Automatica 66 (2016) 34–38

Contents lists available at ScienceDirect

Automatica

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

Technical communique

Maximum entropy properties of discrete-time first-order stable spline kernel[☆]

Tianshi Chen[a,b,1], Tohid Ardeshiri[b], Francesca P. Carli[c], Alessandro Chiuso[d], Lennart Ljung[b], Gianluigi Pillonetto[d]

[a] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[b] Department of Electrical Engineering, Linköping University, Sweden
[c] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[d] Department of Information Engineering, University of Padova, Italy

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
11 July 2015
Accepted 22 November 2015
Available online 15 January 2016

Keywords:
System identification
Regularization method
Kernel structure
Maximum entropy

ABSTRACT

The first order stable spline (SS-1) kernel (also known as the tuned-correlated (TC) kernel) is used extensively in regularized system identification, where the impulse response is modeled as a zero-mean Gaussian process whose covariance function is given by well designed and tuned kernels. In this paper, we discuss the maximum entropy properties of this kernel. In particular, we formulate the exact maximum entropy problem solved by the SS-1 kernel without Gaussian and uniform sampling assumptions. Under general sampling assumption, we also derive the special structure of the SS-1 kernel (e.g. its tridiagonal inverse and factorization have closed form expression), also giving to it a maximum entropy covariance completion interpretation.

1. Introduction

A core issue of system identification is the design of model estimators able to suitably balance structure complexity and adherence to experimental data. This is also known as the bias–variance problem in statistical literature. Traditionally, this problem is tackled by applying the maximum likelihood/prediction error method (ML/PEM), see e.g., Ljung (1999), together with model order selection criteria, such as AIC, BIC and cross validation. Recently, a different method has been introduced in Pillonetto and De Nicolao (2010) and further developed in Pillonetto, Chiuso, and De Nicolao (2011), Chen, Ohlsson, and Ljung (2012), Chen, Andersen, Ljung, Chiuso, and Pillonetto (2014); see also the recent survey (Pillonetto, Dinuzzo, Chen, De Nicolao, & Ljung, 2014). Its key idea is to face the bias–variance problem via well-designed and tuned regularization. More specifically, the impulse response $h(t)$ is modeled as a zero-mean Gaussian process $h(t) \sim \mathrm{GP}(0, k(t,s;\alpha))$, where $k(t,s;\alpha)$ is the covariance (kernel) function, and α is the hyper-parameter vector, see e.g., Rasmussen and Williams (2006). The key step is to design a suitable kernel structure which reflects our prior knowledge on the system to be identified, e.g., stability. Once $k(t,s;\alpha)$ is determined, α is tuned by maximizing the marginal likelihood, and then the conditional mean of $h(t)$ is returned as the impulse response estimate.

Several kernel structures have been proposed, e.g., the stable spline (SS) kernel in Pillonetto and De Nicolao (2010) and the diagonal and correlated (DC) kernel in Chen et al. (2012), which have shown satisfying performance via extensive simulated case studies. In view of this, it seems interesting to investigate how, beyond the empirical evidence, the use of these regularized approaches can be justified by theoretical arguments. Different perspectives can be taken, e.g. deterministic arguments in favor of SS and DC kernels are developed in Chen et al. (2012) while Chiuso, Chen, Ljung, and Pillonetto (2014) discusses the link between the first order stable spline (SS-1) kernel and the Brownian Bridge process suggesting that the SS-1 kernel is indeed a natural description for exponentially decaying impulse responses. In this paper, we will instead work within the Bayesian context, discussing the maximum entropy (MaxEnt) properties of the SS-1 kernel.

The MaxEnt approach has been proposed by Jaynes to derive complete statistical prior distributions from incomplete a priori

---

☆ The material in this paper was partially presented at the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, April 19–24, 2015, Brisbane, Australia. This paper was recommended for publication in revised form by Associate Editor A. Pedro Aguiar under the direction of Editor André L. Tits.

