A second-order generalization of TC and DC kernels

Mattia Zorzi, *Senior Member, IEEE*

***Abstract*— Kernel-based methods have been successfully introduced in system identification to estimate the impulse response of a linear system. Adopting the Bayesian viewpoint, the impulse response is modeled as a zero mean Gaussian process whose covariance function (kernel) is estimated from the data. The most popular kernels used in system identification are the tuned-correlated (TC), the diagonal-correlated (DC) and the stable spline (SS) kernel. TC and DC kernels admit a closed form factorization of the inverse. The SS kernel induces more smoothness than TC and DC on the estimated impulse response, however, the aforementioned property does not hold in this case. In this paper we propose a second-order extension of the TC and DC kernel which induces more smoothness than TC and DC, respectively, on the impulse response and a generalized-correlated kernel which incorporates the TC and DC kernels and their second order extensions. Moreover, these generalizations admit a closed form factorization of the inverse and thus they allow to design efficient algorithms for the search of the optimal kernel hyperparameters. We also show how to use this idea to develop higher oder extensions. Interestingly, these new kernels belong to the family of the so called exponentially convex local stationary kernels: such a property allows to immediately analyze the frequency properties induced on the estimated impulse response by these kernels.**

***Index Terms*— Covariance extension; Gaussian process; kernel methods; maximum entropy; system identification.**

I. Introduction

Linear system identification problems are traditionally addressed by using Prediction Error Methods (PEM), see [1], [2]. Here, the best model is chosen over a fixed parametric model class (e.g. ARMAX, OE, Box-Jenkins). This approach, however, has two issues: first, the parametrization of the predictor is nonlinear which implies that the minimization of the squared prediction error leads to a non-convex optimization problem; second, we have to face a model selection problem (i.e. order selection) which is usually performed by AIC and BIC criteria [3], [4].

Regularized kernel-based methods have been recently proposed in system identification in order to overcome the aforementioned limitations, see [5]–[7]. Here, we search the candidate model, described via the predictor impulse response, in an infinite dimensional nonparametric model class with the help of a penalty term. Adopting the Bayesian viewpoint, this is a Gaussian process regression problem [8]: the impulse response is modeled as a Gaussian process with zero mean and with a suitable covariance function, also called kernel [9]. The latter encodes the a priori knowledge about the predictor impulse response. For instance, the impulse response should be absolutely integrable, i.e. the corresponding system is Bounded Input Bounded Output (BIBO) stable, and with a certain degree of smoothness.

The most popular kernels are the tuned-correlated (TC), the diagonal-correlated (DC), and the stable-spline (SS), see [5], [6]. All these kernels encode the BIBO stability property. Regarding the smoothness, SS is the one inducing more smoothness on the impulse response. These kernels depend on few hyperparameters that are learnt from the data by minimizing the so called negative log-marginal likelihood. This task is computationally expensive especially in the case we want to estimate high dimensional models, e.g. the case of dynamic networks, see [10]–[14].

To reduce the computational complexity different strategies have been proposed, see [15]–[18]. In particular, if the kernel matrix admits a closed form expression for Cholesky factor of its inverse matrix (ant thus also its determinant), then the evaluation of the marginal likelihood can be done efficiently [18]. While it is possible to derive these closed form expressions for TC and DC, see [19], [20], this is not possible for SS. It is worth noting that an efficient algorithm for the SS kernel has been proposed in [17]. The latter, however, can be used only in the case that the input of the system has a prescribed structure, e.g. it cannot be used in the case we collect the data of a system which is in a feedback configuration. It is worth noting that many other kernel design extensions have been proposed, see for instance [21]–[26], however none of them provides a procedure to construct a kernel with both a degree of smoothness comparable with SS and a structure leading to an efficient algorithm for the negative log-marginal likelihood minimization.

The aim of this paper is to introduce a second-order generalization of the TC and DC kernel exploiting the filter-based approach proposed in [27]. These extensions induce more smoothness than TC and DC, respectively. We also introduce a generalized-correlated kernel which incorporates the DC, TC kernels and their second order extensions. Moreover, we show that they admit a closed form expression for the Cholesky of its inverse matrix. It is worth noting that SS is the second-order extension of the TC kernel derived in the continuous time, [28]. In contrast, the extension that we propose here is derived in the discrete time. The derivation of the proposed discrete time version, however, is conceptually different from

M. Zorzi is with the Department of Information Engineering, University of Padova, Padova, Italy; e-mail: zorzimat@dei.unipd.it (M. Zorzi).

the one in the continuous time: this extension is not obtained applying a change of coordinates to the cubic spline kernel. Numerical experiments showed that the new second-oder TC kernel represents an attractive alternative to SS because it leads to an estimation algorithm which outperforms the one using SS (even in the case that the computation of the Cholesky factorization of the kernel exploits the fact that SS is extended 2-semiseparable) in terms of computational time, while the second-order TC and SS are similar in terms of estimation performance. This idea can be also used to higher order extensions and also to generalize the high frequency kernel proposed in [29]. Interestingly, all these new kernels are exponentially convex local stationary (ECLS), [21], [23]. Such a property allows to easily understand the frequency properties of their stationary parts.

The outline of the paper is as follows. In Section II we briefly review the kernel-based PEM method as well as the TC, DC and SS kernels. Section III introduces the second-order extension for the TC kernel, while Section IV the one for the DC kernel. In Section V we introduce the generalized-correlation kernel. In Section VI we derive the closed form expressions for these kernels. In Section VII we extend this idea to higher order generalizations. In Section VIII we show that these kernels are ECLS and we analyze the stationary part of these kernels in the frequency domain. In Section IX we provide an interpretation of the proposed kernels using a system theory perspective. Finally, we draw the conclusions in Section X.

Notation. $\mathcal{S}_T$, with $T \leq \infty$, denotes the cone of positive definite symmetric matrices of dimension $T \times T$. Infinite dimensional matrices, i.e. matrices having an infinite number of columns and/or rows, are denoted using the calligraphic font, e.g. $\mathcal{K}$, while finite dimensional ones are denoted using the normal font, e.g. K. Given $\mathcal{F} \in \mathbb{R}^{p\times\infty}$ and $\mathcal{G} \in \mathbb{R}^{\infty\times m}$, the product $\mathcal{F}\mathcal{G}$ is understood as a $p \times m$ matrix whose entries are limits of infinite sequences [30]. Given $K \in \mathcal{S}_T$, $[K]_{t,s}$ denotes the entry of K in position (t,s), while $[K]_{:,t}$ and $[K]_{t,:}$ denotes the t-th column and row, respectively, of K. Given $K \in \mathcal{S}_T$, $\|v\|_{K^{-1}}$ denotes the weighted Euclidean norm of v with weight K^{-1}; the Euclidean norm of v is denoted by $\|v\|$. Given $v \in \mathbb{R}^T$, $\mathrm{Tpl}(v)$ denotes the lower triangular $T \times T$ Toeplitz matrix whose first column is given by v, while $\mathrm{diag}(v)$ denotes the diagonal matrix whose main diagonal is v.

II. Kernel-based PEM method

Consider the model

$$y(t) = \sum_{k=1}^{\infty} g(k)u(t-k) + e(t), \quad t = 1 \dots N \tag{1}$$

where $y(t)$, $u(t)$, $g(t)$ and $e(t)$ denote the output, the input, the impulse response of the model and a zero-mean white Gaussian noise with variance σ^2, respectively. We can rewrite model (1) as

$$y = \mathcal{A}g + e$$

where $y = [\,y(1) \dots y(N)\,]^\top \in \mathbb{R}^N$, e is defined likewise, $\mathcal{A}^{N\times\infty}$ is the regression matrix whose entries depends on $u(t)$ with $t = 1 \dots N$, $g = [\,g(1)\;\; g(2) \dots]^\top \in \mathbb{R}^\infty$. We want to estimate the impulse response g given the measurements $\{y(t), u(t)\}_{t=1}^N$. Such a problem is ill-posed because we have a finite number of measurements while g contains infinite parameters. The latter can be made well-posed assuming that $g \sim \mathcal{N}(0, \lambda\mathcal{K}(\eta))$ where $\mathcal{K}(\eta) \in \mathcal{S}_\infty$ is the kernel function and η is the vector of hyperparameters characterizing the kernel; in this way, the minimum variance estimator of g is:

$$\hat{g} = \underset{g\in\mathbb{R}^\infty}{\mathrm{argmin}} \|y - \mathcal{A}g\|^2 + \frac{\sigma^2}{\lambda}\|g\|^2_{\mathcal{K}(\eta)^{-1}} \tag{2}$$

where $\hat{g}$ belongs to the reproducing kernel Hilbert space (RKHS) with kernel function $\mathcal{K}(\eta)$ and norm $\|\cdot\|_{\mathcal{K}(\eta)^{-1}}$; $\lambda > 0$ denotes the regularization parameter. It is worth noting that the above problem admits a closed form solution. Moreover, $\mathcal{K}(\eta)$ encodes the a priori information that we have on the impulse response.

The aforementioned problem can be formulated as a finite dimensional problem. Indeed, g can be truncated, obtaining a finite impulse response of length T; the corresponding kernel matrix $K(\eta) \in \mathcal{S}_T$ is defined as $[K(\eta)]_{t,s} = [\mathcal{K}(\eta)]_{t,s}$ for $t, s = 1 \dots T$ and the regression matrix $A \in \mathbb{R}^{N\times T}$ is given by the first T columns of $\mathcal{A}$. Such a truncation, with T sufficiently large, introduces a negligible bias, because g decays to zero. The so called hyperparameters λ and η are estimated by minimizing numerically the negative log-marginal likelihood

$$\begin{aligned}\ell(y; \lambda, \eta) := {}& \log\det(\lambda AK(\eta)A^\top + \sigma^2 I)\\ &+ y^\top(\lambda AK(\eta)A^\top + \sigma^2 I)^{-1}y.\end{aligned} \tag{3}$$

In what follows, we will drop the dependence on η for kernels in order to ease the notation.

