The Harmonic Analysis of Kernel Functions

Mattia Zorzi\textsuperscript{a}, Alessandro Chiuso\textsuperscript{a}

\textsuperscript{a}Dipartimento di Ingegneria dell’Informazione, Università degli studi di Padova, via Gradenigo 6/B, 35131 Padova, Italy

Abstract

Kernel-based methods have been recently introduced for linear system identification as an alternative to parametric prediction error methods. Adopting the Bayesian perspective, the impulse response is modeled as a non-stationary Gaussian process with zero mean and with a certain kernel (i.e. covariance) function. Choosing the kernel is one of the most challenging and important issues. In the present paper we introduce the harmonic analysis of this non-stationary process, and argue that this is an important tool which helps in designing such kernel. Furthermore, this analysis suggests also an effective way to approximate the kernel, which allows to reduce the computational burden of the identification procedure.

Key words: System identification, Kernel-based methods, Power spectral density, Random features.

1 Introduction

Building upon the theory of reproducing kernel Hilbert spaces and statistical learning, kernel-based methods for linear system identification have been recently introduced in the system identification literature, see Pillonetto & De Nicolao (2010); Pillonetto et al. (2011); Chiuso & Pillonetto (2012); Chen et al. (2012); Pil- lonetto et al. (2014); Zorzi & Chiuso (2017, 2015); Chiuso (2016); Fraccaroli et al. (2015).

These methods, framed in the context of Prediction Error Minimization, differ from classical parametric methods Ljung (1999); Söderström & Stoica (1989), in that models are searched for in possibly infinite dimensional model classes, described by Reproducing Kernel Hilbert Spaces (RKHS). Equivalently, in a Bayesian framework, models are described assigning as prior a Gaussian distribution; estimation is then performed following the prescription of Bayesian Statistics, combining the “prior” information with the data in the posteriors distribution. Choosing the covariance function of the prior distribution, or equivalently the Kernel defining the RKHS, is one of the most challenging and important issues. For instance the prior distribution could reflect the fact that the system is Bounded Input Bounded Output (BIBO) stable, its impulse response possibly smooth and so on (Pillonetto et al. 2014).

Within this framework, the purpose of this paper is to discuss the properties of certain kernel choices from the point of view of Harmonic Analysis of stationary processes. The latter is well defined for stationary processes (Lindquist & Picci 2015, Chapter 3). In particular, it defines as Power Spectral Density (PSD) the function describing how the statistical power is distributed over the frequency domain. In this paper, we extend this analysis for a particular class of non-stationary processes modeling impulse responses of marginally stable systems. Accordingly, we define as Generalized Power Spectral Density (GPSD) the function describing how the statistical power is distributed over the decay rate-frequency domain.

Under the assumption that the prior density is Gaussian, the probability density function (PDF) of the prior is linked to the GPSD. The main difference is that while the former is defined over an infinite dimensional space (the underlying RKHS $\mathcal{H}_K$), the latter is defined over the bidimensional decay rate-frequency space. As a consequence, the latter is simple to depict but also to interpret from an engineering point of view. We show experimentally that, over the class of second-order linear systems, the two provide similar information. This class is important because: 1) it contains the simplest systems that exhibit oscillations and overshoot; 2) second order systems are building block of higher order systems and, as such, understanding second order systems helps understanding higher ones. Furthermore, for a special case...
The system identification problem can be frased as that of estimating the impulse response \( \{g(t)\}_{t \in \mathbb{N}} \) from the given data record

\[
Z^N = \{u(1), y(1) \ldots u(N), y(N)\}.
\]

In the Gaussian process regression framework, \( g(t) \) is modeled as a zero-mean discrete time Gaussian process with kernel (covariance) function \( K(t, s) := \mathbb{E}[g(t)g(s)] \). The minimum variance estimator of \( g(t) \) is given by its posterior mean given \( Z^N \), \((\text{Pillonetto } \& \text{ De Nicolao} \ 2010)\). It is clear that the posterior highly depends on the kernel functions. Accordingly, the most challenging part of this system identification procedure is to design \( K \) so that the posterior has some desired properties.

Similarly, in the continuous time case, \( \{g(t)\}_{t \in \mathbb{R}_+} \) is a zero-mean continuous time Gaussian process with kernel function \( K(t, s) := \mathbb{E}[g(t)g(s)] \), with \( t, s \in \mathbb{R}_+ \). In what follows Gaussian processes (both discrete time and continuous time) are always understood with zero-mean.

### 3 Harmonic Analysis: continuous time case

It is well known that the impulse response of a finite dimensional LTI stable (or marginally stable) system can be written as a linear combination of decaying sinusoids \(^\dagger\) (i.e. modes)

\[
g_a(t) = \sum_{l=1}^{N} |c_l| e^{\alpha_l t} \cos(\omega_l t + \zeta_l), \quad t \in \mathbb{R}_+ \tag{2}
\]

where \( \alpha_l \in \mathbb{R}_- \cup \{0\} \) and \( \omega_l \in \mathbb{R} \) are, respectively, the decay rate and the angular frequency of the \( l \)-th damped oscillation, and \( c_l \in \mathbb{C} \). Adopting the Bayesian perspective, \( g(t) \) is modeled as a Gaussian process where the coefficients \( c_l \) are zero mean complex Gaussian random variables such that

\[
\mathbb{E}[c_l c_{l'}^\ast] = \phi_l \delta_{l-l'}, \\
\mathbb{E}[c_l c_{l'}] = 0
\]

\(^\dagger\) For simplicity we exclude here the case of multiple eigenvalues.
with \( l, l' = 1 \ldots N \) and \( \phi_l \geq 0 \). For convenience, we rewrite (2) as

\[
g_a(t) = \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\phi} c_{ik} e^{\alpha t} \cos(\omega_k t + \angle c_{ik})
\]

(3)

that is \((\alpha, \omega_k)\) belongs to a \(N_\alpha \times N_\phi\) grid contained in \(\mathbb{R}_+ \cup \{0\} \times \mathbb{R}\) and \(c_{ik}\) is a complex Gaussian random variable such that

\[
\mathbb{E}[c_{ik} e^{-i k}] = \phi_{ik} \delta_{k-k'}
\]

\[
\mathbb{E}[c_{ik} c_{i'k'}] = 0
\]

with \(\phi_{ik} \geq 0\). It is then natural to generalize (3) as an infinite “dense” sum of decaying sinusoids \(^{[2]}\)

\[
g(t) = \int_{-\infty}^{0} \int_{-\infty}^{\infty} |c(\alpha, \omega)| e^{\alpha t} \cos(\omega t + \angle c(\alpha, \omega))d\omega d\alpha
\]

(4)

where \(c(\alpha, \omega)\) is a bidimensional complex Gaussian process\(^{[2]}\), hereafter called generalized Fourier transform of \(g(t)\), such that

\[
\mathbb{E}[c(\alpha, \omega)c(\alpha', \omega')] = \phi(\alpha, \omega) \delta(\alpha - \alpha') \delta(\omega - \omega')
\]

\[
\mathbb{E}[c(\alpha, \omega)c(\alpha', \omega')] = 0
\]

where \(\phi(\alpha, \omega)\) is a nonnegative function on \(\mathbb{R}_+ \cup \{0\} \times \mathbb{R}\) such that \(\phi(\alpha, \omega) = \phi(\alpha, -\omega)\) and \(\delta(\cdot)\) denotes the Dirac delta function.

