesian interpretation, where the impulse response is modeled as a zero-mean Gaussian process [11] with a covariance $\mathbf{K}(\beta)$, i.e., $\theta \sim \mathcal{N}(0, \mathbf{K}(\beta))$, and independent of the disturbance $v(t)$ which is assumed to be white Gaussian with mean 0 and variance σ^2, i.e., $v(t) \sim \mathcal{N}(0, \sigma^2)$. As a result, θ and Y_N are jointly Gaussian distributed and hence the posterior distribution of θ given Y_N is also Gaussian and its maximum a posterior (MAP) estimate is given as (6b). This interpretation provides an efficient way to estimate the hyperparameters from data following the empirical Bayes approach as follows

$$\begin{aligned} \hat{\beta} &= \arg\max_{\beta} \quad p(Y_N|\beta) \\ &= \arg\min_{\beta} Y_N^\top (\Phi\mathbf{K}(\beta)\Phi^\top + \sigma^2 I_N)^{-1} Y_N \\ &+ \log|\Phi\mathbf{K}(\beta)\Phi^\top + \sigma^2 I_N|. \end{aligned} \tag{7}$$

It is shown in [20] that this approach of tuning the hyperparameters can achieve a well balanced trade-off between the data fit and the high degree of freedom offered by the nonparametric estimators. Moreover, an efficient and accurate way to implement the hyperparameters estimation problem can be found in [21].

B. Kernel structures for impulse response estimation

As mentioned before, the importance of the kernel structure design step comes from the fact that the properties of the kernel function are reflected directly to the associated RKHS that is used as a hypothesis space for the estimation. For impulse response estimation, the kernel function K should reflect what is reasonable to assume about the impulse response, e.g., the exponential stability, smoothness and/or oscillatory response. Hence, it is useful to recall from [6] that the optimal kernel for the estimation problem (5) can be expressed as:

$$K(i, j) = g_i^0 g_j^0. \tag{8}$$

Even if (8) is not possible to be used in practice since the true impulse response is unknown, it provides a guideline to design a suitable kernel function for regularized identification. Inspired by the machine learning literature, many kernel structures have been introduced, e.g., stable spline (SS) kernel [5], diagonal/correlated (DC) [6], tuned/correlated or first order stable spline kernel [6]:

$$\begin{aligned} &\text{DC} \quad K_{\mathrm{i,j}}^{\mathrm{DC}}(\beta) = c\rho^{|i-j|}\lambda^{\frac{i+j}{2}}, \quad \beta = [c\ \rho\ \lambda]^\top \\ &\text{TC} \quad K_{\mathrm{i,j}}^{\mathrm{TC}}(\beta) = c\min(\lambda^i, \lambda^j), \quad \beta = [c\ \lambda]^\top \\ &\text{SS} \quad K_{\mathrm{i,j}}^{\mathrm{SS}}(\beta) = \begin{cases} c\frac{\lambda^{2\mathrm{i}}}{2}(\lambda^{\mathrm{j}} - \frac{\lambda^{\mathrm{i}}}{3}), & i \geq j \\ c\frac{\lambda^{2\mathrm{j}}}{2}(\lambda^{\mathrm{i}} - \frac{\lambda^{\mathrm{j}}}{3}), & i \geq j \end{cases}, \quad \beta = [c\ \lambda]^\top \end{aligned}$$

However, besides of exponential decay of the hypothesised impulse responses, none of these kernels can express other dynamical aspects of impulse response. In the next section we will propose an advanced kernel structure that is capable of expressing these dynamical aspects.

IV. Orthonormal basis functions based kernels

A. Orthonormal basis viewpoint for kernels

Mercer's theorem [22] allows us, under certain conditions, to represent the kernel function and thus any function in $\mathcal{H}_K$ as orthonormal basis in terms of eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ as follows

$$K(i, j) = \sum_{k=1}^{\infty} \lambda_k \phi_k(i)\phi_k(j), \tag{9}$$

where $i, j \in \mathbb{N}$ and the eigenfunctions constitute a basis of $\mathcal{H}_K$ and, as a result, the RKHS space $\mathcal{H}_K$ can be equivalently

defined as linear combinations of the orthonormal basis $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ as follows [11]:

$$\mathcal{H}_K=\left\{f:X\to\mathbb{R}\mid f(x)=\sum_{i=1}^{\infty}a_i\phi_i(x)\ \&\sum_{i=1}^{\infty}\frac{a_i^2}{\lambda_i}<+\infty\right\}$$

This means that any function $f\in\mathcal{H}_K$ can be represented as a linear combination of the orthonormal basis generated by the kernel function K. Next, using LTI system theory, we will show how a proper selection of such orthonormal basis of K can be chosen to define a corresponding RKHS that is suitable for impulse response estimation.

