Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Constructive state space model induced kernels for regularized system identification ⋆

Tianshi Chen and **Lennart Ljung**

Division of Automatic Control, Department of Electrical Engineering,
Linköping University, Sweden. Email:tschen@isy.liu.se,ljung@isy.liu.se.

Abstract: There are two key issues for the kernel-based regularization method: the kernel structure design and the hyper-parameter estimation. In this contribution, we introduce a new family of kernel structures based on state space models. It has more flexible and more general structure, and includes some of stable spline kernels and diagonal correlated kernels as special cases. We also tested a different method for the hyper-parameter estimation by maximizing a profile marginal likelihood and examined three methods dealing with the initialization. Monte Carlo simulations show that the tested kernels are on the average a bit better than the tuned correlated kernel and the profile marginal likelihood maximization and the pre-windowing method work well for hyper-parameter estimation and initialization.

Keywords: System identification, regularization method, kernel structure design, state space model

1. INTRODUCTION

Linear time invariant (LTI) system identification is a classic and fundamental topic in the area of system identification, Ljung (1999); Söderström and Stoica (1989); Pintelon and Schoukens (2012). A new method to this topic is the kernel-based regularization method (KRM) introduced in Pillonetto and De Nicolao (2010) and further studied in Pillonetto et al. (2011); Chen et al. (2012, 2014); Pillonetto et al. (2014). In contrast with the classical prediction error method (PEM) Ljung (1999), equipped with the classical model structure selection technique, such as AIC, BIC, etc., KRM embraces a more reliable way to deal with the bias-variance tradeoff and can often yield more accurate and robust model estimates for short and noisy data.

There are two key issues for KRM: the kernel structure design, i.e., parameterization of the kernel by some parameters often called hyper-parameters, and the hyper-parameter estimation. So far two families of fundamental kernel structures have been introduced: the stable spline (SS) kernel, Pillonetto and De Nicolao (2010) and the diagonal correlated (DC) kernel, Chen et al. (2012). The problem with SS and DC kernels is that these kernels could be improved for systems with rapid oscillation and/or complicated dynamics Pillonetto et al. (2011); Chen et al. (2014). It is thus interesting and important to design kernels with more flexible and more general structure that are suitable for LTI stable system identification. The hyper-parameter estimation issue is often handled by first embedding the regularization problem in a Bayesian framework and then invoking the marginal likelihood maximization method. A tricky problem for this method is how to deal with the unknown noise variance of the measurement noise. In Pillonetto and De Nicolao (2010); Pillonetto et al. (2011); Chen et al. (2012), a low-bias ARX model or a FIR model is first estimated and then its sample variance is used as an estimate and finally the hyper-parameter estimate is yielded by maximizing the marginal likelihood with the noise variance replaced by its estimate.

In this contribution, we introduce a new family of fundamental kernel structure based on stochastic state space models, which is called state space model induced (SSMi) kernel. Under certain conditions on the state space model, the proposed kernel structure ensures that its corresponding reproducing kernel Hilbert space is a subspace of the space of absolutely integrable functions over $[0,\infty)$ and thus is suitable for LTI stable system identification. This kind of SSMi kernels are thus named SSMi stable (SSMiS) kernel. Note that the concept of stable kernel is introduced in Pillonetto and De Nicolao (2010) and further elaborated in Dinuzzo (2012). The SSMiS kernel includes some of SS and DC kernels as special cases. For illustration, a couple preliminary instances of SSMiS kernel are examined here. Besides, we tested a different method for the hyper-parameter estimation. We first maximize the marginal likelihood with respect to the noise variance so that we can express the optimal noise variance as a function of the hyper-parameters and the given data. We then get the estimate of the hyper-parameters by maximizing the profile marginal likelihood, which is obtained by replacing the unknown noise variance with its optimal value in the marginal likelihood. We also examined three methods dealing with the initialization: non-windowing, pre-windowing and estimating the transient as an additional regularized FIR model. Monte Carlo simulations show that the tested SSMiS kernels are on the average a bit better than the TC (tuned correlated) kernel Chen et al. (2012) and moreover, the profile marginal likelihood maximization and the pre-windowing method work well for hyper-parameter estimation and initialization.