A. Diagonal and correlated kernels: an overview

We briefly review the most popular kernels used in system identification, see [6] for a more complete overview. The simplest kernel is diagonal and encodes the a priori information that g should decay to zero exponentially:

$$\mathcal{K}_{DI} = \mathrm{diag}(\beta, \beta^2, \dots, \beta^t, \dots) \tag{4}$$

where $\eta = \beta$ and $0 < \beta < 1$. Indeed, the penalty term $\|g\|^2_{\mathcal{K}_{DI}^{-1}}$ is the squared norm of the weighted impulse response

$$h = [\,h_1\; h_2 \dots h_t \dots]^\top, \quad h_t = \beta^{-t/2} g_t$$

which amplifies in an exponential way the coefficients g_t as t increases. The tuned-correlated (TC, also called first-order stable spline) kernel embeds also the a priori information that g is smooth:

$$[\mathcal{K}_{TC}]_{t,s} = \beta^{\max(t,s)} \tag{5}$$

where $\eta = \beta$ and $0 < \beta < 1$. The smoothness property can be justified as follows. It is well known that

$$\mathcal{K}_{TC} = (1-\beta)(\mathcal{F}\mathcal{D}\mathcal{F}^T)^{-1}$$

where

$$\begin{aligned}\mathcal{F} &= \mathrm{Tpl}(1, -1, 0, \dots)\\ \mathcal{D} &= \mathrm{diag}(\beta^{-1}, \beta^{-2}, \dots, \beta^{-t}, \dots);\end{aligned}$$

then, $\mathcal{F}^\top$ is the prefiltering operator, see [27], performing the first order difference of g and thus the penalty term in (2) penalizes impulses responses for which the weighted norm of the first oder difference of g is large

$$\begin{aligned}\|g\|^2_{\mathcal{K}_{TC}^{-1}} &= (1-\beta)^{-1}\|\mathcal{F}^\top g\|^2_{\mathcal{D}}\\ &= (1-\beta)^{-1}\sum_{t=1}^{\infty}\beta^{-t}(g_t-g_{t+1})^2.\end{aligned}$$

The diagonal-correlated (DC) kernel is defined as

$$[\mathcal{K}_{DC}]_{t,s} = \alpha^{|t-s|}\beta^{\max(t,s)} \tag{6}$$

where $0<\beta<1$, $-\beta^{-1/2}<\alpha<\beta^{-1/2}$ and $\eta = [\alpha\ \beta]^\top$. It is worth noting that we are taking a definition which is not standard, the standard one is $[\mathcal{K}_{DC}]_{t,s} = \rho^{|t-s|}\beta^{\frac{t+s}{2}}$ and $\rho = \alpha\beta^{1/2}$, because the former highlights the following limits:

$$\lim_{\alpha\to 0}\mathcal{K}_{DC} = \mathcal{K}_{DI}, \quad \lim_{\alpha\to 1}\mathcal{K}_{DC} = \mathcal{K}_{TC} \tag{7}$$

that is the DC kernel connects the DI and TC kernel. Indeed, it is not difficult to see that

$$\mathcal{K}_{DC} = (1-\alpha\beta)(\mathcal{F}_\alpha\mathcal{D}\mathcal{F}_\alpha^T)^{-1}$$

with

$$\mathcal{F}_\alpha = \mathrm{Tpl}(1,-\alpha,0,\dots). \tag{8}$$

In plain words, α tunes the behavior of the prefiltering operator: $\mathcal{F}_\alpha^\top$ behaves as the identity operator for α close to zero, while it behaves as the first order difference operator for α close to one. As a consequence the DC kernel allows to tune the degree of smoothness of g.

All these kernels admit a closed form factorization of the inverse and determinant which is an appealing feature for minimizing numerically (3). Moreover, their inverses are banded matrices: $\mathcal{K}_{DI}^{-1}$ is diagonal, $\mathcal{K}_{TC}^{-1}$ and $\mathcal{K}_{DC}^{-1}$ are tridiagonal.

The stable spline (SS, also called second-order stable spline) kernel induces more smoothness than TC:

$$[\mathcal{K}_{SS}]_{t,s} = \frac{\gamma^{t+s}\gamma^{\max(t,s)}}{2} - \frac{\gamma^{3\max(t,s)}}{6} \tag{9}$$

where $\eta=\gamma$ and $0<\gamma<1$. However, it does not admit a closed form factorization of the inverse and determinant. Moreover, since the SS kernel cannot be interpreted as the maximum entropy solution of a covariance extension problem similar to the one in [20] (see also Section VI-A), its inverse is not banded. Finally, a kernel with the aforementioned properties which tunes the degree of smoothness and connects TC with SS does not exist.

III. Second-order TC kernel

In this section we derive a new kernel, hereafter called TC2, which induces more smoothness than TC and represents an alternative to SS. In order to induce more smoothness it is sufficient to take the penalty term as the weighted norm of the second order difference of g:

$$\|g\|^2_{\mathcal{K}_{TC2}^{-1}} = (1-\beta)^{-3}\|(\mathcal{F}^T)^2 g\|^2_{\mathcal{D}},$$

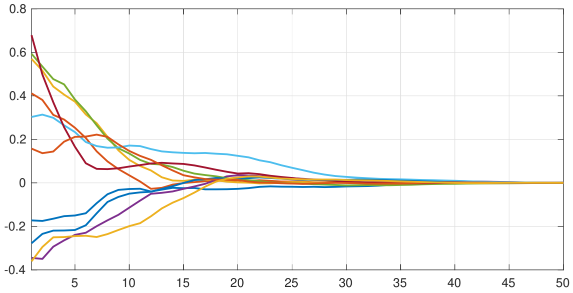

Fig. 1. Ten realizations of $\boldsymbol{g \sim \mathcal{N}(0, \lambda\mathcal{K}_{TC2})}$ with $\boldsymbol{\beta = 0.8}$ and $\boldsymbol{\lambda = \|\mathcal{K}_{TC2}\|^{-1}}$.

thus

$$\mathcal{K}_{TC2} := (1-\beta)^3(\mathcal{F}^2\mathcal{D}(\mathcal{F}^\top)^2)^{-1}$$

where $\eta=\beta$ and $0<\beta<1$. Figure 1 shows ten realizations of g using the TC2 kernel with $\beta=0.8$. We can notice that the degree of smoothness is similar to the one with $\mathcal{K}_{SS}$.

Proposition 3.1: The inverse of $\mathcal{K}_{TC2}$ is a pentadiagonal matrix, that is $[(\mathcal{K}_{TC2})^{-1}]_{t,s}=0$ for any $|t-s|>3$.

Proof: The statement is a particular instance of Proposition 7.1, see Section VII. ■

Throughout the paper we will use the following result.

Lemma 3.1 ([31]): Consider a real infinite lower triangular Toeplitz matrix, defined by the sequence $\{a_k,\ k\geq 0\}$ as follows

$$\mathcal{X} = \mathrm{Tpl}(a_0,a_1,a_2,\dots).$$

If $a_0\neq 0$, $\mathcal{X}$ is invertible and the inverse matrix $\mathcal{Y}=\mathcal{X}^{-1}$ is also a lower triangular Toeplitz matrix with elements $\{b_k,\ k\geq 0\}$ given by the following formula

$$b_0 = \frac{1}{a_0},\ b_k = -\frac{1}{a_0}\sum_{j=0}^{k-1}a_{k-j}b_j \text{ for } k\geq 1.$$

Proposition 3.2: $\mathcal{K}_{TC2}$ admits the following closed form expression:

$$[\mathcal{K}_{TC2}]_{t,s} = 2\beta^{\max(t,s)+1} + (1-\beta)(1+|t-s|)\beta^{\max(t,s)}. \tag{10}$$

Proof: First, $\mathcal{F}$ is a lower triangular Toeplitz matrix which is invertible because the main diagonal is composed by strictly positive elements. Therefore, by Lemma 3.1 we have

$$\mathcal{F}^{-2} = \mathrm{Tpl}(1,2,\dots,t,\dots).$$

Moreover,

$$[\mathcal{F}^{-2}]^\top_{:,t} = [0\ \dots\ 0\ \underbrace{1}_{t\text{-th element}}\ 2\ \ 3\ \dots].$$

Therefore,

$$\begin{aligned}[\mathcal{K}_{TC2}]_{t,s} &= (1-\beta)^3[(\mathcal{F}^{-2})^\top\mathcal{D}^{-1}\mathcal{F}^{-2}]_{t,s}\\ &= (1-\beta)^3[\mathcal{F}^{-2}]^\top_{:,t}\mathcal{D}^{-1}[\mathcal{F}^{-2}]_{:,s}\\ &= \sum_{k=\max(t,s)}^{\infty}\beta^k(k-t+1)(k-s+1).\end{aligned}$$

Finally, it is not difficult to see that the above series converges to (10) by exploiting the identity

$$\sum_{k=0}^{\infty} \beta^k = \frac{1}{1-\beta}. \tag{11}$$

■

It is worth noting that the SS kernel is also a second-order generalization of the TC kernel. Indeed, TC and SS are obtained by applying a “stable” coordinate change to the first and second order, respectively, spline kernel [5]. That extension has been derived in the continuous time domain, while the one proposed here has been derived in the discrete time domain.

IV. Second-order DC kernel

The aim of this section is to introduce a new kernel, hereafter called DC2, which connects the TC and TC2 kernels. The unique difference between TC and TC2 is the prefiltering operator acting on g. Thus, the DC2 kernel should perform a transition from $\mathcal{F}$ to $\mathcal{F}^2$. One possible way is to take

$$\mathcal{F}_{2,\alpha} := (1-\alpha)\mathcal{F} + \alpha\mathcal{F}^2 \tag{12}$$

with $0 \le \alpha \le 1$ and thus we obtain

$$\mathcal{K}_{DC2} := \kappa(\mathcal{F}_{2,\alpha}\mathcal{D}\mathcal{F}_{2,\alpha}^{\top})^{-1} \tag{13}$$

with $\kappa = (1-\beta)(1-\alpha\beta)(1-\alpha^2\beta)$. In this case we have $\eta = [\alpha \;\, \beta]^{\top}$ with $0 < \beta < 1$. From the above definition it follows that

$$\lim_{\alpha\to 0}\mathcal{K}_{DC2} = \mathcal{K}_{TC}, \quad \lim_{\alpha\to 1}\mathcal{K}_{DC2} = \mathcal{K}_{TC2}. \tag{14}$$

Figure 2 shows a realization of the impulse response as a function of α using (13); as expected, the degree of smoothness increases as α increases.

Remark 1: It is worth noting that one could consider other transitions, e.g.

$$\begin{aligned}\mathcal{F}_{2,\alpha} &= \mathrm{Tpl}(1, -1-\alpha^2, \alpha, 0, \dots)\\ \mathcal{F}_{2,\alpha} &= ((1-\alpha)\mathcal{F}^{-1} + \alpha\mathcal{F}^{-2})^{-1}.\end{aligned}$$

However, as we will see, (12) is the unique definition which guarantees that K_{DC2} admits a closed form expression and is the maximum entropy solution of a matrix completion problem.

Proposition 4.1: The inverse of $\mathcal{K}_{DC2}$ is a pentadiagonal matrix, that is $[(\mathcal{K}_{DC2})^{-1}]_{t,s} = 0$ for any $|t-s| > 3$.

Proof: The proof is similar to the one of Proposition 3.1.
■ ■

Proposition 4.2: For $0 \le \alpha < 1$, $\mathcal{K}_{DC2}$ admits the following closed form expression:

$$[\mathcal{K}_{DC2}]_{t,s} = \frac{\beta^{\max(t,s)}(1-(1-\beta)\alpha^{|t-s|+1}) - \alpha^2\beta^{\max(t,s)+1}}{1-\alpha}. \tag{15}$$

Proof: First, we notice that

$$\mathcal{F}_{2,\alpha} = ((1-\alpha)\mathcal{I} + \alpha\mathcal{F})\mathcal{F} = \mathcal{F}_\alpha\mathcal{F}$$

where $\mathcal{F}_\alpha$ has been defined in (8); $\mathcal{I}$ is the identity matrix of infinite dimension. The main diagonal of $\mathcal{F}$ and $\mathcal{F}_\alpha$ is composed by strictly positive elements and thus their inverse exist. By Lemma 3.1, we have

$$\begin{aligned}\mathcal{F}^{-1} &= \mathrm{Tpl}(1, 1, \dots)\\ \mathcal{F}_\alpha^{-1} &= \mathrm{Tpl}(1, \alpha, \alpha^2, \dots).\end{aligned}$$

Therefore,

$$\mathcal{F}_{2,\alpha}^{-1} = \frac{1}{1-\alpha}\,\mathrm{Tpl}(1-\alpha, 1-\alpha^2, 1-\alpha^3, \dots).$$

Finally,

$$\begin{aligned}[\mathcal{K}_{DC2}]_{t,s} &= \kappa[\mathcal{F}_{2,\alpha}^{-1}]_{:,t}^{\top}\mathcal{D}[\mathcal{F}_{2,\alpha}^{-1}]_{:,s}\\ &= \kappa\sum_{k=\max(t,s)}^{\infty}\beta^k\frac{1-\alpha^{k-t+1}}{1-\alpha}\,\frac{1-\alpha^{k-s+1}}{1-\alpha}\end{aligned}$$

where the above series converges to right hand side of (15). The latter fact can be easily proved by using Identity (11). ■

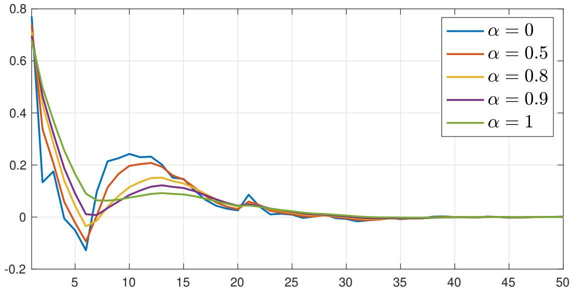

Fig. 2. One realization of $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\mathcal{K}_{DC2})$ for different values of $\boldsymbol{\alpha}$. Here, $\boldsymbol{\lambda} = \|\mathcal{K}_{DC2}\|^{-1}$.