B. OBFs: an overview

The idea is to introduce OBFs [23], [12] to achieve a theoretically sound construction of kernels for impulse responses rather than the choices mentioned in Section III-B. Since OBFs are mainly defined in the frequency-domain-they are generated by inner transfer functions, therefore, in the sequel we will introduce OBFs in the frequency-domain and then define their correspondent basis for impulse responses in time-domain.

Let $\Psi=\{\psi_\tau(z)\}_{\tau=1}^{\infty}$ be a complete basis in $\mathcal{RH}_{2-}(\mathbb{F})$, the Hardy space of strictly proper rational complex functions with real coefficients that are square integrable on $\mathbb{J}$ and analytic in $\mathbb{F}$, with the inner product defined as

$$\langle F_1,F_2\rangle_{\mathcal{RH}_{2-}(\mathbb{F})}=\frac{1}{2\pi i}\oint_{\mathbb{J}}\overline{F_1}(1/\overline{z})F_2(z)\frac{dz}{z}$$

for any $F_1,F_2\in\mathcal{RH}_{2-}(\mathbb{F})$, where $\overline{(\cdot)}$ denotes complex conjugation. For this space, the general OBFs, i.e., Takenaka-Malmquist basis [12] is defined as

$$\psi_\tau(z)=\frac{\sqrt{1-|\xi_\tau|^2}}{z-\xi_\tau}\prod_{i=1}^{\tau-1}\frac{1-\overline{\xi_i}z}{z-\xi_i},\tag{10}$$

with $\{\xi_k\}_{k=1}^{\infty}\subset\mathbb{D}$ being the generating pole locations of Ψ satisfying $\sum_{k=1}^{\infty}(1-|\xi_k|)=\infty$. As $\{\psi_\tau\}_{\tau=1}^{\infty}$ is a complete basis for $\mathcal{RH}_{2-}(\mathbb{F})$, it holds that for all transfer function $F_0\in\mathcal{RH}_{2-}(\mathbb{F})$, there exists a unique sequence of expansion coefficients $\gamma_\tau\in\mathbb{R}$ such that $F_0(z)=\sum_{\tau=1}^{\infty}\gamma_\tau\psi_\tau(z)$. Two interesting special cases of the general Takenaka-Malmquist basis, which have been proven to be useful in system identification, are the Laguerre and the Kautz basis [12]. Laguerre basis in $\mathcal{RH}_{2-}(\mathbb{F})$ are defined as

$$\psi_\tau(z)=\frac{\sqrt{1-\xi^2}}{z-\xi}\left(\frac{1-\xi z}{z-\xi}\right)^{\tau-1},\qquad\xi\in(-1,1)\tag{11}$$

where the parameter ξ is known as the Laguerre parameter or generating real pole. The impulse response of Laguerre basis functions exhibit an exponential decay, however, Laguerre functions can not represent oscillatory behaviour of impulse responses, i.e., complex poles. Therefore, two-parameter Kautz basis functions result in more appropriate structure for this purpose:

$$\begin{aligned}\psi_{2\tau-1}&=\frac{\sqrt{1-c^2}(z-b)}{z^2+b(c-1)z-c}\left(\frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}\right)^{\tau-1}\\ \psi_{2\tau}&=\frac{\sqrt{(1-c^2)(1-b^2)}}{z^2+b(c-1)z-c}\left(\frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}\right)^{\tau-1}\end{aligned}\tag{12}$$

where $b,c\in(-1,1)$. Note that (12) corresponds to a repeated complex pole pair $\xi,\bar{\xi}\in\mathbb{D}$. Since we are interested in impulse response estimation, it is more convenient to define the corresponding OBFs in time-domain. Denote $\{\phi_\tau\}_{\tau=1}^{\infty}$ be the correspondent of $\{\psi_\tau\}_{\tau=1}^{\infty}$ in time-domain, i.e., $\phi_\tau(t)=\mathcal{Z}^{-1}\{\psi_\tau(z)\}$, where $\mathcal{Z}^{-1}\{\cdot\}$ is the inverse z-transform on the appropriate region of convergence, i.e., $\mathbb{F}$. It is an important result that $\{\phi_\tau\}_{\tau=1}^{\infty}$ is a complete basis of ℓ_2 and any impulse response g_0 associated with $F_0\in\mathcal{RH}_{2-}(\mathbb{F})$ can be written as $g_0=\sum_{\tau=1}^{\infty}\gamma_\tau\phi_\tau$. Note also that the expansion coefficients γ_τ decay to zero and based on the choice of the basis functions it can have a rapid convergence rate. We would like to use this property to build appropriate kernels for our Bayesian identification problem.