2. REGULARIZED SYSTEM IDENTIFICATION AND EXISTING KERNELS

Consider a discrete time LTI stable causal system

$$y(t) = G_0(q)u(t) + v(t), \qquad t = 0, \cdots, N-1. \tag{1}$$

where t is the time index, q is the shift operator and $qu(t) = u(t+1)$, and $y(t), u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}$ are the measurement output, the input and the noise at time t, respectively. For simplicity, $v(t)$ is assumed to be white [1] . The transfer function

⋆ The work has been supported by the ERC advanced grant LEARN, no 267381, funded by the European Research Council, the Linnaeus Center CADICS, funded by the Swedish Research Council, VR.

[1] The case where $v(t)$ is colored can be handled in a straightforward way for this kernel-based regularization method, see Pillonetto et al. (2014).

$G_0(q)$ is characterized by $G_0(q)=\sum_{k=1}^{\infty} g_k^0 q^{-k}$, where the coefficients $g_k^0, k=1,\cdots,\infty$ form the impulse response of $G_0(q)$. Our goal is to get an estimate $\hat{G}(q)$ of $G_0(q)$ with the collected data $y(t),u(t),t=0,\cdots,N-1$.

2.1 Regularized impulse response estimation

Since the impulse response of LTI stable systems decays exponentially, it is often enough to truncate $G_0(q)=\sum_{k=1}^{\infty} g_k^0 q^{-k}$ at a certain order n and consider a finite impulse response (FIR) model

$$G(q,\theta)=\sum_{k=1}^{n} g_k q^{-k}, \qquad \theta=[g_1\ g_2\ \cdots\ g_n]^T. \tag{2}$$

Estimating the FIR model (2) by using least squares method is well-known but not often used in practice due to the possibly large variance for large n. As shown in Chen et al. (2012), a suitably designed regularization can be added to curb the large variance and get a FIR model estimate with much smaller mean square error (MSE), leading to the kernel-based regularization method. More specifically, using the FIR model (2), the model of system (1) is described by

$$y(t)=\sum_{k=1}^{n} g(k)u(t-k)+v(t), \quad t=0,\cdots,N-1, \tag{3}$$

which can be rewritten as: $Y_N=\Phi_N^T\theta+V_N$, where the ith row of $Y_N,V_N\in\mathbb{R}^M$ with $M=N-n$ and $\Phi_N^T\in\mathbb{R}^{M\times n}$ are $y(n-1+i),v(n-1+i)$, and $[u(n+i-2)\ \cdots\ u(i-1)]$, respectively. The regularized least squares method to estimate θ is

$$\hat{\theta}_N^R=\arg\min_{\theta}\|Y_N-\Phi_N^T\theta\|_2^2+\sigma^2\theta^T Z^{-1}\theta \tag{4a}$$

$$=Z\Phi_N(\Phi_N^T Z\Phi_N+\sigma^2 I_M)^{-1}Y_N, \tag{4b}$$

where I_M is the M dimensional identity matrix, σ^2 is the noise variance of $v(t)$, $Z\in\mathbb{R}^{n\times n}$ and $Z\succ 0$ is the regularization matrix and also often called kernel matrix, and $\hat{\theta}_N^R$ is the regularized least squares estimate of the impulse response θ.

The quality of $\hat{\theta}_N^R$ depends on Z, Chen et al. (2012). Thus Z has to be designed carefully. It is assumed to take the form of $Z=c\bar{Z}$ with $\bar{Z}\in\mathbb{R}^{n\times n}$ and its (i,j)th element defined as

$$\bar{Z}_{i,j}=K(i,j),\ i,j=1,\cdots,n \tag{5}$$

where $c>0$ is a scaling factor and $K:\{1,\cdots,n\}\times\{1,\cdots,n\}\to\mathbb{R}$ is called a kernel structure.