**Proposition 1** Let \(K(t, s)\) be the kernel function of \(g(t)\) in (4) then,

\[
K(t, s) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{\infty} \phi(\alpha, \omega) e^{\alpha(t+s)} \cos(\omega(t-s))d\omega d\alpha.
\]

(5)

Formula (5) is the harmonic representation of the covariance function of the non-stationary process (4). We refer to \(\phi(\alpha, \omega)\) as generalized power spectral density (GPSD) of \(g(t)\). The latter describes how the “statistical power” of \(g(t)\) (which depends on \(t\)) is distributed over the decay rate-angular frequency space \(\mathbb{R}_+ \cup \{0\} \times \mathbb{R}\) according to:

\[
\mathbb{E} \left[ g(t)^2 \right] = K(t, t) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{\infty} \phi(\alpha, \omega) e^{2\alpha t} d\omega d\alpha.
\]

In particular, when \(t = 0\), the exponential term \(e^{2\alpha t}\) disappears, so that the statistical power of \(g(0)\) is obtained as in the stationary case:

\[
\mathbb{E} [g(0)^2] = K(t, t) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{\infty} \phi(\alpha, \omega) d\omega d\alpha.
\]

It is worth noting that the GPSD can be understood as a function in \(\mathbb{C}\), that is \(\phi(s)\) with \(s = \alpha + j\omega\), and its support is the left half-plane. Now we show that the harmonic representation (5) describes many kernel functions used in system identification.

### 3.1 Stationary kernels

The special case of stationary processes is recaptured when \(\phi(\alpha, \omega) = \delta(\alpha) \phi_1(\omega)\), with \(\phi_1(\omega) = \phi_1(-\omega)\) nonnegative function; in fact, under this assumption, we have

\[
K(t, s) = K_1(t-s) = \frac{1}{2} \int_{-\infty}^{\infty} \phi_1(\omega) \cos(\omega(t-s))d\omega
\]

(7)

which is a stationary kernel and \(\phi_1(\omega)\) is the corresponding power spectral density (PSD). Note that, (7) is the usual harmonic representation of a stationary covariance function which is also exploited in spectral estimation problems (Zorzi, 2015b,a). In this case the stationary process \(g(t)\) is an infinite “dense” sum of sinusoids

\[
g(t) = \int_{-\infty}^{\infty} |c(\omega)| \cos(\omega t + \angle c(\omega))d\omega
\]

which, more rigorously, should be written in terms of the spectral measure \(C(\omega)\)

\[
g(t) = \int_{-\infty}^{\infty} e^{j\omega t} dC(\omega).
\]

Since \(\phi_1(\omega)\) is an even function, we can rewrite it as

\[
\tilde{\phi}_1(\omega) = \tilde{\phi}_1(\omega) + \tilde{\phi}_1(-\omega)
\]

with \(\tilde{\phi}_1(\omega)\) nonnegative function. For instance, choosing \(\tilde{\phi}_1(\omega)\) as one of the following:

\[
\tilde{\phi}_L(\omega) = \frac{1}{\pi} \frac{\beta}{\beta^2 + (\omega - \omega_0)^2}
\]

\[
\tilde{\phi}_C(\omega) = \frac{1}{2\beta} e^{-|\omega - \omega_0|}
\]

\[
\tilde{\phi}_G(\omega) = \frac{1}{\sqrt{2\pi\beta}} e^{-|\omega - \omega_0|^2/\beta}
\]

(8)
we obtain

\[
K_L(t-s) = e^{-\beta|t-s|}\cos(\omega_0(t-s))
\]
\[
K_C(t-s) = \frac{1}{1 + \beta^2(t-s)^2}\cos(\omega_0(t-s))
\]
\[
K_G(t-s) = e^{-\frac{\pi}{2}(t-s)^2}\cos(\omega_0(t-s))
\]

where \(\omega_0\) denotes the angular frequency for which \(\hat{\phi}\) takes the maximum and \(\beta\) is proportional to the bandwidth. Setting \(\omega_0 = 0\) we obtain, respectively, the Laplacian, Cauchy and Gaussian kernel [Rasmussen & Williams 2006]. In particular, the latter is widely used in robotics for the identification of the inverse dynamic [Romeres et al., 2016].

### 3.2 Exponential Convex Locally Stationary (ECLS) kernels

A generalization of stationary kernels, introduced by Silverman [1957], is the so-called class of Exponentially Convex Locally Stationary (ECLS) kernels; this is obtained postulating a separable structure for the GPSD \(\phi(\alpha, \omega) = \phi_1(\omega)\phi_2(\alpha)\), with \(\phi_1(\omega)\) and \(\phi_2(\alpha)\) nonnegative functions, so that the kernel \(K(t,s)\) inherits the decomposition

\[
K(t,s) = K_1(t-s)K_2(t+s)
\]

where \(K_1\) has been defined in (7) and

\[
K_2(t+s) = \int_{-\infty}^{0} \phi_2(\alpha)e^{\alpha(t+s)}d\alpha.
\]

In the case that \(\phi_2(\alpha) = \delta(\alpha - \alpha_0)\) with \(\alpha_0 \in \mathbb{R}\) we obtain the ECLS kernel [Chen & Ljung, 2015b; 2016]:

\[
K(t,s) = e^{\alpha_0(t+s)}K_1(t-s).
\]

Specializing \(K_1(t-s) = K_L(t-s)\) with \(\omega_0 = 0\), in (9) we obtain the so-called diagonal-correlated (DC) kernel. Furthermore, for suitable choices of \(\phi_1(\omega)\) in (9), one can obtain the stable-spline (SS), the diagonal (DI) and the tuned-correlated kernel (TC), see [Chen et al., 2012].