C. OBFs-based kernels

A fundamental result on RKHS is

Proposition 3: ([17]). Let $\mathcal{H}$ be a separable Hilbert space of functions over X with orthonormal basis $\{\varphi_j(\cdot)\}_{j=1}^{\infty}$. Then,

$$\mathcal{H}\text{ is a RKHS}\iff\sum_{j=1}^{\infty}|\varphi_j(x)|^2<\infty,\ \forall x\in X.$$

The unique kernel K that is associated with $\mathcal{H}$ is

$$K(x,y)=\textstyle\sum_{j=1}^{\infty}\varphi_j(x)\varphi_j(y).$$ □

Consider ℓ_2 and its canonical orthonormal basis given by the sequences e_i with all null elements except 1 in the i^{th} component. Using the above given result, it is immediate to conclude that ℓ_2 is an RKHS with kernel given by the infinite-dimensional identity matrix, i.e. $K(i,j)=\delta_{ij}$ where δ_{ij} is the Kronecker delta. Now, the simplest kernel that can be built using Laguerre or Kautz basis functions is

$$K(i,j)=\sum_{k=1}^{\infty}\phi_k(i)\phi_k(j).\tag{13}$$

In the sequel, the suitability of the kernel defined by OBFs (13) for impulse response estimation is assessed. Since Laguerre or Kautz basis correspond to orthonormal basis in ℓ_2, from Proposition 3, and in particular from the unicity of the kernel, it comes that $K(i,j)=\delta_{ij}$. If the system to be identified is stable, the kernel (13) will perform poorly: in fact, the optimal structure (8) suggests that the kernel diagonal elements should decay to zero, instead of being constant. In addition, the off diagonal elements should be different from zero. Also the Bayesian interpretation of regularization, as described, e.g., in [24, Subsection 4.3], helps in understanding the limitations of the kernel (13). The estimator (5) can in fact be seen as the minimum variance estimator of the impulse response when the latter is a zero-mean Gaussian process, independent of the noise, with covariance proportional to K. When (13) is adopted, g becomes proportional to a stationary white noise. But the stable impulse response is expected to decay to zero as time progresses, hence, it should be represented by a noise process with variance decaying to zero. Coming back to our RKHS perspective, the above mentioned problem related to (13) can be expressed as a kernel which is not stable according to the

following definition (which extends to the discrete-time case the one contained in [24, Section 13]).

Definition 4: Let $\mathcal{H}_K$ be the RKHS of functions on $\mathbb{N}$ induced by a kernel K. Then, K is said to be stable (from the impulse response point of view) if $\mathcal{H}_K \subset \ell_1$. □

The following proposition provides a sufficient condition for a kernel to be stable. The proof is omitted since it is derived from the results contained in [25] following the same arguments contained in [24, Section 13]. In particular, one can first think of the function domain as $\mathbb{N}$ equipped with a counting measure. Then, the rationale in [24, Section 13] holds replacing integrals with infinite sums.

Proposition 5: Let $\mathcal{H}_K$ be the RKHS on $\mathbb{N}$ induced by K. Then,

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|K(i,j)| < \infty \implies \mathcal{H}_K \subset \ell_1. \tag{14}$$

□

In view of the above results, to include the stability constraint for (13), the approach proposed in this paper is to consider the following kernel construction

$$K(i,j) = \lambda_s \sum_{\tau=1}^{\infty} q_\tau(\alpha)\phi_\tau(i)\phi_\tau(j), \tag{15}$$

where $|q_\tau(\alpha)| \to 0$ as $\tau \to \infty$. Possible choices of q_τ are

$$q_\tau(\alpha) = \tau^{-\alpha}, \quad \alpha > 0$$

or

$$q_\tau(\alpha) = \alpha^\tau, \quad 0 \le \alpha < 1,$$

so that α becomes a hyperparameter that determines the decay rate of the expansion (15). Note that the other kernel hyperparameters are the scale factor λ_s and the poles used to generate the sequence $\phi_\tau(\cdot)$. It is worth to mention that (15) enables to use a large set of basis to generate K as $q_\tau(\alpha)$ acts as a weighting to specify implicity the number of significant basis. With the marginal likelihood optimization, $q_\tau(\alpha)$ acts as an automatic selection of the number of basis that specify K. The following proposition provides information on the stability of the kernel built using Laguerre functions.