2.2 Existing fundamental kernel structures

There exist two families of fundamental kernel structures (simply called kernels) suitable for impulse response estimation:

Stable spline (SS) kernels [Pillonetto and De Nicolao (2010)] The SS kernels are constructed based on spline kernels Wahba (1990), which are widely used when the unknown function and some of its derivatives are known or assumed to be continuous with bounded energy. The lth order spline kernel is defined as

$$\bar{K}_l(t,s)=\int_0^1 G_l(s,\tau)G_l(t,\tau)d\tau \tag{6}$$

where $0\le t,s\le 1$, $G_l(r,\tau)=(r-\tau)^{l-1}/(l-1)!$ for $r\ge\tau$ and $G_l(r,\tau)=0$ otherwise. The corresponding lth order SS kernel is then defined as $K_l^{ss}(t,s)=\bar{K}_l(e^{-\beta t},e^{-\beta s})$ for $t,s\ge 0$ and $\beta>0$. The role of $e^{-\beta t}$ and $e^{-\beta t}$ is on the one hand to guarantee that $K_l^{ss}(t,s)$ is well-defined and on the other hand to ensure the stability of $K_l^{ss}(t,s)$, which will be discussed in details later in Section 3. In particular, for $l=1,2$,

$$K_1^{ss}(t,s)=\min(e^{-\beta t},e^{-\beta s}), \tag{7a}$$

$$K_2^{ss}(t,s)=e^{-\beta(t+s)}\min(e^{-\beta t},e^{-\beta s})/2-\min(e^{-3\beta t},e^{-3\beta s})/6. \tag{7b}$$

By constraining $t,s=1,\cdots,n$, the stable spline kernels (7) can be used for discrete time system identification.

Diagonal correlated (DC) kernels [Chen et al. (2012)] Under the assumption that the true system $G_0(q)$ in (1) can be described as an nth order FIR model, the optimal kernel Z^{Opt} in the sense of minimizing the MSE[2] of $\hat{\theta}_N^R$ takes the form of $Z^{Opt}=\theta_0\theta_0^T$, where θ_0 is the true impulse response. Although Z^{Opt} cannot be used in practice, making use of the structure of Z^{Opt} and the prior knowledge that the impulse response of LTI stable systems decays exponentially gives some ideas about how to parameterize the kernel, leading to the DC kernel

$$K^{dc}(i,j)=\rho^{|i-j|}\lambda^{(i+j)/2} \tag{8a}$$

where $i,j=1,\cdots,n$, $0\le\lambda\le 1$, $|\rho|\le 1$, λ controls the decay rate of the impulse response and ρ controls the correlation between the impulse response coefficients. When $\rho=\lambda^{1/2}$, the DC kernel becomes the tuned-correlated (TC) kernel

$$K^{tc}(i,j)=\min(\lambda^i,\lambda^j) \tag{8b}$$

Interestingly, if we choose $\lambda=e^{-\beta}$, the TC kernel is same as the first order SS kernel (7) with $t,s=1,\cdots,n$. If ρ is restricted to be positive, the DC kernel becomes the positive DC kernel

$$K^{pdc}(i,j)=\rho^{|i-j|}\lambda^{(i+j)/2},\ \rho>0 \tag{8c}$$

Remark 2.1. Besides the two families of fundamental kernels, there also exists a family of composite kernels, the so-called multiple kernel, which is introduced in Chen et al. (2014) for both model estimation and structure detection. The multiple kernel is a conic combination of some suitably chosen fixed kernels, which can be instances of SS, DC and the rank-1 kernels in Chen et al. (2013). Multiple kernels can yield better model estimates than fundamental kernels for systems with complicated dynamics, e.g. with several widely spread time constants, and have a couple of features, leading to accurate and efficient algorithms, and applications in various structure detection problems in system identification, see Chen et al. (2014) for details. It should be noted that the idea to use multiple kernel is also briefly mentioned in Dinuzzo (2012), see Chen et al. (2014) for comparisons.

