Automatica 94 (2018) 125–137

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

The harmonic analysis of kernel functions☆

Mattia Zorzi *, Alessandro Chiuso

Dipartimento di Ingegneria dell'Informazione, Università degli studi di Padova, via Gradenigo 6/B , 35131 Padova, Italy

ARTICLE INFO

Article history:
Received 15 March 2017
Received in revised form 18 November 2017
Accepted 18 March 2018
Available online 29 May 2018

Keywords:
System identification
Kernel-based methods
Power spectral density
Random features

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.

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 Chen, Ohlsson, and Ljung (2012), Chiuso (2016), Chiuso and Pillonetto (2012), Fraccaroli, Peruffo, and Zorzi (2015), Lataire and Chen (2016), Pillonetto, Chiuso, and De Nicolao (2011), Pillonetto and De Nicolao (2010), Pillonetto, Dinuzzo, Chen, De Nicolao, and Ljung (2014) and Zorzi and Chiuso (2015, 2017).

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 blocks of higher order systems and, as such, understanding second order systems helps understanding higher ones. Furthermore, for a special class of exponentially convex locally stationary processes (ECLS) typically used in system identification (Chen & Ljung, 2015a, b), it is possible to provide (i) a link between the GPSD and the Fourier transform of the exponentially modulated sample trajectory and (ii) characterize the posterior mean in terms of the GPSD. As a consequence, it is possible to outline a simple procedure for the design of the kernel, through the GPSD.

☆ This research has been partially supported by the MIUR FIRB project RBFR12M3AC-Learning meets time: a new computational approach to learning in dynamic systems. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Thomas Bo Schön under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: zorzimat@dei.unipd.it (M. Zorzi), chiuso@dei.unipd.it (A. Chiuso).

https://doi.org/10.1016/j.automatica.2018.04.015
0005-1098/

Another important aspect in kernel-based methods is to reduce the computational burden (Chen & Ljung, 2013). This task can be accomplished by approximating the kernel functions through eigen-decomposition (Carli, Chiuso, & Pillonetto, 2012) or random features, (Rahimi & Recht, 2007) techniques. However, these methods can be applied only to special kernel functions. We show that the GPSD provides a general procedure to approximate a wide class of kernel functions.

The outline of the paper is as follows. In Section 2 we review the Gaussian process regression framework used for kernel-based methods. In Section 3 we present the harmonic representation of the kernel function of continuous time non-stationary processes modeling impulse responses of marginally stable systems and in Section 4 the corresponding discrete time version. In Section 5 we show the relation between the GPSD and the probability density function of the prior over the class of second-order linear systems. In Section 6 we characterize the posterior mean in terms of the GPSD for a special class of ECLS processes. Section 7 regards kernel approximation using the GPSD. In Section 8 we present some numerical examples showing how to exploit the GPSD to design a kernel function. Finally, conclusions are drawn in Section 9. In order to streamline the presentation, all proofs are deferred to the Appendix.

Notation: $\mathbb{N}$ denotes the set of the natural numbers, $\mathbb{Z}$ is the set of the integer numbers, $\mathbb{R}_+$ the set of the nonnegative real numbers, $\mathbb{R}_-$ the set of the negative real numbers, and $\mathbb{C}$ is the set of the complex numbers. Given $x \in \mathbb{C}$, $|x|$ denotes its absolute value, $\angle x$ denotes its phase and $\bar{x}$ denotes its conjugate. Given $A \in \mathbb{C}^{m\times n}$, A^* denotes its transposed conjugate. $\mathbb{E}[\cdot]$ denotes the expectation operator.

2. System identification and Gaussian process regression

For convenience in what follows we consider a discrete time, stable and linear time-invariant (LTI) single input–single output (SISO) in OE form:

$$y(t) = G(z)u(t) + e(t), \quad t \in \mathbb{N} \tag{1}$$

where z^{-1} denotes the backward shift operator; $u(t)$ is the input; $y(t)$ is the noisy output; $e(t)$ is zero mean white noise, that is $\mathbb{E}[e(t)e(s)] = \sigma^2\delta_{t-s}$ where δ_k denotes the Kronecker delta function. The transfer function $G(z)$ is stable and strictly causal, i.e. $G(\infty) = 0$. Expanding $G(z)$ in z^{-1} we obtain the impulse response of the linear system

$$G(z) = \sum_{t=1}^{\infty} g(t)z^{-t}.$$

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

$$\mathcal{Z}^N = \{u(1), y(1) \dots u(N), y(N)\}.$$

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 $\mathcal{Z}^N$ (Pillonetto & 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 (i.e. modes):

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

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 lth 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

$$\mathbb{E}[c_l \overline{c_{l'}}] = \phi_l \delta_{l-l'}$$
$$\mathbb{E}[c_l c_{l'}] = 0$$

with $l, l' = 1 \dots N$ and $\phi_l \geq 0$. Accordingly, the real part and the imaginary part of c_l are independent and with the same variance ϕ_l, see Papoulis and Pillai (2002): roughly speaking, this means that we can rewrite (2) as a weighted sum of sines and cosines with zero mean, real valued and independent random coefficients, with the same variance. For convenience, we rewrite (2) as

$$g_a(t) = \sum_{i=1}^{N_\alpha} \sum_{k=1}^{N_\omega} |c_{ik}| e^{\alpha_i t} \cos(\omega_k t + \angle c_{ik}) \tag{3}$$

that is (α_i, ω_k) belongs to a $N_\alpha \times N_\omega$ grid contained in $\mathbb{R}_- \cup \{0\} \times \mathbb{R}$ and c_{ik} is a complex Gaussian random variable such that

$$\mathbb{E}[c_{ik}\overline{c_{i'k'}}] = \phi_{ik}\delta_{i-i'}\delta_{k-k'}$$
$$\mathbb{E}[c_{ik}c_{i'k'}] = 0$$

with $\phi_{ik} \geq 0$. It is then natural to generalize (3) as an infinite "dense" sum of decaying sinusoids[1]:

$$g(t) = \int_{-\infty}^{0} \int_{-\infty}^{\infty} |c(\alpha,\omega)| e^{\alpha t} \cos(\omega t + \angle c(\alpha,\omega)) \mathrm{d}\omega \mathrm{d}\alpha \tag{4}$$

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

$$\mathbb{E}[c(\alpha,\omega)\overline{c(\alpha',\omega')}] = \phi(\alpha,\omega)\delta(\alpha-\alpha')\delta(\omega-\omega')$$
$$\mathbb{E}[c(\alpha,\omega)c(\alpha',\omega')] = 0$$

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,*

$$K(t,s) = \frac{1}{2} \int_{-\infty}^{0} \int_{-\infty}^{\infty} \phi(\alpha,\omega) e^{\alpha(t+s)} \cos(\omega(t-s)) \mathrm{d}\omega \mathrm{d}\alpha. \tag{5}$$

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

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

[1] Strictly speaking the integral over α should be understood in an open interval of the form $(-\infty, \epsilon)$ with $\epsilon > 0$ and $\epsilon \to 0$.

[2] To be precise, we should work with the "generalized stochastic measure" $C(\omega,\alpha)$, a Gaussian process with orthogonal increments; formally $dC(\omega,\alpha) = c(\omega,\alpha)d\omega d\alpha$.