V. Generalized-correlated kernel

In view of (7) and (14) we can define a general kernel, hereafter called generalized-correlated (GC) kernel, that incorporates the DI, DC, TC, DC2 and TC2 kernels. Let $\mathcal{K}_{DI}(\beta)$, $\mathcal{K}_{DC}(\alpha,\beta)$, $\mathcal{K}_{TC}(\beta)$, $\mathcal{K}_{DC2}(\alpha,\beta)$ and $\mathcal{K}_{TC2}(\beta)$ be the kernels defined in (4), (6), (5), (15) and (10), respectively, where we made explicit their dependence on the hyperparameters $0 < \alpha < 1$ and $0 < \beta < 1$. Then, we define as GC kernel

$$\mathcal{K}_{GC}(\gamma,\beta) = \begin{cases}\mathcal{K}_{DI}(\beta), & \gamma = 0\\ \mathcal{K}_{DC}(\gamma,\beta), & 0 < \gamma < 1\\ \mathcal{K}_{TC}(\beta), & \gamma = 1\\ \mathcal{K}_{DC2}(\gamma-1,\beta), & 1 < \gamma < 2\\ \mathcal{K}_{TC2}(\beta), & \gamma = 2.\end{cases} \tag{16}$$

where γ characterizes the smoothness of the impulse response over a wide range. It is worth noting that $\mathcal{K}_{GC}$ is a continuous function with respect to γ and β, but not differentiable.

It is worth noting that the GC kernels leads to an estimator with less bias and more variance than the one using DI, DC, TC, DC2 or TC2. Accordingly, the GC kernel provides a different bias-variance tradeoff which is better than the one of the other kernels in some specific situations as shown in the next two Monte Carlo studies. The first Monte Carlo study is composed by 200 experiments. In each experiment we

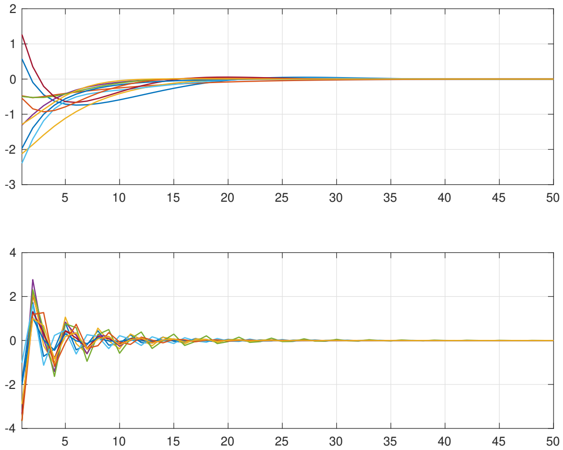

Fig. 3. *Top panel*. Ten realizations of the impulse response in the first Monte Carlo study. *Bottom panel*. Ten realizations of the impulse response in the second Monte Carlo study.

generate the impulse response g with practical length $T = 50$ as follows:

$$g_t = \sum_{k=1}^{10} a_k \cos(b_k t + c_k)$$

where its parameters are drawn as follows: $a_k \in \mathcal{U}([0.2, 0.9])$, $b_k \in \mathcal{U}([10^{-6}\pi, 10^{-1}\pi])$ and $c_k \in \mathcal{U}([0, \pi])$. Figure 3 (top) shows ten realizations drawn from such process. Then, we generate the input of length $N = 500$ using the MATLAB function `idinput.m` as a realization drawn from a Gaussian noise with band [0, 0.6]. Then, we feed the corresponding system (1) with it obtaining the dataset $\mathrm{D}^N := \{y(t), u(t)\}_{t=1}^N$. Here, σ^2 is chosen in such a way that the signal to noise ratio is equal to two. Then, we estimate the impulse response using the following estimators:

- $\hat{g}_{DI}$ is the estimator in (2) using the diagonal kernel (4);
- $\hat{g}_{DC}$ is the estimator in (2) using the DC kernel (6);
- $\hat{g}_{TC}$ is the estimator in (2) using the TC kernel (5);
- $\hat{g}_{D2}$ is the estimator in (2) using the DC2 kernel (15);
- $\hat{g}_{T2}$ is the estimator in (2) using the TC2 kernel (10);
- $\hat{g}_{SS}$ is the estimator in (2) using the SS kernel (9);
- $\hat{g}_{GC}$ is the estimator in (2) using the GC kernel (16).

The hyperparameters of the kernels used in the aforementioned estimators are estimated by minimizing numerically the negative log-marginal likelihood in (3). Finally, for each estimator we compute the impulse response fit

$$\mathrm{FIT} = 100\left(1 - \frac{\|g - \hat{g}\|}{\|g - \bar{g}\|}\right) \tag{17}$$

where $\bar{g} = \sum_{k=1}^{T} g(k)$ and $\hat{g}$ is the corresponding estimator. Clearly, the more FIT is close to 100, the better the estimator performance is. Figure 4 (top) shows the boxplot of FIT for the estimators over the 200 experiments: D2, T2, SS and GC are the best estimators, while DI is the worst one. In plain words, the best estimators are the ones that are able to induce a sufficient degree of smoothness on the impulse response.

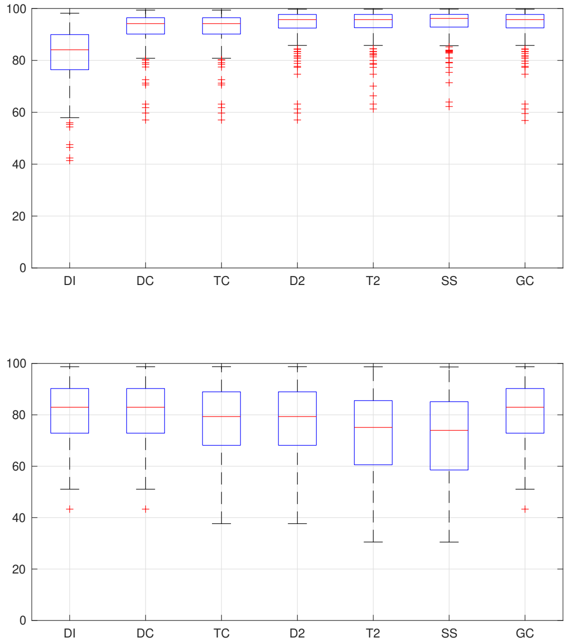

Fig. 4. Impulse response fit in the first (top) and second (bottom) Monte Carlo study composed by 200 experiments.

The second Monte Carlo study is likewise to the previous one, but $b_k \in \mathcal{U}([0.6\pi, 0.7\pi])$. In this case, the realizations of the process g_t are less smooth than before, see Figure 3 (bottom). Figure 4 (bottom) shows the boxplot of FIT for the estimators: DI, DC and GC are the best estimators, while T2 and SS are the worst ones. We conclude that if the estimators using DI, TC, DC, TC2, DC2 are with too much bias, then it is better to use GC.

VI. Efficient Implementation to Estimate the Hyperparameters

The minimization of (3) is typically performed through the nonlinear optimization solver `fmincon.m` of Matlab. Thus, the crucial aspect is to consider an efficient algorithm to evaluate (3). We show that the proposed kernels are suitable for this aim. Recall that $K \in \mathcal{S}_T$ denotes the finite dimensional kernel corresponding to $\mathcal{K}$ and defined as

$$[K]_{t,s} = [\mathcal{K}]_{t,s}, \quad t, s = 1 \ldots T.$$

If K^{-1} admits a closed form expression of its Cholesky factor, then the negative log-marginal likelihood in (3) can be evaluated efficiently as follows, see [20]:

$$\frac{r^2}{\sigma^2} + (N - T)\log\sigma^2 + \log\det(\lambda K) + 2\log\det R_1 \tag{18}$$

where L is the Cholesky factor of $K^{-1} = LL^T$ and R_1 is given by the QR factorization

$$\begin{bmatrix} R_{d1} & R_{d2} \\ \sigma\sqrt{\lambda^{-1}}L^\top & 0 \end{bmatrix} = QR = Q\begin{bmatrix} R_1 & R_2 \\ 0 & r \end{bmatrix}$$

where $Q \in \mathbb{R}^{2T+1\times T+1}$, $R_1 \in \mathbb{R}^{T\times T}$, $R_2 \in \mathbb{R}^{T}$ and $r \in \mathbb{R}$. Moreover, R_{d1} and R_{d2} is given by the QR factorization $[A\ \ y] = Q_d[R_{d1}\ \ R_{d2}]$ which can be computed “offline” before to start the optimization task. In what follows we show that TC2, DC2 and GC admit a closed form expression for L and thus also $\log\det(\lambda K)$.

Proposition 6.1: The inverse of $K_{TC2} \in \mathcal{S}_T$ admits the following decomposition

$$K_{TC2}^{-1} = (1-\beta)^{-3} F_T^2 D_T (F_T^2)^\top$$

where

$$\begin{aligned} F_T &= \mathrm{Tpl}(1,-1,0,\dots 0) \in \mathbb{R}^{T\times T}\\ D_T &= \begin{bmatrix} D_{1,T} & 0\\ 0 & B_T\end{bmatrix}\\ D_{1,T} &= \mathrm{diag}(\beta^{-1},\beta^{-2},\dots \beta^{T-2})\\ B_T &= (1-\beta)\beta^{-T}\begin{bmatrix} \beta+\beta^2 & 2\beta^2\\ 2\beta^2 & 1-3\beta+4\beta^2\end{bmatrix}. \end{aligned}$$

Thus, K_{TC2}^{-1} is a pentadiagonal matrix.