3. STATE-SPACE MODEL INDUCED KERNELS

The problem with SS and DC kernels is that these kernels could be improved for systems with rapid oscillation and/or complicated dynamics Pillonetto et al. (2011); Chen et al. (2014). It is thus interesting and important to design kernels with more flexible and more general structure that are suitable for impulse response estimation. To address this issue, we introduce the so-called state space model induced (SSMi) kernels.

Consider the following state space model

$$dx=Axdt+B(t)dW, \quad x(0)\sim\mathcal{N}(0,O), \quad t\ge 0, \tag{9a}$$

$$g=Cx \tag{9b}$$

which contains an Itô stochastic differential equation (SDE) (9a) and an algebraic output equation (9b). Here, $x\in\mathbb{R}^p, W\in$

[2] More discussions about the MSE of the kernel-based regularization method can be found in Carli et al. (2012); Aravkin et al. (2012).

$\mathbb{R}^m, g \in \mathbb{R}$ are the state, disturbance and output, respectively, and $A \in \mathbb{R}^{p\times p}, C \in \mathbb{R}^{1\times p}$ [3] and $B(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^{p\times m}$. While x, g and W are all functions of time t, it is customary to omit their dependence on t when writing the state space model (9). Throughout the paper, assume that $B(t)$ is measurable and bounded for $t \geq 0$, $x(0)$ is independent of the disturbance $W(t)$, which is a Wiener process (also called Brownian motion in Physics). Recall that an $\mathbb{R}^m$-valued stochastic process $W(t)$ is called a Wiener process if

- $W(0) = 0$ a.s.;
- $W(t) - W(s) \sim \mathcal{N}(0, (t-s)I_m)$ for all $t \geq s \geq 0$, where $\mathcal{N}(0, (t-s)I_m)$ is the m-dimensional Gaussian distribution with mean zero and covariance matrix $(t-s)I_m$;
- $W(t)$ has independent increment property, i.e., for all time instants $0 < t_1 < t_2 < \cdots < t_l$, $W(t_1)$, $W(t_2) - W(t_1), \cdots, W(t_l) - W(t_{l-1})$ are independent.

Lemma 3.1. Consider the state space model (9). The following results hold:

1) The SDE (9a) has a unique solution $x(t)$, which is continuous in t, adapted to the filtration generated by x_0 and $W(s)$ with $0 \leq s \leq t$, and satisfies $E\int_0^t ||x(s)||_2^2 ds < \infty$.
2) The output $g(t)$ is a zero mean Gaussian process with covariance function (also often called kernel)
$$K(t,s) = Eg(t)g(s) = C\Big\{\exp(At)O\exp(As)^T + \int_0^{\min\{t,s\}} \exp(A(t-\delta))B(\delta)B(\delta)^T \exp(A(s-\delta))^T d\delta\Big\}C^T \quad (10)$$
which is referred to as the SSMi kernel.

Remark 3.1. In (9), if we let
$$A = \begin{bmatrix} 0_{(l-1)\times 1} & I_{l-1} \\ 0 & 0_{1\times(l-1)} \end{bmatrix},\ B(t) = \begin{bmatrix} 0_{(l-1)\times 1} \\ 1 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 0_{1\times(l-1)} \end{bmatrix},\ O = 0,$$
then $g(t)$ is in particular the so-called $l-1$ fold integrated Wiener process Shepp (1966) and moreover, its covariance function (10) is nothing but the lth order spline kernel (6). Noting Section 2.2 and the fact that integrated Wiener processes associated with spline kernels have state space model representations, it is natural and interesting to ask whether there exist state space model representations for the stochastic processes associated with the SS and DC kernel structures. We will come back to this question in the following sections.