### 3.3 Integrated kernels

Consider the GPSD

\[
\phi(\alpha, \omega) = 2\frac{-\alpha}{\pi|\omega^2 + \alpha^2|}\mathbf{1}_{[\alpha_m/2, \alpha_M/2]}(\alpha)
\]

where

\[
\mathbf{1}_{[\alpha_m/2, \alpha_M/2]}(\alpha) = \begin{cases} 1, & \alpha \in [\alpha_m/2, \alpha_M/2] \\ 0, & \text{otherwise} \end{cases}
\]

with \(\alpha_m < \alpha_M < 0\). Then, the corresponding kernel function is

\[
K(t,s) = \int_{-\infty}^{0} \frac{1}{\alpha_m}e^{\alpha(t+s)}\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega - \frac{\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)
\]

with \(\alpha_m < \alpha_M < 0\). Then, the corresponding kernel function is

\[
K(t,s) = \int_{\alpha_m}^{\alpha_M} \frac{1}{\alpha_m}e^{\alpha(t+s)}\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega - \frac{\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)\left(\frac{-\alpha}{\pi(\omega^2 + \alpha^2)}\cos(\omega(t-s))d\omega\right)
\]

which is similar to the integrated TC kernel introduced in Pillonetto et al. [2016]. In general, taking

\[
\phi(\alpha, \omega) = \phi_1(\omega)\mathbf{1}_{[\alpha_m, \alpha_M]}(\alpha),
\]

where we made explicit the dependence of \(\phi_1\) upon \(\alpha\), we have

\[
K(t,s) = \int_{\alpha_m}^{\alpha_M} e^{\alpha(t+s)}K_1(t-s; \alpha)d\alpha
\]

where \(K_1(t-s; \alpha)\) is the stationary kernel corresponding to \(\phi_1(\omega; \alpha)\). Notice that, kernel (11) is obtained by integrating the ECLS kernel \(e^{\alpha(t+s)}K_1(t-s; \alpha)\) over the interval \([\alpha_m, \alpha_M]\), which justifies the name “integrated”. Another possible way to construct an integrated kernel is choosing

\[
\phi(\alpha, \omega) = \phi_1(\omega)\mathbf{1}_{[\alpha_m, \alpha_M]}(\alpha)
\]

where \(\phi_1\) does not depend on \(\alpha\). Then, the corresponding kernel is

\[
K(t,s) = K_1(t-s)\frac{e^{\alpha_M(t+s)} - e^{\alpha_m(t+s)}}{t+s},
\]

which is an ECLS kernel with

\[
K_2(t+s) = \frac{e^{\alpha_M(t+s)} - e^{\alpha_m(t+s)}}{t+s}.
\]

### 4 Harmonic analysis: discrete time case

Following the same argumentations of Section 3, a Gaussian process describing a discrete time causal impulse response can be understood as

\[
g(t) = \int_{0}^{1} \int_{-\pi}^{\pi} |c(\lambda, \theta)|\lambda^t \cos(\theta(t-s))d\theta d\lambda, \quad t \in \mathbb{N}
\]

\[\text{See Remark 3 for more details.}\]
where \( c(\lambda, \vartheta) \) is the generalized Fourier transform of \( g(t) \), \( \lambda \) is the decay rate and \( \vartheta \) is the normalized angular frequency. Moreover, the kernel function of \( g(t) \) admits the following harmonic representation

\[
K(t, s) = \frac{1}{2} \int_0^1 \int_{-\pi}^{\pi} \phi(\lambda, \vartheta) \lambda^{t+s} \cos(\vartheta(t - s)) d\vartheta d\lambda
\]

(12)

and \( \phi(\lambda, \vartheta) = \phi(\lambda, -\vartheta) \) is the GPSD of \( g(t) \). The latter is a nonnegative function over the decay rate-normalized angular frequency space \([0, 1] \times [-\pi, \pi]\). Also in this case the GPSD can be understood as a function in \( \mathbb{C} \), that is \( \phi(z) \) with \( z = \lambda e^{j\vartheta} \), and its support is the unit circle.

In system identification, it is usual to design a kernel function for a continuous time Gaussian process \( g_c(t) \), \( t \in \mathbb{R}_+ \). Then, the “discrete time” kernel is obtained by sampling the “continuous time” kernel with a certain sampling time \( T \). The latter corresponds to the discrete time Gaussian process \( g_d(k) \), \( k \in \mathbb{N} \), obtained by sampling \( g_c(t) \).

**Proposition 2** Let \( \phi_c(\alpha, \omega) \) and \( \phi_d(\lambda, \vartheta) \) be GPSD of \( g_c(t) \), \( t \in \mathbb{R}_+ \), and \( g_d(k) \), \( k \in \mathbb{N} \), respectively. Then,

\[
\phi_d(\lambda, \vartheta) = \frac{1}{\lambda T^2} \sum_{k \in \mathbb{Z}} \phi_c(T^{-1} \log \lambda, T^{-1} (\vartheta - 2\pi k)).
\]

(13)

Accordingly, if the continuous time GPSD is such that

\[
\phi_c(\alpha, \omega) \approx 0, \quad |\omega| > \frac{\pi}{T}
\]

then its discretized version is such that

\[
\phi_d(\lambda, \vartheta) \approx \frac{1}{\lambda T^2} \phi_c(T^{-1} \log \lambda, T^{-1} \vartheta).
\]

**Remark 3** Discretizing (10) with \( T = 1 \), we obtain

\[
K_{\text{ITC}}(t, s) = \frac{\lambda_M^{\max\{t, s\}} - \lambda_m^{\max\{t, s\}}}{2 \max\{t, s\}},
\]

with \( 0 < \lambda_m < \lambda_M < 1 \) and \( t, s \in \mathbb{N} \). However, the integrated TC kernel derived in [Pillonetto et al. 2016] is slightly different:

\[
\hat{K}_{\text{ITC}}(t, s) = \frac{\lambda_M^{\max\{t, s\}+1} - \lambda_m^{\max\{t, s\}+1}}{\max\{t, s\} + 1}.
\]

(15)

Indeed, the latter has been derived by discretizing the TC kernel, and then the integration along the decay rate has been performed in the discrete domain. Note that, the integration along the decay rate in \( K_{\text{ITC}} \) is uniform, while in \( \hat{K}_{\text{ITC}} \) such integration is warped according to (13).