Proof: Consider

$$X := (1-\beta)^3 (F_T^2 \tilde{D}_T (F_T^2)^\top)^{-1}$$

where

$$\tilde{D}_T = \mathrm{diag}(\beta^{-1},\beta^{-2}\dots,\beta^{-T}). \tag{19}$$

It is not difficult to see that

$$F_T^{-2} = \mathrm{Tpl}(1,2,\dots,T).$$

Thus, by arguments similar to ones used in the proof of Proposition 3.2, we have

$$[X]_{t,s} = (1-\beta)^3 \sum_{k=\max(t,s)}^{T} \beta^k (k-t+1)(k-s+1).$$

Without loss of generality, we assume that $t \geq s$; hence,

$$\begin{aligned} [X]_{t,s} &= (1-\beta)^3 \sum_{k=t}^{T} \beta^k (k-t+1)(k-s+1).\\ &= (2\beta + (1-\beta)(1+t-s))\beta^t + \eta(t,s) \end{aligned}$$

where

$$\begin{aligned} \eta(t,s) = (1-\beta)(&\beta^{T+2}(T-t+1)(T-s+3)\\ &-\beta^{T+1}(T-t+2)(T-s+2))\\ &+2\beta^{T+2}((T-t+1)\beta-(T-t+2)) \end{aligned}$$

and we have exploited the fact that

$$\sum_{k=0}^{T} \beta^k = \frac{1-\beta^{T+1}}{1-\beta}.$$

Notice that

$$[K_{TC2}]_{t,s} = [X]_{t,s} - \eta(t,s)$$

and we can rewrite η in the shorthand way

$$\eta(t,s) = \gamma_1 t + \gamma_1 s + \gamma_2 ts + \gamma_3 \tag{20}$$

where γ_k's are constants not depending on t and s. On the other hand, if we take

$$\begin{aligned} Y :&= (1-\beta)^3 (F_T^2 \Delta^{-1} (F_T^2)^\top)^{-1}\\ &= (1-\beta)^3 (F_T^{-2})^\top \Delta F_T^{-2}, \end{aligned}$$

with

$$\Delta = \begin{bmatrix} 0 & \dots & 0 & \dots & 0\\ \vdots & \ddots & \vdots & & \vdots\\ \vdots & & 0 & z & y\\ 0 & \dots & 0 & y & x \end{bmatrix}, \tag{21}$$

then it is not difficult to see that

$$\begin{aligned} [Y]_{t,s} =& (1-\beta)^3[-(T(x+2y+z)+x+y)(t+s)\\ &+(x+2y+z)ts + 2T(x+y)+x]. \end{aligned}$$

By taking into account (20), we can impose that x, y, z obey the conditions

$$\begin{aligned} \gamma_1 &= -(1-\beta)^3(T(x+2y+z)+x+y)\\ \gamma_2 &= (1-\beta)^3(x+2y+z)\\ \gamma_3 &= (1-\beta)^3[2T(x+y)+x]. \end{aligned}$$

In this way, $\eta(t,s) = [Y]_{t,s}$. With this choice, we have

$$\begin{aligned} K_{TC2} = X - Y &= (1-\beta)^3 (F_T^{-2})^\top (\tilde{D}_T^{-1} - \Delta)(F_T^{-2})\\ &= (1-\beta)^3 (F_T^2 (\tilde{D}_T^{-1} - \Delta)^{-1} (F_T^2)^\top)^{-1} \end{aligned}$$

where it is not difficult to see that $(\tilde{D}_T^{-1} - \Delta)^{-1}$ coincides with D_T. Finally, the fact that K_{TC2}^{-1} is pentadiagonal follows from Proposition 7.2, see Section VII. ■

Proposition 6.2: The inverse of $K_{DC2} \in \mathcal{S}_T$ admits the following decomposition

$$K_{DC2}^{-1} = \kappa^{-1} F_{2,\alpha,T} D_T F_{2,\alpha,T}^\top$$

where

$$\begin{aligned} F_{2,\alpha,T} &= (1-\alpha)F_T + \alpha F_T^2\\ D_T &= \begin{bmatrix} D_{1,T} & 0\\ 0 & B_T\end{bmatrix}\\ D_{1,T} &= \mathrm{diag}(\beta^{-1},\beta^{-2},\dots \beta^{T-2})\\ B_T &= (1-\alpha\beta)\beta^{-T}\\ &\times \begin{bmatrix} \beta(1+\alpha\beta) & \alpha\beta^2(1+\alpha)\\ \alpha\beta^2(1+\alpha) & (1-\beta-\alpha^2\beta)(1-\alpha\beta)+2\alpha^2\beta^2\end{bmatrix}. \end{aligned}$$

Thus, K_{DC2}^{-1} is a pentadiagonal matrix.

Proof: Consider

$$X := \kappa(F_{2,\alpha,T}\tilde{D}_T F_{2,\alpha,T}^\top)^{-1}$$

where $\tilde{D}_T$ has been defined in (19). Notice that $F_{2,\alpha,T} = F_{\alpha,T}F_T$ where

$$F_{\alpha,T} = \mathrm{Tpl}(1,-\alpha,0\dots,0) \in \mathbb{R}^{T\times T}$$

and

$$\begin{aligned} F_{\alpha,T}^{-1} &= \mathrm{Tpl}(1,\alpha^2,\dots,\alpha^T)\\ F_{2,\alpha,T}^{-1} &= F_T^{-1}F_{\alpha,T}^{-1}\\ &= \frac{1}{1-\alpha}\mathrm{Tpl}(1-\alpha,1-\alpha^2,\dots,1-\alpha^T). \end{aligned}$$

Without loss of generality, we assume that $t \geq s$, then it is not difficult to see that

$$\begin{aligned}[X]_{t,s} &= [(F_T^{-1})^\top (F_{\alpha,T}^{-1})^\top \tilde{D}_T^{-1} F_{\alpha,T}^{-1} F_T^{-1}]_{t,s} \\ &= \frac{1}{(1-\alpha)^2} \sum_{k=t}^{T} \beta^k (1-\alpha^{k-t+1})(1-\alpha^{k-s+1}) \\ &= [K_{DC2}]_{t,s} + \eta(t,s)\end{aligned}$$

where

$$\eta(t,s) = \gamma_1 \alpha^{-t} + \gamma_1 \alpha^{-s} + \gamma_2 \alpha^{-(t+s)} + \gamma_3 \tag{22}$$

and γ_k's are constants not depending on t and s. On the other hand, if we take

$$Y := \kappa(F_{2,\alpha,T} \Delta^{-1} F_{2,\alpha,T}^\top)^{-1} = \kappa(F_{2,\alpha,T}^{-1})^\top \Delta F_{2\alpha,T}^{-1}$$

where Δ is defined as in (21), then it is not difficult to see that

$$\begin{aligned}[Y]_{t,s} =& \kappa[-\alpha^T (z + y + \alpha x + \alpha y)(\alpha^{-t} + \alpha^{-s}) \\ &+ \alpha^{2T}(z + 2\alpha y + \alpha^2 x)\alpha^{-(t+s)} + (z + 2y + x)].\end{aligned}$$

By taking into account (22), we can impose that x, y, z obey the conditions

$$\begin{aligned}\gamma_1 &= -\kappa\alpha^T (z + y + \alpha x + \alpha y) \\ \gamma_2 &= \kappa\alpha^{2T}(z + 2\alpha y + \alpha^2 x) \\ \gamma_3 &= \kappa(z + 2y + x).\end{aligned}$$

In this way, $\eta(t,s) = [Y]_{t,s}$. With this choice, we have

$$\begin{aligned}K_{DC2} = X - Y &= \kappa(F_{2,\alpha,T}^{-1})^\top (\tilde{D}_T^{-1} - \Delta) F_{2,\alpha,T}^{-1} \\ &= \kappa(F_{2,\alpha,T}(\tilde{D}_T^{-1} - \Delta)^{-1} F_{2,\alpha,T}^\top)^{-1}\end{aligned}$$

where it is not difficult to see that $(\tilde{D}_T^{-1} - \Delta)^{-1}$ coincides with D_T. Finally, the fact that K_{DC2}^{-1} is pentadiagonal follows by Proposition 7.4 in Section VII. ■

By Proposition 6.1 and 6.2 we have following corollaries.

Corollary 6.1: Let L denote the Cholesky factor of K_{TC2}^{-1}, then

$$[L]_{t,s} = \begin{cases} \frac{1}{\sqrt{(1-\beta)^3 \beta^t}}, & 1 \leq t = s \leq T-2 \\ \frac{-2}{\sqrt{(1-\beta)^3 \beta^{t-1}}}, & 2 \leq t = s+1 \leq T-1 \\ \frac{1}{\sqrt{(1-\beta)^3 \beta^{t-2}}}, & 3 \leq t = s+2 \leq T \\ \frac{\sqrt{\beta^{-T+1}(1+\beta)}}{1-\beta}, & t = s = T-1 \\ \frac{-2\sqrt{\beta^{-T+1}}}{(1-\beta)\sqrt{1+\beta}}, & t = s+1 = T \\ \sqrt{\frac{\beta^{-T}}{1+\beta}}, & t = s = T \\ 0, & \text{otherwise.} \end{cases}$$

Moreover,

$$\det K_{TC2} = \beta^{\frac{T(T+1)}{2}} (1-\beta)^{3T-4}.$$

Corollary 6.2: Let L denote the Cholesky factor of K_{DC2}^{-1}, then

$$[L]_{t,s} = \begin{cases} \frac{1}{\sqrt{\kappa\beta^t}}, & 1 \leq t = s \leq T-2 \\ \frac{-(1+\alpha)}{\sqrt{\kappa\beta^{t-1}}}, & 2 \leq t = s+1 \leq T-1 \\ \frac{\alpha}{\sqrt{\kappa\beta^{t-2}}}, & 3 \leq t = s+2 \leq T \\ \sqrt{\frac{(1+\alpha\beta)\beta^{-T+1}}{(1-\beta)(1-\alpha^2\beta)}}, & t = s = T-1 \\ \frac{-(1+\alpha)\sqrt{\beta^{-T+1}}}{\sqrt{(1+\alpha\beta)(1-\beta)(1-\alpha^2\beta)}}, & t = s+1 = T \\ \sqrt{\frac{\beta^{-T}}{1+\alpha\beta}}, & t = s = T \\ 0, & \text{otherwise.} \end{cases}$$

Moreover,

$$\det K_{DC2} = \beta^{\frac{T(T+1)}{2}} (1-\alpha\beta)^{T-2} (1-\beta)^{T-1} (1-\alpha^2\beta)^{T-1}.$$

In view of the above properties, we have

$$\begin{aligned}\log\det(\lambda K_{TC2}) &= T\log\lambda + \frac{T(T+1)}{2}\log\beta \\ &+ (3T-4)\log(1-\beta) \\ \log\det(\lambda K_{DC2}) &= T\log\lambda + \frac{T(T+1)}{2}\log\beta \\ &+ (T-2)\log(1-\alpha\beta) + (T-1)\log(1-\beta) \\ &+ (T-1)\log(1-\alpha^2\beta).\end{aligned}$$

In view of the above corollaries and since K_{DI}^{-1}, K_{DC}^{-1}, K_{TC}^{-1} admit a closed form expression for the Cholesky factor, see [20], then it follows that the Cholesky factor of K_{GC}^{-1} admits a closed form expression. Moreover, the Cholesky factor L of K_{TC2}^{-1}, K_{DC2}^{-1} and K_{GC}^{-1} is a lower triangular and pentadiagonal matrix. Therefore, the construction of L requires the computation of $3(T-1)$ elements. Hence, the computational complexity for the computation of L is $O(T)$. It is worth noting that the most efficient way to compute the Cholesky factor of K_{SS} is to exploit the fact that the SS kernel is extended 2-semiseparable: first it is required to compute the generators of K_{SS}, say $U, W \in \mathbb{R}^{T\times 2}$; the latter can be computed through [32, Algorithm 4.2]; finally, $L_{SS} = \mathrm{tril}(UW^T)$ where $\mathrm{tril}(UW^T)$ denotes the lower triangular part of UW^T. The computational complexity of the resulting algorithm is $O(T^2)$. Accordingly, the computation of the Cholesky factor of K_{TC2}^{-1} is more efficient that the computation of the Cholesky factor of K_{SS}. Figure 5 confirms this analysis: it shows the average time (over 5000 runs) needed to compute the Cholesky factors of K_{SS} (blue line) and K_{TC2}^{-1} (red line) for different sizes T of the kernel matrices. As a consequence, the minimization of the log-marginal likelihood using TC2 and GC can be efficiently performed by means of the previous algorithm equipped with the closed form expressions. In order to test it, we consider a Monte Carlo study composed by 50 experiments where the models and the data are generated likewise to the first Monte Carlo study of Section V, but $a_k \in \mathcal{U}[0.2, 0.9995]$ and $N = 5000$. We consider the following algorithms to estimate the impulse response:

- **T2** is the algorithm in [20], i.e. the one explained before, to compute $\hat{g}_{T2}$ which exploits the fact that TC2 admits the closed form expression for L and $\log\det(\lambda K)$;

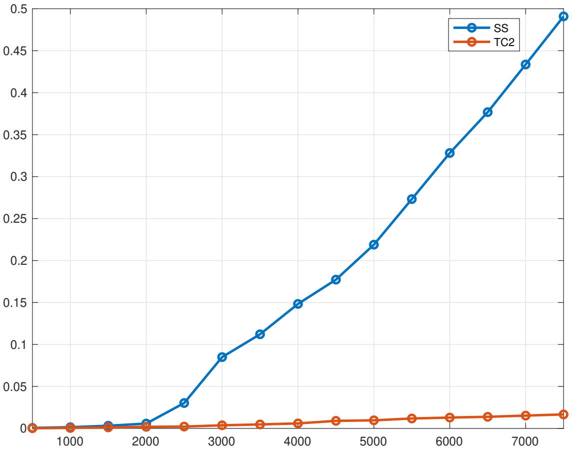

Fig. 5. Average computational time (in seconds) for different values of $\boldsymbol{T} \in [\mathbf{500}, \mathbf{7500}]$ to compute the Cholesky factor of $\boldsymbol{K_{SS}}$ (blue line) and the Cholesky factor of $\boldsymbol{K_{TC2}^{-1}}$ (red line).

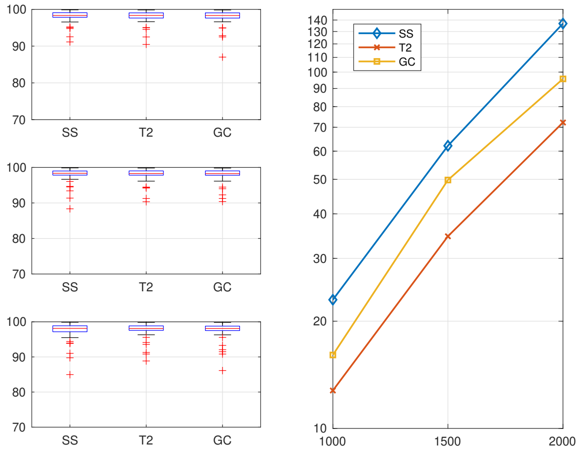

Fig. 6. *Left panels.* Impulse response fit for $\boldsymbol{T} = \mathbf{1000}$ (top), $\boldsymbol{T} = \mathbf{1500}$ (middle) and $\boldsymbol{T} = \mathbf{2000}$ (bottom). *Right panel.* Average computational time (in seconds) for $\boldsymbol{T} = \mathbf{1000}, \mathbf{1500}, \mathbf{2000}$.

- **GC** is the algorithm in [20], i.e. the one explained before, to compute $\hat{g}_{GC}$ which exploits the fact that GC admits the closed form expression for L and $\log\det(\lambda K)$;
- **SS** is the algorithm in [18] to compute $\hat{g}_{SS}$ which exploits the fact that the Cholesky factor, say L_{SS}, of K_{SS} can be computed efficiently through [32, Algorithm 4.2]; it is worth noting this algorithm neither computes L_{SS}^{-1} nor K_{SS}^{-1}.

In all the above algorithms `fmincon.m` is used with the default options with the exception of the maximum number of function evaluations and the maximum number of iterations which were set equal to 100000 and 10000. For any experiment we measure the computational time (in seconds) of these algorithms through the functions `tic` and `toc` in Matlab. The simulation is performed with MatlabR2021a and run on a MacBook Air with operating system macOS Big Sur, 3.2GHz Apple M1 processor and 8GB 4266 LPDDR4 memory. Figure 6 shows the average computational time for the three algorithms using as practical length $T = 1000, T = 1500, T = 2000$ (right panel) and the corresponding impulse response fit (17) (left panel). While the performance of the estimators is similar, **T2** exhibits the best computational time and **SS** the worst one. It is worth noting that the computational time of **GC** is worse than the one of **T2** because in the former we have to optimize three hyperparameters (i.e. λ, γ and β) while in the latter only two (i.e. λ and β). Finally, for the SS kernel we also considered the algorithm proposed in [20] where the Cholesky factor of K_{SS} is computed by [32, Algorithm 4.2]: the computational time was worse than the one of **SS**. In order to have an additional assessment between **SS** and **T2**, which is independent from the number of iterations executed by the fmincon routine, we have considered an additional Monte Carlo study composed by 50 trials with $N = 5000$ and T fixed. In each trial we generate randomly A, y, σ^2, λ and η. Then, we compute the execution time to evaluate the likelihood for **SS** and **T2** using these parameters. The execution time does not include the preliminary one-time operations (e.g. the computation of R_{d1} and R_{d2} for **T2**). Figure 7 shows the average computational time (in seconds) for different values of T. Still, **T2** is more efficient than **SS** in terms of computational time. We also tried different values of $N \geq T$ and still we have found similar results.

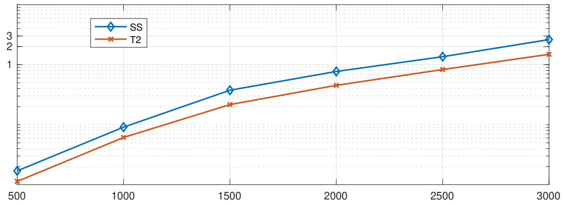

Fig. 7. Average computational time (in seconds) to evaluate the negative log-likelihood for different values of $\boldsymbol{T}$.

All these numerical experiments showed that TC2 makes more efficient the minimization of the negative log-marginal likelihood than SS, while the corresponding estimators exhibit a similar performance.

Remark 2: Although, we showed empirically that **T2** is more efficient than **SS**, it is also worth noting both these two algorithms exhibits the same asymptotic cost to evaluate the likelihood which is equal to $O(T^3)$. In addition, the computational efficiency of the algorithms is heavily dependent on the particular set-up and implementation. For instance, all the experiments have been performed without the use of code written in C or C++; the use of it can improve the performance of both the algorithms, however further investigation is required in order to draw some conclusion.

Remark 3: The fact that the inverse kernel matrix is pentadiagonal can be also used to compute efficiently (2) in the case that $T \gg N$ through the alternating direction method of multipliers (ADMM) proposed in [33]. Indeed, although that paper considers the case of tridiagonal inverse kernel matrices (e.g. TC and DC kernels) that idea holds also for banded inverse kernel matrices and the computational flops do not change.

A. Maximum Entropy interpretation

Proposition 6.1 and Proposition 6.2 are also important to show that the kernel matrices $K_{TC2} \in \mathcal{S}_T$ and $K_{DC2} \in \mathcal{S}_T$, with $T \geq 4$, are the maximum entropy solution of a matrix completion problem of the following form.

Problem 1 (Band extension problem): Given $m \in \mathbb{N}$ and $c_{t,s}$, with $|t-s| \leq m$, find the covariance matrix $\Sigma \in \mathcal{S}_T$ of a zero mean Gaussian random vector such that

$$[\Sigma]_{t,s} = c_{t,s}, \quad |t-s| \leq m.$$

Such an interpretation is important because, as pointed out by Dempster in [34], see also [35]–[38], "the principle of seeking maximum entropy is a principle of seeking maximum simplicity of explanation". Accordingly, these kernels represent the simplest way of embedding in the prior the fact that the impulse response describes a BIBO stable system and with certain degree of smoothness. Recall that the maximum entropy solution (or extension) of the above problem is defined as

$$\begin{aligned} &\max\nolimits_{\Sigma \in \mathcal{S}_T} \log\det\Sigma \\ &\text{subject to } [\Sigma]_{t,s} = c_{t,s}, \quad |t-s| \leq m. \end{aligned} \qquad (23)$$

The following statements can be proved using arguments similar to the ones in [20].

Theorem 6.1: Consider Problem 1 with $m = 2$ and

$$c_{t,s} = 2\beta^{\max(t,s)+1} + (1-\beta)(1+|t-s|)\beta^{\max(t,s)}, \qquad (24)$$

with $|t-s| \leq 2$. Then, the maximun entropy extension solution to (23) is K_{TC2}.

Theorem 6.2: Consider Problem 1 with $m = 2$ and

$$c_{t,s} = \frac{\beta^{\max(t,s)}(1-(1-\beta)\alpha^{|t-s|+1}) - \alpha^2\beta^{\max(t,s)+1}}{1-\alpha}, \qquad (25)$$

with $|t-s| \leq 2$. Then, the maximun entropy extension solution to (23) is K_{DC2}.

Accordingly, these two theorems tell us that given the covariance lags up to the second order (24) and (25) of the impulse response, then K_{TC2} and K_{DC2}, respectively, describe the correlation function of the process that best represents the current state of knowledge about the impulse response. Moreover, unlike the TC and DC kernels, the previous property also involves the second order covariance lags of the impulse response.

VII. Higher-order extensions

Drawing inspiration from Section III we can define the TC kernel of order $\delta \in \mathbb{N}$ as

$$\mathcal{K}_{TC\delta} = \kappa_\delta(\mathcal{F}^\delta \mathcal{D}(\mathcal{F}^\delta)^\top)^{-1} \qquad (26)$$

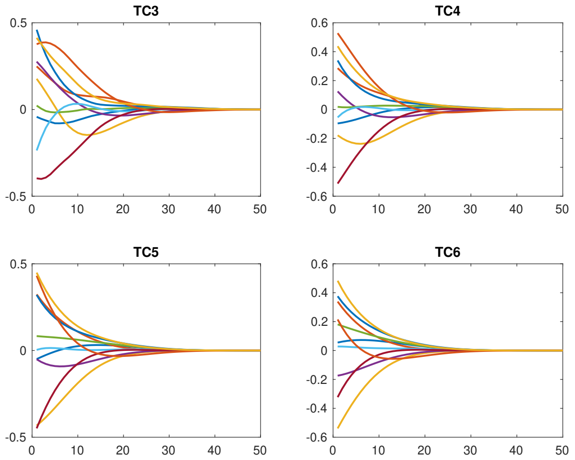

Fig. 8. Ten realizations of $\boldsymbol{g \sim \mathcal{N}(0, \lambda\mathcal{K}_{TC\delta})}$ with $\boldsymbol{\beta = 0.8}$, $\boldsymbol{\delta = 3, 4, 5, 6}$ and $\boldsymbol{\lambda = \|\mathcal{K}_{TC\delta}\|^{-1}}$.

where κ_δ is a suitable normalization constant. Here, $\eta = \beta$ with $0 < \beta < 1$. Figure 8 shows ten realizations of g using the TCδ kernel with $\beta = 0.8$ and for different values of δ. As expected, the larger δ is the more smoothness is induced on g.

Proposition 7.1: The inverse of $\mathcal{K}_{TC\delta}$ is a banded matrix of bandwidth δ, that is $[\mathcal{K}_{TC\delta}^{-1}]_{t,s} = 0$ for any $|t-s| > \delta$.