3.1 Stable kernels

Our goal is to construct kernels suitable for impulse response estimation of LTI stable systems. To achieve this goal, we need some concepts and tools from the theory of reproducing kernel Hilbert space (RKHS), see Aronszajn (1950) for details and also Dinuzzo (2012); Bottegal and Pillonetto (2013).

Definition 3.1. An RKHS over $\mathbb{R}_{\geq 0}$ is a Hilbert space of functions $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that
$$\forall t \geq 0, \quad \exists 0 < d_t < \infty \quad s.t. \quad |g(t)| \leq d_t ||g||_{\mathscr{H}}, \forall g \in \mathscr{H}$$
where $||g||_{\mathscr{H}} = \langle g, g\rangle_{\mathscr{H}}$ is the induced norm with $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ being the inner product over $\mathbb{R}$ associated with $\mathscr{H}$.

Definition 3.2. A function $K : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is called a positive semidefinite kernel if $K(s,t) = K(t,s)$ for any $t, s \geq 0$ and if for any $l \in \mathbb{N}$,
$$\sum_{i=1}^{l}\sum_{j=1}^{l} d_i d_j K(t_i, t_j) \geq 0, \forall t_k \geq 0, d_k \in \mathbb{R}, \quad k = 1, \cdots, l. \quad (11)$$
If, in addition, the first inequality holds in (11) only when $d_k = 0, k = 1, \cdots, l$, then K is called a positive kernel. Given a positive semidefinite kernel K and $t \geq 0$, the kernel section K_t of K located at t is defined as $K_t(s) = K(t,s)$, for $s \geq 0$.

Theorem 3.1. (Moore-Aronszajn Theorem). To every RKHS $\mathscr{H}$ there corresponds a unique positive semidefinite kernel K, called the reproducing kernel, such that
$$g(t) = \langle g, K_t\rangle_{\mathscr{H}}, \qquad \forall t \geq 0, g \in \mathscr{H} \quad (12)$$
Conversely, given a positive semidefinite kernel K, there exists a unique RKHS of functions $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$, whose reproducing kernel is K.

In what follows, we regard the functions $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ as impulse responses of LTI stable systems. As well-known, an LTI system is stable if and only if
$$\int_0^\infty |g(t)| dt < \infty, \quad (13)$$
or in other words, $g \in \mathscr{L}^1$, where $\mathscr{L}^1$ is the space of functions $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ that satisfy (13). Moore-Aronszajn Theorem affirms the connection between RKHS and positive semidefinite kernels. So instead of searching for the functions $g \in \mathscr{L}^1$ directly, we can try to search or design a kernel K such that its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$. In this regard, the following lemma (Dinuzzo, 2012, Lemma 2), which is an immediate corollary of (Carmeli et al., 2006, Propostion 4.4) is very useful.

Lemma 3.2. (Carmeli et al. (2006), Dinuzzo (2012)). For a given positive semidefinite kernel K, its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$ if and only if
$$\int_0^\infty \left|\int_0^\infty K(t,s)h(t)dt\right| ds < \infty, \quad \forall h \in \mathscr{L}^\infty, \quad (14)$$
where $\mathscr{L}^\infty$ is the space of functions $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that there exists a positive number $0 < \varepsilon < \infty$, with $|h(t)| < \varepsilon$ for all $t \geq 0$.

Now we go back to the state space model (9) and consider the kernel (10). For simplicity, define
$$P(t,s) = \int_0^{\min\{t,s\}} \exp(A(t-\delta))B(\delta)B(\delta)^T \exp(A(s-\delta))^T d\delta.$$
Then by Lemma 3.2, conditions on A and $B(t)$ in (9) can be given such that its corresponding RKHS is a subspace of $\mathscr{L}^1$.