### 4.1 Filtered kernels

In order to account for high frequency components of predictor impulse responses, [Pillonetto et al. 2011] have introduced a class of priors obtained as filtered versions of stable spline kernels, using second order filters of the form:

\[
F(z) = \frac{z^2}{(z - \rho_0 e^{j\vartheta_0})(z - \rho_0 e^{-j\vartheta_0})}
\]

with \( |\rho_0| < 1 \) and \( \vartheta_0 \in [-\pi, \pi] \). The latter filter is fed by a Gaussian process \( \tilde{g}(t) \) with kernel function \( \tilde{K}(t, s) \), for instance in the class of “stable-spline” kernels (see [Pillonetto et al. 2010]), which have most of the statistical power (6) concentrated around \( \vartheta = 0 \); in this paper \( \tilde{K} \) is chosen as TC kernel. \( F(z) \) plays the role of a shaping filter, which concentrates the statistical power of the stationary part around \( \vartheta = \vartheta_0 \). It is not difficult to see that the kernel function of \( \tilde{g}(t) \) is

\[
K(t, s) = \sum_{i=1}^{t} \sum_{m=1}^{s} f_{t-i} f_{s-m} \tilde{K}(l, m)
\]

where \( \{f_t\}_{t \in \mathbb{N}} \) is the impulse response of the filter \( F(z) \), i.e. \( F(z) = \sum_{s=0}^{\infty} f_s z^{-s} \). Assuming that \( \tilde{K}(t, s) \) admits the harmonic representation

\[
\tilde{K}(t, s) = \frac{1}{2} \int_0^1 \int_{-\pi}^{\pi} \phi(\lambda, \vartheta) \lambda^{t+s} \cos(\vartheta(t - s)) d\vartheta d\lambda
\]

(16)

then we have

\[
K(t, s) = \frac{1}{2} \int_0^1 \int_{-\pi}^{\pi} \phi(\lambda, \vartheta) f(\lambda, \vartheta, t, s) d\vartheta d\lambda
\]

(17)

Indeed, the latter has been derived by discretizing the TC kernel, and then the integration along the decay rate has been performed in the discrete domain. Note that, the integration along the decay rate in \( K_{\text{ITC}} \) is uniform, while in \( \hat{K}_{\text{ITC}} \) such integration is warped according to (13).

### 5 Probability Density Function of Gaussian Processes and their GPSD

Given a discrete time Gaussian process \( g(t) \) with kernel \( K(t, s) \), in this Section we shall study the relation between the associated Gaussian probability distribution and the Generalized Spectral Density introduced
We consider the ECLS kernels applied to the decay rate, which relates to band (i.e. the phase of energy around specific frequency bands are well suited that indeed GPSDs whose stationary part concentrate in (19)). Similarly, the same applies to the decay rate, which relates to \(|p|\) in (19).

5.1 ECLS kernels

We consider the ECLS kernels

\[
\begin{align*}
K^\text{ECLS}_L(t,s) &= e^{\omega_0(t+s)} K_L(t-s) \\
K^\text{ECLS}_C(t,s) &= e^{\omega_0(t+s)} K_C(t-s) \\
K^\text{ECLS}_G(t,s) &= e^{\omega_0(t+s)} K_G(t-s)
\end{align*}
\]

where \(\alpha_0 = -0.1, \beta = 0.1\) and \(\omega_0 = 3\pi/5\). In Figure 1 (top) we show the corresponding PSDs of the stationary part. Then, we discretize these kernels with \(T = 1\). The corresponding GPSDs have as support a circle with radius \(r_0 = e^{\omega_0} T = 0.9\) centered in zero, see left picture of Figure 2. Since condition (14) holds, the shape of the discretized versions essentially reflect the continuous time version in Figure 1 (top). This means that the statistical power at time \(t = 0\) is concentrated in a neighborhood of \(z_0 = \lambda_0 e^{j\theta_0}\), with \(\theta_0 = \omega_0\), along the normalized angular frequency domain. In particular, the one of \(K^\text{ECLS}_C\) is more concentrated in \(z_0\) than the one of \(K^\text{ECLS}_L\) and the latter is more concentrated than the one of \(K^\text{ECLS}_G\). Finally we also consider the filtered kernel with \(p_0 = 0.93\) and \(\theta_0 = 3\pi/5\). The corresponding PDFs are depicted in Figure 3(a). For all these kernels the impulse responses taking high probability have a pole in a neighborhood of \(p_0 = \lambda_0 e^{j\theta_0}\). Consistently with the GPSDs, the neighborhood of high-probability of \(K^\text{ECLS}_G\) is more spread along \(\theta\) than the one of \(K^\text{ECLS}_L\), and the latter is more

\[\begin{align*}
\exp\left(-\frac{g^T K^{-1} g}{2}\right)
\end{align*}\]
Fig. 3. (a) PDF for processes having kernel $K^{ECLS}_L$ (first), $K^{ECLS}_C$ (second) and $K^{ECLS}_G$ (third) and filtered kernel (fourth). Here, $\beta = 0.1, \lambda_0 = 0.9, \varphi_0 = 3\pi/5$ and $\rho_0 = 0.93$. (b) PDF for processes having kernel $K^{ECLS}_L$ (first), $K^{ECLS}_C$ (second) and $K^{ECLS}_G$ (third) and filtered kernel (fourth). Here, $\beta = 0.05, \lambda_0 = 0.9, \varphi_0 = 3\pi/5$ and $\rho_0 = 0.96$. (c) PDF for processes having kernel $K^{INT}_L$ (first), $K^{INT}_C$ (second) and $K^{INT}_G$ (third) and integrated filtered kernel (fourth). Here, $\beta = 0.05, \lambda_m = 0.4, \lambda_M = 0.9, \varphi_0 = 3\pi/5$ and $\rho_0 = 0.96$.

spread along $\vartheta$ than the one of $K^{ECLS}_C$. It is worth noting that for the filtered kernel there is a neighborhood of high-probability around a pole with phase close to zero. Indeed, process $\tilde{g}(t)$ at the input of $F(z)$ gives high-probability to impulse responses with pole whose phase is close to zero, and $F(z)$ does not sufficiently attenuate these impulse responses at the output.

Decreasing $\beta$ to 0.05, the statistical power at time $t = 0$ is more concentrated around $z_0$ along the normalized angular frequency domain, see Figure 1 (bottom). The corresponding PDFs are depicted in Figure 3 (b). Consistently with the GPSDs, the neighborhood of high-probability is more squeezed along $\vartheta$. Regarding the filtered kernel, $\rho$ has been increased to 0.96; then the bandwidth of $F(z)$ is decreased. As a consequence, the neighborhood of high-probability around the pole with phase close to zero disappeared because $F(z)$ now drastically attenuate these impulse responses at the output.