Proposition 6: Consider the kernel (15) built using the Laguerre basis functions. Then, the kernel is stable if $q_\tau(\alpha) = \tau^{-\alpha}$ and $\alpha > 3$ or $q_\tau(\alpha) = \alpha^\tau$ and $0 \le \alpha < 1$.

Proof: The proof is a simple application of Proposition 5 and of the following inequality taken from [14]:

$$\|\phi_\tau\|_1 \le \tau A,$$

where $\|\cdot\|_1$ indicates the norm in ℓ_1 while $A \in \mathbb{R}$ is a constant that depends only on the poles of the OBFs. Then,

$$\begin{aligned}
&\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\sum_{\tau=1}^{\infty} q_\tau(\alpha)\phi_\tau(i)\phi_\tau(j)| \\
&\quad\le \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{\tau=1}^{\infty} q_\tau(\alpha)|\phi_\tau(i)||\phi_\tau(j)| \\
&\quad= \sum_{\tau=1}^{\infty} q_\tau(\alpha)\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\phi_\tau(i)||\phi_\tau(j)| \\
&= \sum_{\tau=1}^{\infty} q_\tau(\alpha)\sum_{i=1}^{\infty}|\phi_\tau(i)|\sum_{j=1}^{\infty}|\phi_\tau(j)| \le A^2\sum_{\tau=1}^{\infty}\tau^2 q_\tau(\alpha)
\end{aligned}$$

■

In the sequel, we will limit our attention to the case $q_\tau(\alpha) = \tau^{-\alpha}$, leaving other investigations to future work.

D. Hyperparameters estimation

In case of OBFs-based kernel defined in (15), the hyperparameters that need to be estimated from data are the scaling parameter λ_s, decay parameter α and the generating poles. Note that in case of Laguerre-based (LOBF) kernel, only one real pole, i.e., ξ, is needed to generate the full sequence of basis and for Kautz-based (KOBF) kernel, two conjugate complex poles defined by b and c in (12), are needed to generate that sequence. Hence, the estimation of these hyperparameters following the empirical Bayes approach can be accomplished by solving the optimization (7).

V. Numerical Simulation

A. Simulation studies

To test the proposed OBFs-based kernels in the considered Bayesian identification settings, five simulation studies are accomplished for the following scenarios:

1) `S1D1`: fast systems, data sets with $N = 500$, SNR=10.
2) `S1D2`: fast systems, data sets with $N = 375$, SNR=1.
3) `S2D1`: slow systems, data sets with $N = 500$, SNR=10.
4) `S2D2`: slow systems, data sets with $N = 375$, SNR=1.
5) `S3`: oscillatory systems, data sets with $N = 400$, SNR=10.

Each scenario 1) to 4) corresponds to 100 randomly generated 30^{th} order discrete-time systems which are generated by the `drss` Matlab function. The fast systems have all poles inside the circle centered at the origin and radius 0.95 and the slow systems have at least one pole outside this circle but inside $\mathbb{D}$, i.e., slow dominant poles. These systems are used to generate data sets for a white u, with $u \sim \mathcal{N}(0,1)$ and v being additive white Gaussian noise. The variance of v is set such that the *signal-to-noise* ratio (SNR), i.e.,

$$10\log_{10}\left(\frac{\sum_{k=1}^{N}\tilde{y}^2(k)}{\sum_{k=1}^{N}v^2(k)}\right)$$

where $\tilde{y}(k)$ denote the noise-free system output. Whereas, scenario 5) generated as reported in [26], but with only one dominant complex conjugate pole pair.

B. Identification setting

In all of the five scenarios, we estimate FIR models (3) with $n = 125$ and with three different stable kernels:

1) First order stable spline denoted by TC
2) Laguerre-based kernel (LOBF)
3) Kautz-based kernel (KOBF)

The performance index that is used to measure the quality of the impulse response estimation with different estimators is the best fit rate index

$$\text{BFR}=100\left[1-\left(\frac{\sum_{k=1}^{125}|g_k^0-\hat{g}_k|^2}{\sum_{k=1}^{125}|g_k^0-\bar{g}^0|^2}\right)^{\frac{1}{2}}\right], \bar{g}^0 = \frac{1}{125}\sum_{k=1}^{125}g_k^0, \tag{16}$$

TABLE I: Average of the BFR of each kernel

	TC	LOBF
S1D1	89.43	**90.75**
S1D2	74.27	**75.10**
S2D1	82.96	**86.47**
S2D2	60.48	**65.72**

where, $\hat{g}_k$ are the estimated impulse response coefficients and g_k^0 are the true coefficients values. The hyperparameters have been estimated by the discussed marginal likelihood maximization, i.e., (7).