Proposition 3.1. Consider the state space model (9) with stable A, i.e., A has all eigenvalues with strictly negative real parts. The output $g(t)$ is a zero mean Gaussian process with kernel $K(t,s)$ as defined in (10). Then its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$ if and only if A and $B(t)$ are such that
$$\int_0^\infty \left|\int_0^\infty C\left\{\exp(At)O\exp(As)^T + P(t,s)\right\}C^T h(t)dt\right| ds < \infty,\ \forall h \in \mathscr{L}^\infty \quad (15)$$

Corollary 3.1. Consider the state space model (9) with stable A. The output $g(t)$ is a zero mean Gaussian process with kernel $K(t,s)$ as defined in (10). Then the following results hold:

1) Its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$ if A and $B(t)$ are such that
$$\int_0^\infty \int_0^\infty \left|CP(t,s)C^T\right| dtds < \infty \quad (16)$$
2) If $CP(t,s)C^T \geq 0$ for any $t, s \geq 0$, its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$ if and only if A and $B(t)$ are such that

[3] The output matrix C can of course contain multiple rows, which is critical for multiple output system identification to be studied in the near future.

$$\int_0^\infty \int_0^\infty CP(t,s)C^T dtds < \infty \tag{17}$$

3) If $x \in \mathbb{R}$, its corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$ if and only if A and $B(t)$ are such that

$$\int_0^\infty \int_0^\infty P(t,s)dtds < \infty \tag{18}$$

It can happen that $\mathscr{H}_K \nsubseteq \mathscr{L}^1$ for a given kernel K, but $\mathscr{H}_K \cap \mathscr{L}^1$ may not be empty. The problem of using such kernels for impulse response estimation is that it is tricky to deal with the functions of $\mathscr{H}_K$ that do not belong to $\mathscr{L}^1$. In contrast, it is more convenient to work only on kernels K, whose corresponding RKHS $\mathscr{H}_K \subset \mathscr{L}^1$. This kind of kernels are referred to as SSMi stable (SSMiS) kernels below. Note that the concept of stable kernel is introduced in Pillonetto and De Nicolao (2010) and further elaborated in Dinuzzo (2012).

3.2 Some preliminary 1st order SSMiS kernels

Proposition 3.2. Consider the state space model (9) with $x \in \mathbb{R}$. Then the following results hold:

1) Let $A = -\beta, B(t) = \exp(-\alpha t)$, and $O = \gamma$ with $\alpha, \beta, \gamma > 0$. Then $g(t)$ is a zero mean Gaussian process with *stable* kernel

$$K(t,s) = C^2\gamma e^{-\beta(t+s)} + C^2 \frac{1}{2(\beta-\alpha)} e^{-\beta(t+s)}(e^{2(\beta-\alpha)\min\{t,s\}} - 1) \tag{19a}$$

When $\beta = 2\alpha = \gamma^{-1} = C^2$, the kernel (19a) becomes $K(t,s) = \min\{e^{-\beta t}, e^{-\beta s}\}$, which is the 1st order SS kernel (7a), and TC kernel (8b) with $\lambda = e^{-\beta}$ and $t,s = 1,\cdots,n$. When $\alpha, \beta - \alpha > 0$ and $2(\beta-\alpha) = \gamma^{-1} = C^2$, the kernel (19a) becomes the positive DC kernel (8c) with $\lambda = e^{-2\alpha}$ and $\rho = e^{-(\beta-\alpha)}$ and $t,s = 1,\cdots,n$.

2) Let $A = -\beta, B(t) = 0, C = 1$ and $O = 1$ with $\beta > 0$. Then $g(t)$ is a zero mean Gaussian process with *stable* kernel $K(t,s) = e^{-\beta(t+s)}$, which is the exponential kernel studied in Dinuzzo (2012).