E-mail addresses: tschen@cuhk.edu.cn, tschen@isy.liu.se (T. Chen), tohid@isy.liu.se (T. Ardeshiri), fpc23@cam.ac.uk (F.P. Carli), chiuso@dei.unipd.it (A. Chiuso), ljung@isy.liu.se (L. Ljung), giapi@dei.unipd.it (G. Pillonetto).

[1] Tel.: +46 013284726.

http://dx.doi.org/10.1016/j.automatica.2015.12.009
0005-1098/

information (Jaynes, 1982). Among all distributions that satisfy some constraints, e.g. in terms of the value taken by a few expectations, the MaxEnt criterion chooses the distribution maximizing the entropy. The justification underlying this choice is that the MaxEnt distribution, subject to available knowledge, is the one that can be realized in the greatest number of ways, see also Jaynes' Concentration Theorem (Jaynes, 1982). A preliminary study on the MaxEnt property of kernels for system identification was developed in Pillonetto and De Nicolao (2011). Working in continuous time (CT) and under Gaussian assumption, the problem was to derive the MaxEnt prior using only information on the smoothness and exponential stability of the impulse response. The arguments in Pillonetto and De Nicolao (2011) were however quite involved, mainly due to the infinite-dimensional nature of the problem and the fact that the differential entropy rate of a generic CT stochastic process is not well-defined. Another recent contribution is Carli (2014) where, under Gaussian and uniform sampling assumption, it is shown that the SS-1 kernel matrix can be given a MaxEnt covariance completion interpretation (Dempster, 1972), that is then exploited to derive its special structure (namely that it admits a tridiagonal inverse with closed form representation as well as factorization).

In this paper, we study the MaxEnt properties of the *discrete-time* (DT) SS-1 kernel. We first formulate the MaxEnt problem solved by the DT SS-1 kernel without Gaussian and uniform sampling assumptions. Then, we extend the result of Carli (2014) and link it to our former result: under general sampling assumption, we show that the inverse of the SS-1 kernel matrix is tridiagonal and moreover, the inverse and its factorization have closed form expression, which can be used to derive numerically more stable and efficient algorithms for regularized system identification (Chen & Ljung, 2013, Carli, Chen, & Ljung, 2014, Section 5). What is more, we show that the SS-1 kernel matrix is the solution of a maximum entropy covariance extension problem (Dempster, 1972) with band constraints (the same holds for the kernel matrix associated with the DT Wiener process).

2. MaxEnt property of Wiener and SS-1 kernels

We are dealing with real-valued DT stochastic processes defined on an ordered index set $\mathcal{T} = \{t_i | 0 \leq t_i < t_{i+1} \leq \infty, i = 0, 1, 2, \ldots, \}$. A real-valued DT stochastic process $w(i)$ with $i = 0, 1, 2, \ldots,$ is called a white Gaussian noise if the $w(i)$'s are independent identically Gaussian distributed with zero mean and constant variance. For a real valued random variable X, we let $p(x)$, $\mathbb{E}(X)$ and $\mathbb{V}(X)$ to denote its probability density function, mean and variance of X, respectively. Moreover, the differential entropy $H(X)$ of X is defined as $H(X) = -\int_S p(x)\log p(x)\mathrm{d}x$, where S is the support set of X.

2.1. DT Wiener process

Let $w(i)$ with $i = 0, 1, 2, \ldots,$ be a white Gaussian noise with constant variance c and construct $g(t)$ as follows:

$$g(t_0) = 0 \quad \text{with } t_0 = 0,$$
$$g(t_k) = \sum_{i=1}^{k} w(i-1)\sqrt{t_i - t_{i-1}}, \quad k = 1, 2, \ldots \tag{1}$$

Then it is easy to see that $g(t)$ is a Gaussian process with zero mean and covariance (kernel) function:

$$\text{Wiener: } K^{\text{Wiener}}(t, s; c) = c\min(t, s), \quad t, s \in \mathcal{T} \tag{2}$$

and moreover, $g(t)$ is the DT Wiener process since $g(t_0) = 0$, $g(t)$ is Gaussian distributed with zero mean, and has independent increments with $g(t_i) - g(t_j) \sim \mathcal{N}\left(0, c(t_i - t_j)\right)$ for $0 \leq t_j < t_i$. We will now show that $g(t)$ has MaxEnt property and this result will then be used to derive the MaxEnt property of the SS-1 kernel.