Proof: We prove the claim by induction. First, for $\delta = 1$ we have that TCδ is the standard TC and its inverse is tridiagonal, i.e. the claim holds. Assume that $\mathcal{K}_{TC\delta-1}^{-1}$ is a banded matrix of bandwidth $\delta - 1$. Then,

$$\mathcal{K}_{TC\delta}^{-1} = \frac{\kappa_{\delta-1}}{\kappa_\delta}\mathcal{F}\mathcal{K}_{TC\delta-1}^{-1}\mathcal{F}^\top.$$

Notice that $\mathcal{F} = \mathcal{I} - \mathcal{S}$ where $\mathcal{S}$ is the lower shift matrix and $\mathcal{I}$ the identity matrix, both infinite dimensional. Hence,

$$\begin{aligned} \mathcal{K}_{TC\delta}^{-1} &= \kappa_{\delta-1}\kappa_\delta^{-1}[\mathcal{K}_{TC\delta-1}^{-1} \\ &\quad + \mathcal{S}\mathcal{K}_{TC\delta-1}^{-1}\mathcal{S}^\top - \mathcal{K}_{TC\delta-1}^{-1}\mathcal{S}^\top - \mathcal{S}\mathcal{K}_{TC\delta-1}^{-1}]. \end{aligned} \qquad (27)$$

It is well known that premultiplying a matrix A by a lower shift matrix results in the elements of A being shifted downward by one position, with zeroes appearing in the top row. Thus, in view of (27), we have that $\mathcal{S}\mathcal{K}_{TC\delta-1}^{-1}\mathcal{S}^\top$ is a band matrix with bandwidth $\delta - 1$, while $\mathcal{K}_{TC\delta-1}^{-1}\mathcal{S}^\top + \mathcal{S}\mathcal{K}_{TC\delta-1}^{-1}$ and thus $\mathcal{K}_{TC\delta-1}$ are band matrices with bandwidth δ. ■

Also in this case one could try to find the closed form expression for $\mathcal{K}_{TC\delta}$, however its derivation is not straightforward from the case $\delta = 2$. On the other hand, we can define the corresponding finite dimensional kernel matrix $K_{TC\delta} \in \mathcal{S}_T$ as

$$[K_{TC\delta}]_{t,s} = [\mathcal{K}_{TC\delta}]_{t,s}, \quad t, s = 1 \ldots T.$$

Proposition 7.2: The finite dimensional kernel $K_{TC\delta}$ admits the following decomposition:

$$K_{TC\delta}^{-1} = \kappa_\delta^{-1} F_T^\delta D_T (F_T^\delta)^\top$$

where

$$D_T = \begin{bmatrix} D_{1,T} & 0 \\ 0 & B_T \end{bmatrix}$$
$$D_{1,T} = \mathrm{diag}(\beta^{-1}, \beta^{-2}, \dots \beta^{T-\delta}),$$

and B_T is a $\delta \times \delta$ matrix. Thus, $K_{TC\delta}^{-1}$ is banded of bandwidth δ.

Proof: Let $\mathcal{V}^{(j)} \in \mathbb{R}^{\infty \times T}$ denote a matrix whose first $j-1$ columns coincide with the null sequence and the remaining ones do not, thus $\mathcal{V}^{(T+1)}$ is the null matrix. We use $\sim$ to denote the equivalence relation $\mathcal{X} \sim \mathcal{Y}$ which means that $\mathcal{X} \in \mathbb{R}^{\infty \times T}$ and $\mathcal{Y} \in \mathbb{R}^{\infty \times T}$ have the first columns (in the same number) equal to the null sequence and the other ones do not. Thus, the latter induces a splitting of $\mathbb{R}^{\infty \times T}$ through the corresponding equivalence classes $[\mathcal{V}^{(j)}] = \{\mathcal{X} \in \mathbb{R}^{\infty \times T} \text{ s.t. } \mathcal{X} \sim \mathcal{V}^{(j)}\}$ with $1 \leq j \leq T+1$. In what follows, in order to ease the exposition (and thus with some abuse of notation) we use the symbol $=$ instead of $\sim$ in all the (submatrix) relations involving $\mathcal{V}^{(j)}$, with $j = 1 \dots T+1$.

First, notice that $\mathcal{F} = \mathcal{I} - \mathcal{S}$ and $F_T = I_T - S$ where $\mathcal{S}$ and S denote, respectively, the infinite and finite dimensional lower shift matrix. Recall that postmultiplying $\mathcal{V}^{(j)}$, with $1 \leq j \leq T$, by S results in the columns of $\mathcal{V}^{(j)}$ being shifted left by one position with a null sequence appearing in the last column position, thus

$$\mathcal{V}^{(j)} F_T = \mathcal{V}^{(j-1)}; \tag{28}$$

premultiplying $\mathcal{V}^{(j-1)}$, with $1 \leq j \leq T$, by $\mathcal{S}$ results in the rows of $\mathcal{V}^{(j-1)}$ being shifted downward by one position with a null row vector appearing in the first top row, thus

$$\mathcal{V}^{(j-1)} = \mathcal{F}\mathcal{V}^{(j-1)}. \tag{29}$$

Combining (28)-(29), we obtain

$$\mathcal{V}^{(j)} F_T = \mathcal{F}\mathcal{V}^{(j-1)}$$

and thus

$$\mathcal{F}^{-1}\mathcal{V}^{(j)} F_T = \mathcal{V}^{(j-1)}, \;\; 1 \leq j \leq T. \tag{30}$$

Then, we have

$$\mathcal{F}^{-1} \begin{bmatrix} I_T \\ \mathcal{V}^{(j)} \end{bmatrix} F_T = \begin{bmatrix} I_T \\ \mathcal{O} + \mathcal{F}^{-1}\mathcal{V}^{(j)} F_T \end{bmatrix} \tag{31}$$

where $\mathcal{O} \in \mathbb{R}^{\infty \times T}$ is a matrix whose last column is a sequence of ones, while the other columns are null sequencess, i.e. $\mathcal{O} = \mathcal{V}^{(T-1)}$. Accordingly, by (30)-(31) we have

$$\mathcal{F}^{-1} \begin{bmatrix} I_T \\ \mathcal{V}^{(j)} \end{bmatrix} F_T = \begin{bmatrix} I_T \\ \mathcal{V}^{(j-1)} \end{bmatrix}, \;\; 1 \leq j \leq T+1. \tag{32}$$

Notice that

$$\begin{aligned} K_{TC\delta} &= \begin{bmatrix} I_T & 0 \end{bmatrix} \mathcal{K}_{TC\delta} \begin{bmatrix} I_T \\ 0 \end{bmatrix} \\ &= \kappa_\delta \begin{bmatrix} I_T & 0 \end{bmatrix} (\mathcal{F}^{-\delta})^\top \mathcal{D}^{-1} \mathcal{F}^{-\delta} \begin{bmatrix} I_T \\ 0 \end{bmatrix}. \end{aligned}$$

Consider

$$\begin{aligned} Y :&= (F_T^\delta)^\top K_{TC\delta} F_T^\delta \\ &= \kappa_\delta \mathcal{W}_\delta^\top \mathcal{D}^{-1} \mathcal{W}_\delta \end{aligned}$$

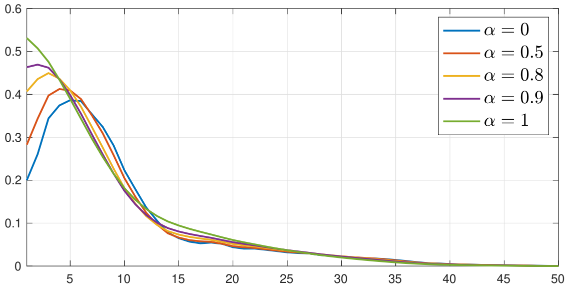

Fig. 9. One realization of $\boldsymbol{g \sim \mathcal{N}(0, \lambda\mathcal{K}_{DC3})}$ for different values of $\boldsymbol{\alpha}$. Here, $\boldsymbol{\beta = 0.8}$ and $\boldsymbol{\lambda = \|\mathcal{K}_{DC3}\|^{-1}}$.

where

$$\mathcal{W}_\delta = \mathcal{F}^{-\delta} \begin{bmatrix} F_T^\delta \\ 0 \end{bmatrix} = \mathcal{F}^{-\delta} \begin{bmatrix} F_T^\delta \\ \mathcal{V}^{(T+1)} \end{bmatrix}.$$

Then, it remains to prove that $Y = \kappa_\delta D_T^{-1}$. Indeed,

$$\begin{aligned} \mathcal{W}_\delta &= \mathcal{F}^{-(\delta-1)} \mathcal{F}^{-1} \begin{bmatrix} I_T \\ \mathcal{V}^{(T+1)} \end{bmatrix} F_T F_T^{\delta-1} \\ &= \mathcal{F}^{-(\delta-1)} \begin{bmatrix} I_T \\ \mathcal{V}^{(T)} \end{bmatrix} F_T^{\delta-1} \\ &= \dots = \begin{bmatrix} I_T \\ \mathcal{V}^{(T+1-\delta)} \end{bmatrix} \end{aligned}$$

where we exploited (32). Thus,

$$\begin{aligned} Y &= \kappa_\delta \begin{bmatrix} I_T & (\mathcal{V}^{(T+1-\delta)})^\top \end{bmatrix} \mathcal{D}^{-1} \begin{bmatrix} I_T \\ \mathcal{V}^{(T+1-\delta)} \end{bmatrix} \\ &= \kappa_\delta \begin{bmatrix} I_T & (\mathcal{V}^{(T+1-\delta)})^\top \end{bmatrix} \begin{bmatrix} D_{1,T}^{-1} & 0 \\ 0 & \tilde{\mathcal{D}}^{-1} \end{bmatrix} \begin{bmatrix} I_T \\ \mathcal{V}^{(T+1-\delta)} \end{bmatrix} \\ &= \kappa_\delta (D_{1,T}^{-1} + (\mathcal{V}^{(T+1-\delta)})^\top \tilde{\mathcal{D}}^{-1} \mathcal{V}^{(T+1-\delta)}) = \kappa_\delta D_T^{-1} \end{aligned}$$

where $\tilde{\mathcal{D}} = \mathrm{diag}(\beta^{T-\delta-1}, \beta^{T-\delta-2}, \dots)$. ■

It remains to design the DC kernel of oder δ connecting $\mathcal{K}_{TC\delta-1}$ and $\mathcal{K}_{TC\delta}$. Drawing inspiration from Section IV we define it as

$$\mathcal{K}_{DC\delta} = \kappa_\delta (\mathcal{F}_{\delta,\alpha} \mathcal{D} \mathcal{F}_{\delta,\alpha}^\top)^{-1} \tag{33}$$

where

$$\mathcal{F}_{\delta,\alpha} := (1-\alpha)\mathcal{F}^{\delta-1} + \alpha \mathcal{F}^\delta$$

and κ_δ is the normalization constant. Here, $\eta = [\beta \;\alpha]^\top$ with $0 < \beta < 1$ and $0 \leq \alpha \leq 1$. In Figure 9 we show a realization of the impulse response using (33) with $\delta = 3$ as a function of α; as expected, the degree of smoothness increases as α increases.

Proposition 7.3: The inverse of $\mathcal{K}_{DC\delta}$ is a banded matrix of bandwidth δ, that is $[\mathcal{K}_{DC\delta}^{-1}]_{t,s} = 0$ for any $|t-s| > \delta$.