5.2 Integrated kernels

We consider the integrated versions of (20)

$$K^{INT}_T(t, s) = \frac{e^{\alpha_M(t+s)} - e^{\alpha_m(t+s)}}{t+s}K_T(t-s)$$

where $T \in \{L, C, G\}$, $\alpha_m = -0.92$, $\alpha_M = -0.1$, $\beta = 0.05$ and $\omega_0 = 3\pi/5$. Similarly as before, we discretize these kernels with $T = 1$. The corresponding GPSDs have as support an annulus inside the unit circle, where the lower radius is $\lambda_m = e^{\alpha_m} = 0.4$ and the upper radius $\lambda_M = e^{\alpha_M} = 0.9$, see Figure 2 (right). If we develop any circle in the support, then we find the function of Figure 1 (bottom) up to a scaling factor. Therefore, the statistical power at time $t = 0$ is concentrated in a neighborhood of $z_0$ which now is spread along the decay rate interval $[\lambda_m, \lambda_M]$. The corresponding PDFs are depicted in Figure 3 (c). Consistently with the GPSDs, the PDFs are such that the neighborhood of high-probability prob-
ability around $p_0$ is now also spread over the decay rate domain. Note that, the integrated version of the filtered kernel is obtained by feeding $\hat{F}(z)$, here $p_0 = 0.96$, with process $\hat{g}(t)$ with kernel the integrated TC kernel (15).

6 GPSD design for a special class of ECLS kernels

In order to study in more detail the relation among the (sample) properties of Bayes estimators, the frequency domain description of the unknown impulse responses (Fourier transform) and the properties of the kernel, we shall now specialize to a particular class of priors on $g(t)$, admitting an ECLS kernel as in (9). Under this restriction, the generalized Fourier transform takes the form

$$c(\alpha, \omega) = c_1(\omega)\delta(\alpha - \alpha_0) \quad (21)$$

where $c_1(\omega)$ is a complex Gaussian process such that

$$E[c_1(\omega)c_1(\omega')] = \phi_1(\omega)\delta(\omega - \omega')$$
$$E[c_1(\omega)c_1(\omega')] = 0, \quad (22)$$

and $\phi_1(\omega) \geq 0$ denotes its PSD. Indeed,

$$E[c(\alpha, \omega)c(\alpha', \omega')] = E[c_1(\omega)c_1(\omega')][\delta(\alpha - \alpha_0)\delta(\alpha' - \alpha_0)] = \phi_1(\omega)\delta(\omega - \omega')\delta(\alpha - \alpha_0)\delta(\alpha - \alpha')$$

and $\phi(\alpha, \omega) = \phi_1(\omega)\delta(\alpha - \alpha_0)$. By (4) and (21), it is not difficult to see that

$$g(t) = \frac{1}{2}e^{\alpha_0 t} \int_{-\infty}^{\infty} G_{\alpha_0}(\omega)e^{j\omega t}d\omega. \quad (23)$$

where $G_{\alpha_0}(\omega) := (c_1(\omega) + c_1(-\omega))/2$. Accordingly, $G_{\alpha_0}(\omega)$ is the Fourier transform of $\pi^{-1}e^{-\alpha_0 t}g(t)$:

$$G_{\alpha_0}(\omega) = \int_{0}^{\infty} \pi^{-1}e^{-\alpha_0 t}g(t)e^{-j\omega t}dt. \quad (24)$$

The next Proposition characterizes the posterior mean of $g(t)$ in terms of the PSD $\phi_1(\omega)$.

**Proposition 4** Consider the continuous time model

$$v(t) = \int_{0}^{\infty} g(s)u(t-s)ds, \ t \in \mathbb{R}^+ \quad (25)$$

$u(t)$ is the (known) measured input. Let $y^N = [y(1) \ldots y(N)]^T$ be sampled noisy measure-ments\(^6\)

$$y(k) := v(k) + e(k) \quad k = 1, \ldots, N$$

where $e(k), \ k = 1, \ldots, N$ are i.i.d. zero mean Gaussian with variance $\sigma^2$, independent of $g(t)$. Assuming the prior distribution on $g(t)$ is Gaussian with kernel (9), the posterior mean $E[g(t)|y^N]$ of $g(t)$ given $y^N = [y(1) \ldots y(N)]^T$ satisfies

$$E[g(t)|y^N] = \frac{1}{2} e^{\alpha_0 t} \int_{-\infty}^{\infty} G_{\alpha_0}(\omega)[y^N]e^{j\omega t}d\omega \quad (26)$$

where

$$E[G_{\alpha_0}(\omega)|y^N] = \phi_1(\omega)U_{\alpha_0}(\omega)^*V_{\alpha_0}^{-1}y^N \quad (27)$$

is the posterior mean of $G_{\alpha_0}(\omega)$ with

$$V_{\alpha_0} = \int_{-\infty}^{\infty} \phi_1(\omega)U_{\alpha_0}(\omega)U_{\alpha_0}(\omega)^*d\omega + \sigma^2 I_N$$

$$U_{\alpha_0}(\omega) = \frac{1}{2} \int_{0}^{\infty} e^{\alpha_0 s}u_s^N e^{j\omega s}ds$$

$$u_s^N = [u(1-s) \ldots u(N-s)]^T. \quad (28)$$

Proposition 4 can be also adapted to the discrete time case as follows:

**Proposition 5** Consider a discrete time process $g(t)$ with ECLS kernel $K(t,s) = \lambda_0^{t+s}K_1(t-s)$, with $0 < \lambda_0 < 1$ and $t, s \in \mathbb{N}$. Then,

$$g(t) = \frac{1}{2} \lambda_0^t \int_{-\pi}^{\pi} G_{\lambda_0}(\vartheta) e^{j\vartheta t}d\vartheta$$

$$G_{\lambda_0}(\vartheta) = \sum_{t=1}^{\infty} \pi^{-1} \lambda_0^{-t} g(t)e^{j\vartheta t}. \quad (29)$$

Consider the discrete time OE model

$$y(t) = \sum_{s=1}^{\infty} g(s)u(t-s) + e(t), \ t \in \mathbb{N}$$

where $u(t)$ is the measured input and $e(t)$ is zero mean white Gaussian noise with variance $\sigma^2$, independent of $g(t)$ and assume $\phi_1(\vartheta)$ is the PSD of $K_1(t-s)$. Then the posterior mean $E[g(t)|y^N]$ of $g(t)$ given $y^N = [y(1) \ldots y(N)]^T$ is given by

$$E[g(t)|y^N] = \frac{1}{2} \lambda_0 \int_{-\pi}^{\pi} E[G_{\lambda_0}(\vartheta)|y^N]e^{j\vartheta t}d\vartheta$$

For simplicity, here we assume that the sampling time is $T = 1$. 