Lemma 1. [2] *Let $h(t)$ be any stochastic process with $h(t_0) = 0$ for $t_0 = 0$. For any $n \in \mathbb{N}$, the DT Wiener process (1) is the solution to the MaxEnt problem*

$$\begin{aligned} \underset{h(\cdot)}{\text{maximize}} \quad & H(h(t_1), h(t_2), \ldots, h(t_n)) \\ \textit{subject to} \quad & \mathbb{V}(h(t_i) - h(t_{i-1})) = c\,(t_i - t_{i-1}) \\ & \mathbb{E}(h(t_i)) = 0, \quad i = 1, \ldots, n \end{aligned} \tag{3}$$

where for simplicity $H(h(t_1), h(t_2), \ldots, h(t_n))$ denotes the differential entropy of $[h(t_1)\; h(t_2) \; \cdots \; h(t_n)]^T$.

2.2. The first order SS kernel

Based on Lemma 1, we can derive the MaxEnt property of the SS-1 kernel:

$$\begin{aligned} &\text{SS-1: } K^{\text{SS-1}}(t, s; \alpha) = c\min(e^{-\beta t}, e^{-\beta s}), \\ &\alpha = [c\;\beta]^T, \quad c \geq 0, \beta > 0, t, s \in \mathcal{T} \end{aligned} \tag{4}$$

which was introduced independently in a deterministic argument in Chen et al. (2012) and called the tuned correlated (TC) kernel. It is fair to call (4) the SS-1 kernel here, since the "stable" time transformation involved in deriving the SS-1 kernel plays a key role in the following theorem.

Theorem 1. *Define a stochastic process as follows:*

$$\begin{aligned} &h^o(t_k) = \sum_{i=k}^{n-1} w(n-1-i)\sqrt{e^{-\beta t_i} - e^{-\beta t_{i+1}}}, \\ &k = 0, \ldots, n-1, \; h^o(t_n) = 0 \text{ with } t_n = \infty \end{aligned} \tag{5}$$

where $w(i)$ with $i = 0, 1, 2, \ldots,$ is a white Gaussian noise with constant variance c. Then $h^o(t)$ is a Gaussian process with zero mean and the SS-1 kernel (4) as its covariance function. Further let $h(t)$ be any stochastic process with $h(t_n) = 0$ for $t_n = \infty$. For any $n \in \mathbb{N}$, the Gaussian process (5) is the solution to the MaxEnt problem

$$\begin{aligned} \underset{h(\cdot)}{\text{maximize}} \quad & H(h(t_0), h(t_1), \ldots, h(t_{n-1})) \\ \textit{subject to} \quad & \mathbb{V}(h(t_{i+1}) - h(t_i)) = c\left(e^{-\beta t_i} - e^{-\beta t_{i+1}}\right) \\ & \mathbb{E}(h(t_i)) = 0, i = 0, \ldots, n-1 \end{aligned} \tag{6}$$

Remark 1. In the cost function of (3) and (6), if we divide the differential entropy of a *finite* sequence of the stochastic process by n and let n go to ∞, then the limit (if exists) becomes the differential entropy rate of the stochastic process (Cover & Thomas, 2012). However, this limit does not exist for Gaussian processes (1) and (5), see Ardeshiri and Chen (2015, Remark 1) for details, and thus the differential entropy of a *finite* sequence of stochastic processes is used instead.

3. Special structure of Wiener and SS-1 kernels and their MaxEnt interpretation

It is easy to see that both stochastic processes in (1) and (5) are not only Gaussian processes but also Markov processes with order-1 Markov property, see e.g., Rasmussen and Williams (2006, Appendix B), Chen and Ljung (2015, Section 5). This observation implies the kernel matrix of the Wiener and SS-1 kernels has

[2] All proofs have been deferred to the Appendix.