Proof: First, for $\delta = 1$ $K_{DC\delta}$ is the standard DC kernel whose inverse is tridiagonal, i.e. the statement holds. Finally, notice that

$$\mathcal{F}_{\delta,\alpha} := \mathcal{F}((1-\alpha)\mathcal{F}^{\delta-2} + \alpha\mathcal{F}^{\delta-1}) = \mathcal{F}\mathcal{F}_{\delta-1,\alpha},$$

thus

$$\mathcal{K}_{DC\delta}^{-1} = \kappa_{\delta-1}\kappa_{\delta}^{-1}\mathcal{F}\mathcal{K}_{DC\delta-1}^{-1}\mathcal{F}^{\top}.$$

Accordingly, the remaining part of the proof is similar to the one of Proposition 7.1. ■

Also in this case the finite dimensional kernel $K_{DC\delta} \in \mathcal{S}_T$ is defined as

$$[K_{DC\delta}]_{t,s} = [\mathcal{K}_{DC\delta}]_{t,s}, \ \ t, s = 1 \dots T.$$

Proposition 7.4: The finite dimensional kernel $K_{DC\delta}$ admits the following decomposition:

$$K_{DC\delta}^{-1} = \kappa_{\delta} F_{\delta,\alpha,T} D_T (F_{\delta,\alpha,T})^{\top}$$

where

$$F_{\delta,\alpha,T} = (1-\alpha)F_T^{\delta-1} + \alpha F_T^{\delta}$$
$$D_T = \begin{bmatrix} D_{1,T} & 0 \\ 0 & B_T \end{bmatrix}$$
$$D_{1,T} = \mathrm{diag}(\beta^{-1}, \beta^{-2}, \dots \beta^{T-\delta})$$

and B_T is a $\delta \times \delta$ matrix; Thus, $K_{DC\delta}^{-1}$ is banded of bandwidth δ.

Proof: The proof is similar to the one of Proposition 7.2. ■

Finally, this extension can be applied also to the high-frequency (HF) kernel, see [29]:

$$[\mathcal{K}_{HF}]_{t,s} = (-1)^{|t-s|}\beta^{\max(t,s)} = (-1)^{|t-s|}[\mathcal{K}_{TC}]_{t,s}$$

where $0 < \beta < 1$. We define the high frequency kernel of oder $\delta \in \mathbb{N}$ as

$$[\mathcal{K}_{HF\delta}]_{t,s} = (-1)^{|t-s|}[\mathcal{K}_{TC\delta}]_{t,s}.$$

Moreover, we can define the high frequency diagonal-correlated (HC) kernel connecting HF$\delta - 1$ and HFδ as

$$[\mathcal{K}_{HC}]_{t,s} = (-1)^{|t-s|}[\mathcal{K}_{DC\delta}]_{t,s}.$$

It is straightforward to see that $\mathcal{K}_{HF\delta}^{-1}$ and $\mathcal{K}_{HC\delta}^{-1}$ are banded of bandwidth δ, as well as their finite dimensional matrices $K_{HF\delta}^{-1}$ and $K_{HC\delta}^{-1}$. It is possible to find the closed form expression for the Cholesky factor and the determinant of K_{HF2}^{-1} and K_{HC2}^{-1}. Finally, K_{HF2} and K_{HC2} are, respectively, the maximum entropy solution of a band extension problem similar to the ones introduced in Section VI.

VIII. Frequency analysis

An amplitude modulated kernel locally stationary (AMLS) kernel is a particular type of exponentially convex local stationary (ECLS) kernel $\mathcal{K} \in \mathcal{S}_{\infty}$ and it admits the following decomposition

$$[\mathcal{K}]_{t,s} = \beta^{\frac{t+s}{2}}[\mathcal{W}]_{t,s} \tag{34}$$

where $\mathcal{W} \in \mathcal{S}_{\infty}$ is a stationary kernel, i.e. the covariance function of a stationary process and thus $[\mathcal{W}]_{t,s} = [\mathcal{W}]_{t+k,s+k}$ for any $k \in \mathbb{N}$. Recall that TC, DC and SS are ECLS kernels. It is straightforward to see that TC2 and DC2 are ECLS kernel whose stationary parts are, respectively,

$$[\mathcal{W}_{TC2}]_{t,s} = 2\beta^{\frac{|t-s|}{2}+1} + (1-\beta)(1+|t-s|)\beta^{\frac{|t-s|}{2}}$$
$$[\mathcal{W}_{DC2}]_{t,s} = \frac{\beta^{\frac{|t-s|}{2}}(1-(1-\beta)\alpha^{|t-s|+1}) - \alpha^2\beta^{\frac{|t-s|}{2}+1}}{1-\alpha}.$$

Theorem 8.1: TCδ and DCδ kernels with $\delta > 2$ are ECLS, that is

$$\mathcal{K}_{TC\delta} = \beta^{\frac{t+s}{2}}[\mathcal{W}_{TC\delta}]_{t,s}, \ \ \mathcal{K}_{DC\delta} = \beta^{\frac{t+s}{2}}[\mathcal{W}_{DC\delta}]_{t,s}$$

where $\mathcal{W}_{TC\delta}$ and $\mathcal{W}_{DC\delta}$ are stationary kernels.

Proof: We only prove the claim for TCδ because the one for DCδ is similar. By (26), we have that

$$\mathcal{K}_{TC\delta} = \kappa_{\delta}\mathcal{X}^{\top}\mathcal{D}^{-1}\mathcal{X} \tag{35}$$

where $\mathcal{X} = \mathcal{F}^{-\delta} = (\mathcal{F}^{-1})^{\delta}$. Since $\mathcal{F}$ is lower triangular, Toeplitz and invertible, then by Lemma 3.1 we know that $\mathcal{F}^{-1}$ is lower triangular and Toeplitz. Accordingly, $\mathcal{X}$ is lower triangular and Toeplitz because it is given by a product of lower triangular and Toeplitz matrices. Hence, let

$$\mathcal{X} = \mathrm{Tpl}(x_1, x_2, x_3, \dots).$$

Moreover,

$$[\mathcal{X}]_{t,:} = [0 \ \dots \ 0 \ \underbrace{x_1}_{t\text{-th element}} \ x_2 \ x_3 \dots].$$

Taking into account (35), we have

$$\begin{aligned}[\mathcal{K}_{TC\delta}]_{t,s} &= \kappa_{\delta}[\mathcal{X}]_{t,:}\mathcal{D}^{-1}[\mathcal{X}]_{s,:}^{\top} \\ &= \kappa_{\delta}\sum_{k=1}^{\infty}\beta^{\max(t,s)+k-1}x_k x_{k+|t-s|} \\ &= \beta^{\frac{t+s}{2}}\underbrace{\kappa_{\delta}\sum_{k=1}^{\infty}\beta^{\frac{|t-s|}{2}+k-1}x_k x_{k+|t-s|}}_{:=[\mathcal{W}_{TC\delta}]_{t,s}}\end{aligned} \tag{36}$$

where we have exploited the fact that $\max(t,s) = (t+s)/2 + |t-s|/2$. It is straightforward to see that $\mathcal{W}_{TC\delta}$ is a stationary kernel. In view of (34) and (36), we conclude that TCδ is ECLS. ■

Although it is not immediate to derive the closed form expression for $W_{TC\delta}$ and $W_{DC\delta}$, we can compute them numerically:

$$[W_{TC\delta}]_{t,s} \approx \beta^{-\frac{t+s}{2}}[K_{TC\delta}]_{t,s}$$

and likewise for DCδ. Clearly, the larger T is, the better the approximation above is.

Therefore, it is interesting to compare the frequency content in their stationary parts. In doing that, we recall that

$$[W]_{t,s} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\phi(\vartheta)\cos(\vartheta(t-s))\mathrm{d}\vartheta$$

where $\phi(\vartheta)$, with $\vartheta \in [0, 2\pi]$, is the power spectral density of the (stationary) process. In order to compare SS with the others we need to choose $\gamma = \sqrt[3]{\beta}$ in (9); in this way the latter has the exponential part as in (34). Figure 10 shows the power

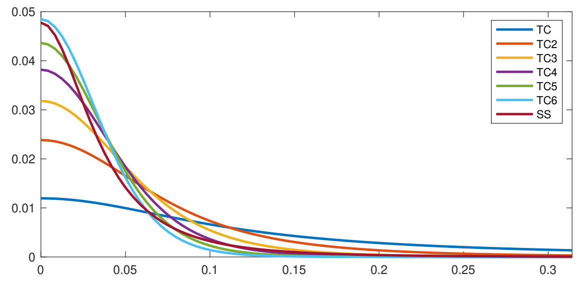

Fig. 10. Power spectral density of the stationary part of TC, TC$\boldsymbol{\delta}$, with $\boldsymbol{\delta = 2 \dots 6}$, and SS with $\boldsymbol{\beta = 0.8}$. All those power spectral densities are normalized to one in order to ease the comparison.

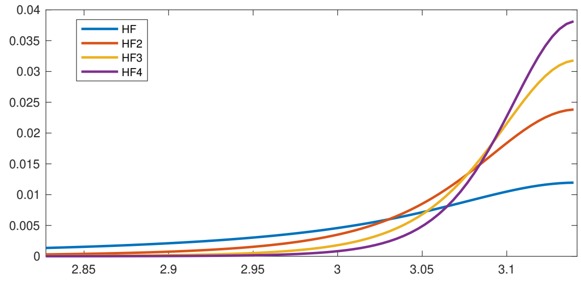

Fig. 11. Power spectral density of the stationary part of HF and HF$\boldsymbol{\delta}$, with $\boldsymbol{\delta = 2 \dots 4}$, with $\boldsymbol{\beta = 0.8}$. All those power spectral densities are normalized to one in order to ease the comparison.

spectral densities of the stationary part of TC, TCδ, with $\delta = 2 \dots 6$ and SS: the higher TCδ is, the more statistical power is concentrated for frequencies close to zero. The frequency content of SS is more similar to the one of TC2 than the one of TC. It is worth noting that we can plot also the power spectral density corresponding to DCδ. The latter smoothly changes from the one of TC$\delta - 1$, with $\alpha = 0$, to the one of TCδ, with $\alpha = 1$.

Finally, also HFδ and HCδ are ECLS kernels. Figure 11 shows the power spectral density of the stationary part of HF and HFδ for $\delta = 2 \dots 4$. The higher δ is, the more the statistical power is concentrated for frequencies close to π.

In order to understand the use of the TCδ kernel we consider a Monte Carlo study composed by 200 experiments. In each experiment the models and the data are generated likewise to the first Monte Carlo study of Section V, but the input u is a realization drawn from a Gaussian noise with band [0, 0.2]. We consider the following additional estimators for the impulse response:

- $\hat{g}_{TC\delta}$ is the estimator in (2) using the TCδ kernel (26) with $\delta = 2 \dots 6$.

Figure 12 shows the boxplot of FIT for the estimators: the higher δ is, the better the performance of the estimator is. In view of the fact that the true impulse responses in this Monte Carlo study are enough smooth, see Figure 3 (top), this study suggests the following design guideline. The more smooth the impulse response is expected, the higher the parameter δ should be chosen. Similar conclusions hold for the DCδ kernel.