---

6 For simplicity, here we assume that the sampling time is $T = 1$. 

---

8
where

\[ \mathbb{E}[G_{\lambda_0}(\vartheta) | y^N] = \phi_1(\vartheta) U_{\lambda_0}(\vartheta) * V_{\lambda_0}^{-1} y^N \]

is the posterior mean of \( G_{\lambda_0}(\vartheta) \) with

\[
V_{\lambda_0} = \int_{-\pi}^{\pi} \phi_1(\vartheta) U_{\lambda_0}(\vartheta) U_{\lambda_0}(\vartheta)^* d\vartheta + \sigma^2 I_N
\]

\[
U_{\lambda_0}(\vartheta) = \frac{1}{2} \sum_{s=1}^{\infty} e^{\lambda_0 s} u_s^N e^{j\vartheta s}
\]

\[
u_s^N = [u(1-s) \ldots u(N-s)]^T.
\]

For simplicity in what follows we consider the discrete time case, but the same observations hold for the continuous time case. Proposition 5 shows that the absolute value of the posterior mean of \( G_{\lambda_0}(\vartheta) \) is proportional to \( \phi_1(\vartheta) \) (frequency wise). To understand better this fact, we consider a data record \( Z^N \) of length \( N = 500 \) generated by the discrete time model (1) with transfer function having poles in 0.936, -0.45 ± 0.8, -0.25 ± 0.85 and zeros in 0.16, -0.8 ± 0.4. Since the dominant pole is 0.936, we take \( \lambda_0 = 0.94 \). In this way \( \pi^{-1} \lambda_0^{-1} g(t) \) admits

Fig. 4. Fourier transform of \( \pi^{-1} \lambda_0^{-1} g(t) \).

Fourier transform \( G_{\lambda_0}(\vartheta) \), see Figure 4.

We consider the posterior mean estimators of \( g(t) \) with three different sampled version of kernel (9) with \( T = 1 \). In particular we choose \( \alpha_0 = \log(\lambda_0) \), and \( \phi_1(\omega) \) takes one of the following three shapes:

- \( \phi_1(\omega) = \phi_{LO}(\omega) = 2\beta / (\pi(\beta^2 + \omega^2)) \)
- \( \phi_1(\omega) = \phi_{L}(\omega) = \tilde{\phi}_{L}(\omega) + \tilde{\phi}_{L}(-\omega) \) with \( \tilde{\phi}_{L} \) as in (8)
- \( \phi_1(\omega) = \phi_{M}(\omega) = \phi_{LO}(\omega) + \phi_{L}(\omega) \) (a mixture kernel inspired by [Chen et al., 2014]).

We shall denote, correspondingly, with \( g_{LO}(t), g_{L}(t) \) and \( g_{M}(t) \) the three estimators.

The hyperparameters of the kernel are estimated from the data by minimizing the negative log-likelihood (e.g. the ones for \( g_{M}(t) \) are \( \beta, \omega_0 \) and the scaling factor). In Figure 5, the Fourier transform of the estimates, as well as the absolute errors are depicted. It is clear that \( g_{M}(t) \) provides the best approximation of \( g(t) \). In particular, compared to \( g_{LO}(t) \) and \( g_{L}(t) \), it improves the approximation at low (\( \vartheta \approx 0 \)), medium (\( \vartheta \approx 1 \)) and high (\( \vartheta \approx \pi \)) normalized angular frequencies. In Figure 6, the PSDs \( \phi_{LO}(\vartheta), \phi_{L}(\vartheta) \) and \( \phi_{M}(\vartheta) \) of the stationary part of the three discretized kernels are depicted. Note that \( \phi_{LO}(\vartheta), \phi_{L}(\vartheta) \) and \( \phi_{M}(\vartheta) \) are the periodic repetition (up to a scaling factor) of \( \phi_{LO}(\omega), \phi_{L}(\omega) \) and \( \phi_{M}(\omega) \), respectively, with period \( 2\pi \). It can be noticed that, only \( \phi_{M} \) follows the shape of \( |G_{\lambda_0}(\vartheta)| \). This confirms the intuition that, if the PSD of the stationary part of the kernel has a similar shape of \( |G_{\lambda_0}(\vartheta)| \), then we expect the corresponding estimate of \( g(t) \) is good. Hence this provides guidelines as to how the (stationary part of the) kernel should be designed; in particular, it should mimic the frequency response function \( G_{\lambda_0}^{\dagger}(\vartheta) \) of \( \pi^{-1} \lambda_0^{-1} g(t) \) where \( g^{\dagger}(t) \) is the “true” impulse response function.

Note that that optimality of the stationary part of the kernel is coupled with the choice of the decay \( \lambda_0 \); in practice, estimating the hyperparameters using the marginal likelihood estimator allows to optimize jointly \( \lambda_0 \) and the stationary part \( \phi_1(\omega) \).

Note also that, for the discussion above to make sense, \( \lambda_0^{-1} g^{\dagger}(t) \) should admit a Fourier transform, which imposes constraints on \( \lambda_0 \). To the purpose of illustration, let us postulate that \( g^{\dagger}(t) = \sum_{i=1}^{\infty} a_i e^{i\alpha_i t} \cos(\delta t) \) with \( \alpha_i \in \mathbb{R}_- \), i.e. \( g^{\dagger}(t) \) is a sum of damped sinusoids, then (23)-(24) hold if and only if

\[
\alpha_{\max} := \max_{i \geq 2} \alpha_i < \alpha_0 := \log(\lambda_0).
\]

For instance, the TC kernel, i.e. stable spline kernel of order one, is defined as \( K_{TC}(t,s) = e^{-\gamma \max(t, s)} \) with \( \gamma > 0 \) which can be written in the form (9) choosing \( \alpha_0 = -\gamma / 2 \) and \( \phi_L(\omega) \) in (8) with \( \beta = \gamma / 2 \), therefore condition (29) becomes \( \alpha_{\max} < -\gamma / 2 \). Furthermore, the SS kernel, i.e. stable spline kernel of order two, is defined

\[
K_{SS}(t,s) = \frac{e^{-\gamma (t+s)} - e^{-\gamma \max(t, s)}}{2} - \frac{e^{-3\gamma \max(t, s)}}{6};
\]

we can rewrite it as (9) by choosing \( \alpha_0 = -3\gamma / 2 \) and

\[
\phi_1(\omega) = \frac{e^{-\gamma |t-s|}}{2} - \frac{e^{-2\gamma |t-s|}}{6},
\]

therefore condition (29) becomes \( \alpha_{\max} < 3\gamma / 2 \). It is interesting to note that the two conditions above coincide with the conditions derived in [Pillonetto et al., 2010] for
the posterior mean estimator to be statistically consistent, namely:

$$\gamma < -\frac{2\alpha_{\text{max}}}{2m-1},$$

where $m$ is the order of the stable spline kernel. Indeed, if (23)-(24) hold then the hypothesis space is endowed by a probability density which is strictly positive at the “true” system $g^\dagger(t)$; under this condition it can be proved that the posterior mean estimator almost surely converges to $g^\dagger(t)$. A similar reasoning can be applied to the discrete time case.