C. Identification results

The average model fits over the first four data sets are reported in Table I, where the largest average model fit is written in bold. For illustration, the distribution of the model fits over the data sets S2D1, S2D2 and S3 is shown by boxplots in Fig. 1.

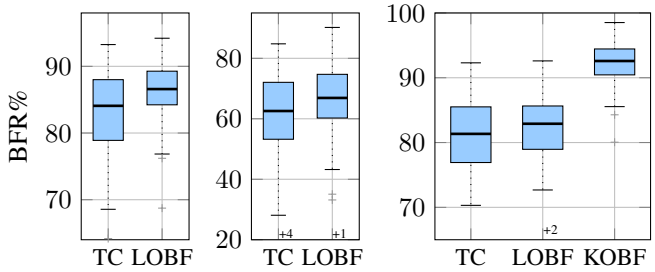
Based on Fig. 1, it is obvious that the proposed OBFs-based kernels perform well compared to other kernel structures, i.e., the TC kernel. This is due to the fact that the kernel is built with OBFs which are directly linked to the dynamical system behavior compared to existing kernel structures. Moreover, the LOBF kernel is capable of capturing the dynamics of the systems with a dominant real pole, whereas KOBF has the advantage over other kernels to deal with oscillatory systems.

VI. Conclusion and Future work

In this work, we have introduced a novel idea of using OBFs to build a RKHS in the time-domain and use it as a hypothesis space for impulse response estimation. Different weights on different OBFs are imposed by a decaying term. In this way, hyperparameters estimation replaces the difficulty of selecting the number of basis functions that should be introduced in the kernel. Two special cases are shown, the LOBF and the KOBF kernel structure. The performance of both of them is evaluated and compared with the TC kernel by means of Monte-Carlo simulations. Results show that the LOBF kernel performs well compared with the TC kernel especially for slow systems. Moreover, KOBF performs significantly better on resonant systems compared with TC and LOBF.

VII. Acknowledgments

The authors are grateful to Dr. Tianshi Chen for valuable discussions on this topic.



(a) S2D1 (b) S2D2 (c) S3

Fig. 1: Boxplot for model fits over S2D1, S2D2 and S3.

References

[1] L. Ljung, *System Identification, theory for the user*, 2nd ed. Prentice-Hall, 1999.

[2] T. Söderström and P. Stoica, *System identification*. Prentice-Hall, Inc., 1988.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Prediction error identification of linear systems: a nonparametric gaussian regression approach," *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.

[5] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[6] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes–revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[7] G. Kitagawa and W. Gersch, *Smoothness priors analysis of time series*. Springer, 1996.

[8] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. on Automatic Control*, vol. 37, no. 7, pp. 913–928, 1992.

[9] T. Chen and L. Ljung, "On kernel structure for regularized system identification (i): a machine learning perspective," in *Proc. of the 17th IFAC SYSID, to appear*, 2015.

[10] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*. CRC Press, 2000.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. The MIT Press, 2006.

[12] P. S. Heuberger, P. M. Van den Hof, and B. Wahlberg, *Modelling and identification with rational orthogonal basis functions*. Springer, 2005.

[13] T. Chen and L. Ljung, "Regularized system identification using orthonormal basis functions," in *Proc. of the 14th European Control Conference (ECC), to appear*, 2015.

[14] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. on Automatic Control*, vol. 36, no. 5, pp. 551–562, 1991.

[15] ——, "System identification using Kautz models," *IEEE Trans. on Automatic Control*, vol. 39, no. 6, pp. 1276–1282, 1994.

[16] G. Wahba, *Spline models for observational data*. Siam, 1990.

[17] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, pp. 337–404, 1950.

[18] G. Pillonetto and G. De Nicolao, "Kernel selection in linear system identification part i: A Gaussian process perspective," in *Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 4318–4325.

[19] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[20] G. Pillonetto and A. Chiuso, "Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator," *Automatica*, vol. 58, pp. 106–117, 2015.

[21] T. Chen and L. Ljung, "Implementation of algorithms for tuning parameters in regularized least squares problems in system identification," *Automatica*, vol. 49, no. 7, pp. 2213–2220, 2013.

[22] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 209, pp. 415–446, 1909.

[23] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra, "A generalized orthonormal basis for linear dynamical systems," *IEEE Trans. on Automatic Control*, vol. 40, no. 3, pp. 451–465, 1995.

[24] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[25] C. Carmeli, E. D. Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Analysis and Applications*, vol. 4, no. 4, pp. 377–408, 2006.

[26] A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto, "On the design of multiple kernels for nonparametric linear system identification," in *Proc. of the 53rd IEEE Conference on Decision and Control*, 2014, pp. 3346–3351.