IX. A System Theory Perspective

Since TC2 and DC2 are AMLS kernels, then the Gaussian proccess g_t with kernel function $\mathcal{K}_{TC2}$ or $\mathcal{K}_{DC2}$ can be understood as the output of linear time invariant (LTI) system, [21]. Indeed, the stationary part of the TC2 kernel $[\mathcal{W}_{TC2}]_{t,s}$ corresponds to a stationary stochastic process whose covariance function is

$$r(t) := 2\beta^{\frac{|t|}{2}+1} + (1-\beta)(1+|t|)\beta^{\frac{|t|}{2}}$$

and it is not difficult to see that the corresponding power spectral density is

$$\phi(\vartheta) = \frac{-\gamma z + \delta - \gamma z^{-1}}{(z - \sqrt{\beta})^2 (z^{-1} - \sqrt{\beta})^2}$$

where $\gamma = \sqrt{\beta}(1-\beta)^2$ and $\delta = 2(1-\beta)^2$. Then, by Proposition 4 in [21], process g_t with kernel TC2 can be understood as the output of the state space model

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ g_t &= Cx(t) + Du(t) \end{aligned}$$

where $u(t) = \sqrt{\beta}^t w(t)$ is the input, $w(t)$ is a zero-mean normalized white Gaussian noise process, the initial state $x(0) \sim \mathcal{N}(0, Q)$ is independent from $w(t)$ and Q is solution to the algebraic equation $Q = \beta^{-1}(AQA^T + BB^T)$. In this case, it is not difficult to see that

$$\begin{aligned} &A = \sqrt{\beta} \begin{bmatrix} \sqrt{\beta} & 1 \\ 0 & \sqrt{\beta} \end{bmatrix}, \; B = \sqrt{\beta} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ &C = \sqrt{\frac{\gamma}{p}} \begin{bmatrix} 2\sqrt{\beta} - p & 2\sqrt{\beta} - p - \beta + \sqrt{\beta} p \end{bmatrix}, \; D = \sqrt{\frac{\gamma}{p}} \end{aligned} \tag{37}$$

where $p = (1 - \sqrt{1-\beta})/\sqrt{\beta}$. It is worth noting that the proposed kernel has the same state space dimension of the SS kernel, however the eigenvalues of the state transition matrix are different. In the TC2 kernel there is only one eigenvalue whose index is two, while in the SS kernel there are two different eigenvalues and their index is one.

In a similar way, it is possible to prove that if process g_t is characterized by kernel DC2, then its state space representation is with

$$\begin{aligned} &A = \sqrt{\beta} \begin{bmatrix} \sqrt{\beta} & 0 \\ 0 & \alpha\sqrt{\beta} \end{bmatrix}, \; B = \sqrt{\beta} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ &C = \frac{\mu\sqrt{\beta}}{1-\alpha} \begin{bmatrix} 1 & -\alpha^2 \end{bmatrix}, \; D = \mu \end{aligned} \tag{38}$$

$\mu = \sqrt{(1-\alpha^2\beta)(1-\alpha\beta)(1-\beta)}$. In particular, the dimension of the state space realization is larger than the one of the DC kernel.

X. Conclusions

We have introduced a second-order extension to TC and DC kernels called TC2 and DC2, respectively. The latter induce more smoothness than the former. This idea can be also extended to higher-orders. We also have introduced a

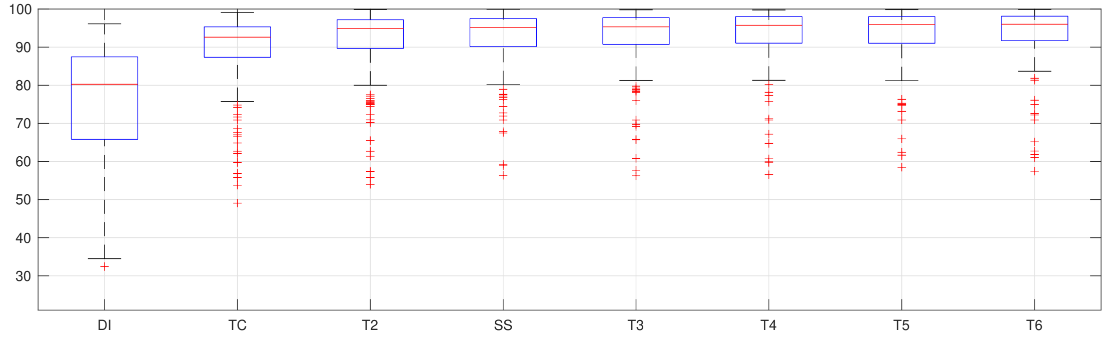

Fig. 12. Impulse response fit in the Monte Carlo study composed by 200 experiments.

generalized-correlated (GC) kernel which incorporates the DI, DC, TC kernels, i.e. the most popular kernels in system identification, and the DC2 and TC2 kernels. We have derived the closed form expression for the determinant and the Cholesky factorization of the inverse matrix of TC2, DC2 and GC. The log-likelihood calculation with the TC2 kernel is straightforward to implement efficiently in Matlab, and the asymptotic complexity matches that of the same calculations with kernels such as the TC and SS kernels. Numerical experiments showed that TC2 and SS kernels produce similar performances for estimating the impulse response, but the search of the optimal hyperparameters through marginal likelihood is more efficient (in terms of execution time) using TC2 than SS. The runtime performance of these algorithms can be improved with the use of code written in C or C++, however further investigation is required in order to draw some conclusion. Finally, we have also shown that these new kernels are exponentially convex local stationary and thus it is possible to understand easily their frequency properties.

[1] L. Ljung, *System Identification: Theory for the User*. New Jersey: Prentice Hall, 1999.

[2] T. Söderström and P. Stoica, *System Identification*. Hemel Hempstead, UK: Prentice-Hall International, 1989.

[3] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, Dec. 1974.

[4] G. Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics*, vol. 6, pp. 461–464, Mar. 1978.

[5] G. Pillonetto and G. De Nicolao, “A new kernel-based approach for linear system identification,” *Automatica*, vol. 46, 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. 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.

[8] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[9] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[10] M. Zorzi and A. Chiuso, “Sparse plus low rank network identification: A nonparametric approach,” *Automatica*, vol. 76, pp. 355–366, 2017.

[11] A. Chiuso and G. Pillonetto, “A Bayesian approach to sparse dynamic network identification,” *Automatica*, vol. 48, no. 8, pp. 1553–1565, 2012.

[12] M. Zorzi and A. Chiuso, “A Bayesian approach to sparse plus low rank network identification,” in *Proceedings of the IEEE Conference on Decision and Control*, (Osaka), pp. 7386–7391, 2015.

[13] M. Zorzi, “Nonparametric identification of Kronecker networks,” *Automatica*, vol. 145, p. 110518, 2022.

[14] M. Zorzi, “Autoregressive identification of Kronecker graphical models,” *Automatica*, vol. 119, p. 109053, 2020.

[15] F. P. Carli, A. Chiuso, and G. Pillonetto, “Efficient algorithms for large scale linear system identification using stable spline estimators,” *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 119–124, 2012.

[16] T. Chen, M. S. Andersen, B. Mu, F. Yin, L. Ljung, and S. J. Qin, “Regularized LTI system identification with multiple regularization matrix,” *Ifac-papersonline*, vol. 51, no. 15, pp. 180–185, 2018.

[17] T. Chen and M. S. Andersen, “On semiseparable kernels and efficient implementation for regularized system identification and function estimation,” *Automatica*, vol. 132, p. 109682, 2021.

[18] 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.

[19] F. P. Carli, “On the maximum entropy property of the first-order stable spline kernel and its implications,” in *IEEE Conference on Control Applications (CCA)*, pp. 409–414, 2014.

[20] F. P. Carli, T. Chen, and L. Ljung, “Maximum entropy kernels for system identification,” *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1471–1477, 2017.

[21] T. Chen, “On kernel design for regularized LTI system identification,” *Automatica*, vol. 90, pp. 109–122, 2018.

[22] F. Dinuzzo, “Kernels for linear time invariant system identification,” *SIAM Journal on Control and Optimization*, vol. 53, no. 5, pp. 3299–3317, 2015.

[23] M. Zorzi and A. Chiuso, “The harmonic analysis of kernel functions,” *Automatica*, vol. 94, pp. 125–137, 2018.

[24] M. Zorzi, “A new kernel-based approach for spectral estimation,” in *European Control Conference (ECC)*, pp. 534–539, 2020.

[25] T. Chen and L. Ljung, “On kernel structures for regularized system identification (i): a machine learning perspective,” *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 1035–1040, 2015.

[26] T. Chen and L. Ljung, “On kernel structures for regularized system identification (ii): A system theory perspective,” *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 1041–1046, 2015.

[27] A. Marconato, M. Schoukens, and J. Schoukens, “Filter-based regularisation for impulse response modelling,” *IET Control Theory & Applications*, vol. 11, no. 2, pp. 194–204, 2017.

[28] G. Pillonetto, A. Chiuso, and G. De Nicolao, “Regularized estimation of sums of exponentials in spaces generated by stable spline kernels,” in *Proceedings of the 2010 American Control Conference*, 2010.

[29] G. Pillonetto and G. De Nicolao, “Kernel selection in linear system identification part i: A gaussian process perspective,” in *50th IEEE Conference on Decision and Control and European Control Conference*, pp. 4318–4325, 2011.

[30] P. Jorgesen, K. Kornelson, and K. Shuman, *Iterated Function Systems, Moments, and Transformations of Infinite Matrices*. American Mathematical Society, 2011.

[31] N. J. Ford, D. V. Savostyanov, and N. L. Zamarashkin, “On the decay of the elements of inverse Triangular toeplitz matrices,” *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 4, pp. 1288–1302, 2014.

[32] M. S. Andersen and T. Chen, “Smoothing splines and rank structured matrices: Revisiting the spline kernel,” *SIAM Journal on Matrix Analysis and Applications*, vol. 41, no. 2, pp. 389–412, 2020.

[33] Y. Fujimoto, “Efficient implementation of kernel regularization based on ADMM,” in *SYSID*, 2021.

[34] A. P. Dempster, “Covariance selection,” *Biometrics*, vol. 28, pp. 157–175, 1972.

[35] T. Chen, T. Ardeshiri, F. P. Carli, A. Chiuso, L. Ljung, and G. Pillonetto, “Maximum entropy properties of discrete-time first-order stable spline kernel,” *Automatica*, vol. 66, pp. 34–38, 2016.

[36] T. Chen, “Continuous-time dc kernel—a stable generalized first-order spline kernel,” *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4442–4447, 2018.

[37] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci, “A maximum entropy solution of the covariance extension problem for reciprocal processes,” *IEEE Transactions on Automatic Control*, vol. 56, no. 9, pp. 1999–2012, 2011.

[38] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci, “An efficient algorithm for maximum entropy extension of block-circulant covariance matrices,” *Linear Algebra and its Applications*, vol. 439, no. 8, pp. 2309–2329, 2013.

Mattia Zorzi received the M.S. degree in Automation Engineering and the Ph.D. degree in Information Engineering from the University of Padova, Padova, Italy, in 2009 and 2013, respectively. He held Postdoctoral appointments with the Department of Electrical Engineering and Computer Science, University of Liege, Liege, Belgium, and with the Human Inspired Technology Research Centre, University of Paodva, Padova, Italy. He held visiting positions with the Department of Electrical and Computer Engineering, University of California, Davis, USA, and with the Department of Engineering, University of Cambridge, Cambridge, U.K., in 2011 and 2013-2014, respectively. He is currently an Associate Professor with the Department of Information Engineering, University of Padova. His current research interests include machine learning, robust estimation, identification theory.

Dr. Zorzi has been an Associate Editor of Automatica since 2021 and IEEE Control Systems Letters since 2019. He serves as an Associate Editor on the IEEE Control System Society Conference Editorial Board since 2017 and the EUCA Conference Editorial Board since 2020. He is an IEEE Senior member and a member of the IFAC Technical Committee on Modelling, Identification and Signal Processing.