### 7 Kernel approximation

Kernel approximation is widely used in machine learning and system identification to reduce the computational burden. Next, we show that the GPSD represents a powerful tool for this problem. To this purpose, note that the process $g_a(t)$ in (3) can be understood as an approximation of process $g(t)$ in (4). In particular, its kernel function is

$$K_a(t,s) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{N} \phi_{ik} z_{\alpha_i,\omega_k}(t) \top z_{\alpha_i,\omega_k}(s).$$

We define $z_{\alpha,\omega}(t) = e^{\alpha t}[\cos(\omega t) \; \sin(\omega t)]^\top$, then

$$K_a(t,s) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \phi_{ik} z_{\alpha_i,\omega_k}(t) \top z_{\alpha_i,\omega_k}(s).$$

The computational burden of the identification procedure based on $K_a(t,s)$ is related to the number of points

$$z_{\alpha,\omega}(t) = e^{\alpha t}[\cos(\omega t) \; \sin(\omega t)]^\top,$$
\(N_\alpha\) and \(N_\omega\), which can thus be chosen to trade off kernel approximation and computational cost. For fixed \(N_\alpha\) and \(N_\omega\), the quality of the approximation \(K_\alpha(t,s) \approx K(t,s)\) depends on the choice of the \(\phi_{ik}'s\), which will be now discussed. First of all, let us observe that (6) can be approximated as follows:

\[
K(0,0) \approx \frac{1}{2} \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\omega} \Delta \alpha_i \Delta \omega_k \phi(\alpha_i, \omega_k) \tag{31}
\]

where \(\Delta \alpha_i = \alpha_{i+1} - \alpha_i\), \(\Delta \omega_k = \omega_{k+1} - \omega_k\) with \(\alpha_{i+1} > \alpha_i\), \(\omega_{k+1} > \omega_k\). On the other hand, we have

\[
K_\alpha(0,0) = \frac{1}{2} \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\omega} \phi_{ik}. \tag{32}
\]

Matching (31) and (32), we obtain

\[
\phi_{ik} = \Delta \alpha_i \Delta \omega_k \phi(\alpha_i, \omega_k). \tag{33}
\]

An alternative way to approximate \(K(t,s)\) takes inspiration from the random features approach for stationary kernels [Rahimi & Recht, 2007]. Observe that (5) can be rewritten as

\[
K(t,s) = \frac{K(0,0)}{2} \mathbb{E}[z_{\alpha,\omega}(t)^\top z_{\alpha,\omega}(s)] \tag{33}
\]

where \(\mathbb{E}[\cdot]\) is the expectation operator taken with respect to the PDF \(\tilde{p}(\alpha, \omega) = K(0,0)^{-1} \phi(\alpha, \omega)\). Then, we can approximate (33) with

\[
K_\alpha(t,s) = \frac{\gamma}{2} \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\omega} \phi_{ik} z_{\alpha,\omega_k}(t)^\top z_{\alpha,\omega_k}(s)
\]

where \(\phi_{ik}'s\) are drawn from \(\tilde{p}(\alpha, \omega)\) and the constant \(\gamma\) satisfies \(\gamma = \frac{1}{2} \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\omega} \phi_{ik}\).

7.1 Numerical experiments

Data set. We generate 1000 discrete time SISO systems of order 30. The poles are randomly generated as follows: 75% have phase randomly generated over an interval of size \(\pi/6\) centered in \(\vartheta_0 \sim \mathcal{U}[\pi/4,3\pi/4]\) and absolute value \(\sim \mathcal{U}[0.8,0.95]\); the remaining poles are generated uniformly inside the closed unit disc of radius \([0,0.95]\). For each system a data set of 230 points is obtained feeding the linear system with zero mean, unit variance, white Gaussian noise and corrupting the output with additive zero mean white Gaussian noise so as to guarantee a signal to noise ratio equal to 10.

Simulation setup and results. We consider model (1) with \(G(z) = \sum_{t=1}^{n} g(t) z^{-t}\) where \(n = 100\) is the practical length. We consider several estimators of \(g(t)\) which differ on the choice of the kernel describing the prior distribution on \(g(t)\). In particular we shall use the following subscripts:

- \(L0\) for estimator which uses the kernel \(K_{L}^{INT}\) with \(\omega_0 = 0\);
- \(L0A\) as above but with \(K_{L}^{INT}\) approximated using (30) with \(N_\omega = 5\) and \(N_\alpha = 3\);
- \(L\) for estimator which uses the kernel \(K_{L}^{INT}\);
- \(LA\) as above but with \(K_{L}^{INT}\) approximated using (30) with \(N_\omega = 5\) and \(N_\alpha = 3\);

The subscripts \(G0, G0A, G, GA\) and \(C0, C0A, C, CA\) will have the same meaning as above but with respect to kernel \(K_{L}^{INT}\) and \(K_{C}^{INT}\) respectively. Note that all hyperparameters (e.g. \(\alpha_m, \alpha_M, \beta, \omega_0\) and the scaling factor for \(K_{L}^{INT}\)) are estimated from the data by minimizing the negative log-likelihood. To evaluate the various kernels, the impulse response estimates \(\hat{g}(t), t = 1 \ldots n\), are compared to the true one, i.e. \(g(t)\), by the average fit

\[
AF = 100 \left(1 - \frac{\sqrt{\sum_{t=1}^{n} |\hat{g}(t) - g(t)|^2}}{\sum_{t=1}^{n} |\hat{g}(t) - \bar{g}|^2}\right), \quad \bar{g} = \frac{1}{n} \sum_{t=1}^{n} g(t).
\]

The distribution of the fits are shown by box-plots in Figure 7: L0 and C0 estimators are outperformed by their approximated versions. In the remaining cases the approximated versions provides similar performance to the exact version.
8 Conclusions

In this paper we have introduced the harmonic representation of kernel functions used in system identification. In doing that, we have introduced the GPSD which represents the generalization of the power spectral density used for the harmonic representation of stationary kernels. We have showed by simulation that the GPSD carries similar information that we can find in the PDF of the process over the class of second-order systems. Moreover, we have characterized the posterior mean in terms of the GPSD for a special class of ECLS kernels. Finally, we have showed that the GPSD provides a powerful tool to approximate kernels function, and thus to reduce the computational burden in the system identification procedure.

Appendix

Proof of Proposition 1

Rewriting

\[ g(t) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{t} e^{\alpha(t+s)} \{ c(\alpha, \omega) e^{j\omega t} + c(\alpha, \omega) e^{-j\omega t}\} d\omega d\alpha \]

then we have

\[ K(t, s) = \mathbb{E}[g(t)g(s)] \]

\[ = \frac{1}{4} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{0} \int_{-\infty}^{0} e^{\alpha(t+\alpha')} \{ e^{j\omega t+j\omega'} + e^{-j\omega t-j\omega'} \} d\omega d\alpha \]

Then, by (5) we have

\[ h_s(t) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{\infty} \phi_c(\alpha, \omega)e^{\alpha s}e^{j\omega t}d\omega d\alpha \]

\[ = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{0} \pi\phi_c(\alpha, \omega)e^{\alpha s}\omega e^{j\omega t}d\omega d\alpha \]

\[ = H_s(\omega) \]

accordingly \( H_s(\omega) \) is the Fourier transform of \( h_s(t) \).

Proof of Proposition 4

By (23), we trivially have (26). Then,

\[ \mathbb{E}[G_{\alpha\omega}(\omega)|y^N] = \mathbb{E} \left[ \frac{c_1(\omega) + c_1(-\omega)}{2} \right] \frac{y^N}{2} \]

\[ = \mathbb{E}[c_1(\omega)|y^N] + \mathbb{E}[c_1(-\omega)|y^N] \]

(37)
where
\[
\begin{align*}
E[c_1(\omega)|y^N] &= E[c_1(\omega)(y^N)^T]E[y^N(y^N)^T]^{-1}y^N \\
E[c_1(-\omega)|y^N] &= E[c_1(-\omega)(y^N)^T]E[y^N(y^N)^T]^{-1}y^N.
\end{align*}
\]

By (25), we have
\[
v(t) = \int_{-\infty}^{\infty} c_1(\omega) + c_1(-\omega) U_{ao,t}(\omega)d\omega,
\]
where we exploited relation (23) and
\[
U_{ao,t}(\omega) := \frac{1}{2} \int_{0}^{\infty} e^{ao} u(t-s) e^{i\omega s} ds.
\]

Let \(k, n = 1 \ldots N\). It follows that
\[
E[c_1(\omega)y(k)] = \int_{-\infty}^{\infty} \left[ E[c_1(\omega)c_1(\omega')] + E[c_1(\omega)c_1(-\omega')] \right] \times U_{ao,k}(\omega')d\omega' + E[c_1(\omega)e(k)]
\]
\[
= \frac{1}{2} \phi_1(\omega)U_{ao,k}(\omega)
\]
where we exploited (22) and the fact that \(U_{ao,t}(\omega)\) is an Hermitian function. Moreover,
\[
E[y(k)y(n)] = \int_{-\infty}^{\infty} g(t')u(k-t')g(s')u(n-s')dt'ds' + \sigma^2 \delta_{k-n}.
\]

Substituting (23) in (38) and using (22), we obtain
\[
E[y(k)y(n)] = \frac{1}{2} \int_{-\infty}^{\infty} \phi_1(\omega')U_{ao,k}(\omega')U_{ao,n}(\omega')d\omega' + \sigma^2 \delta_{k-n}
\]
where we exploited the fact that \(\phi_1(\omega) = \phi_1(-\omega)\).

Since \(U_{ao}(\omega) = [U_{ao,1}(\omega) \ldots U_{ao,N}(\omega)]^T\), we have that \(E[c_1(\omega)^2] = \phi_1(\omega)U_{ao}(\omega)^* / 2\) and
\[
E[c_1(\omega)|y^N] = \phi_1(\omega)U_{ao}(\omega)^* V_{ao}^{-1} y^N
\]
where \(V_{ao}\) has been defined in (28). In similar way, it can be proven that
\[
E[c_1(-\omega)|y^N] = \phi_1(\omega)U_{ao}(\omega)^* V_{ao}^{-1} y^N.
\]

Finally, substituting (39) and (40) in (37) we obtain (27).

**Proof of Proposition 5**

The proof is similar to the one of Proposition 4. \(\Box\)

**References**

Carli, F., Chiuso, A., & Pillonetto, G. (2012). Efficient algorithms for large scale linear system identification using stable spline estimators. In *Proc. of SYSID 2012*.

Chen, T., Andersen, M. S., Ljung, L., Chiuso, A., & Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control, 59*, 2933–2945.

Chen, T., & Ljung, L. (2013). Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica, 49*, 2213–2220.

Chen, T., & Ljung, L. (2015a). On kernel structures for regularized system identification (i): a machine learning perspective. *IFAC, 48*, 1035–1040.

Chen, T., & Ljung, L. (2015b). On kernel structures for regularized system identification (ii): a system theory perspective. *IFAC, 48*, 1041–1046.

Chen, T., & Ljung, L. (2016). On kernel design for regularized lti system identification. *arXiv preprint arXiv:1612.03542*.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and gaussian processes revisited. *Automatica, 48*, 1525–1535.

Chiuso, A. (2016). Regularization and Bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control, 41*, 24–38.

Chiuso, A., & Pillonetto, G. (2012). A Bayesian approach to sparse dynamic network identification. *Automatica, 48*, 1553–1565.

Fraccaroli, F., Peruffo, A., & Zorzi, M. (2015). A new recursive least squares method with multiple forgetting schemes. In *54th IEEE Conference on Decision and Control (CDC)* (pp. 3367–3372).

Lindquist, A., & Picci, G. (2015). *Linear Stochastic Systems*. Springer.

Ljung, L. (1999). *System Identification: Theory for the User*. New Jersey: Prentice Hall.

Pillonetto, G., Chen, T., Chiuso, A., Nicolao, G. D., & Ljung, L. (2016). Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica, 69*, 137–149.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2010). Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *American Control Conference, 2010*.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011b). Prediction error identification of linear systems: A nonparametric gaussian regression approach. *Automatica, 47*, 291–305.

Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica, 46*, 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identifi-
cation, machine learning and function estimation: A survey. *Automatica*, 50, 657–682.
Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems* (pp. 1177–1184).
Rasmussen, C., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.
Romeres, D., Zorzi, M., Camoriano, R., & Chiuso, A. (2016). Online semi-parametric learning for inverse dynamics modeling. In *Proceedings of the IEEE Conference on Decision and Control*. Las Vegas.
Silverman, R. (1957). Locally stationary random processes. *IRE Transactions on Information Theory*, 3, 182–187.
Söderström, T., & Stoica, P. (1989). *System Identification*. London, UK: Prentice-Hall International.
Zorzi, M. (2015a). An interpretation of the dual problem of the three-like approaches. *Automatica*, 62, 87 – 92.
Zorzi, M. (2015b). Multivariate spectral estimation based on the concept of optimal prediction. *IEEE Transactions on Automatic Control*, 60, 1647–1652.
Zorzi, M., & Chiuso, A. (2015). A Bayesian approach to sparse plus low rank network identification. In *Proceedings of the IEEE Conference on Decision and Control*. Osaka.
Zorzi, M., & Chiuso, A. (2017). Sparse plus low rank network identification: a nonparametric approach. *Automatica*, 53.