3) Let $A = -\beta, B(t) = \exp(-\beta t)/(t+1), C = 1, O = \gamma$ with $\beta, \gamma > 0$. Then $g(t)$ is a zero mean Gaussian process with *stable* kernel

$$K(t,s) = \gamma e^{-\beta(t+s)} + e^{-\beta(t+s)}(1 - \frac{1}{\min\{t,s\}+1}) \tag{19b}$$

Remark 3.2. The 1st order SSMiS kernels are constructed based on 1st order state space model (9), where the system without the stochastic disturbance has a negative real pole. Intuitively, this may not be favorable for identification of systems with strong oscillations. In this case, it is straightforward to introduce a 2nd order SSMiS kernel where the system in (9) without the stochastic disturbance has a pair of complex conjugate poles with negative real parts. Consider the state space model (9) with $x \in \mathbb{R}^2$. Let $A = \begin{bmatrix} 0 & 1 \\ -\alpha^2-\beta^2 & -2\alpha \end{bmatrix}$, $B(t) = [0 \quad \exp(-\gamma t)]^T$, $C = [1\ 0]$ and $O = 0$ with $\alpha, \beta, \gamma > 0$. Then $g(t)$ is a zero mean Gaussian process with *stable* kernel

$$\begin{aligned} K(t,s) = {} & \frac{1}{4\beta^2(\alpha-\gamma)} e^{-\alpha(t+s)} \cos\beta(t-s)(e^{2(\alpha-\gamma)\min\{t,s\}} - 1) \\ & - \frac{1}{2\beta^2\sqrt{4\beta^2+4(\alpha-\gamma)^2}} e^{-\alpha(t+s)} \\ & \times [e^{2(\alpha-\gamma)\min\{t,s\}} \sin(2\beta\min\{t,s\} + \theta - \beta(t+s)) - \sin(\theta - \beta(t+s))] \end{aligned} \tag{20}$$

Our preliminary simulation results show that the kernel (20) can give more accurate and robust model estimates than the existing SS and DC kernels for systems with strong oscillations. Due to limitation of space, we are unable to give the details here and will include the complete discussions in an independent paper.

4. HYPER-PARAMETER ESTIMATION BY PROFILE MARGINAL LIKELIHOOD MAXIMIZATION

In this section, we study the hyper-parameter estimation problem for a given kernel K. Denote the hyper-parameter used to parameterize K by η. Here, η is composed of β for (7), λ, ρ for (8), and α, β, γ for (19). Then the kernel matrix Z in (4a) can be written as $Z = c\bar{Z}(\eta)$ where from (5), the (i,j)th element of $\bar{Z}(\eta)$ is $K(i,j)$. There exist several ways to estimate c, η. Currently, the most widely used one is to embed the regularization term $\theta^T Z^{-1}\theta$ in (4a) in a Bayesian framework and estimate c, η by maximizing the marginal likelihood.

To be specific, assume $v(t)$ in (1) is Gaussian distributed and $\theta \sim \mathscr{N}(0, c\bar{Z}(\eta))$. Then the maximum a posteriori (MAP) estimation problem $\arg\max_\theta p(\theta|Y_N)$ is equivalent to (4a). Note that $p(Y_N|c,\eta) = \mathscr{N}(0, \Phi_N^T c\bar{Z}(\eta)\Phi_N + \sigma^2 I_M)$, then the marginal likelihood maximization method $\arg\max_{c,\eta} p(Y_N|c,\eta)$ to estimate c, η is equivalent to

$$\arg\min_{c,\eta} Y_N^T(\Phi_N^T c\bar{Z}(\eta)\Phi_N + \sigma^2 I_M)^{-1} Y_N + \log|\Phi_N^T c\bar{Z}(\eta)\Phi_N + \sigma^2 I_M|. \tag{21}$$

A tricky problem for (21) is how to handle the unknown σ^2. In Pillonetto and De Nicolao (2010); Chen et al. (2012), a low-bias ARX model or a FIR model is first estimated and then its sample variance $\hat{\sigma}^2$ is used as an estimate of σ^2 and finally η is estimated according to (21) with σ^2 replaced by $\hat{\sigma}^2$.

Here we treat σ^2 as an additional "hyper-parameter" and estimate it together with η by maximizing the marginal likelihood, MacKay (1992). Instead of directly estimating it, we define $\bar{c} = c/\sigma^2$. The cost function of (21) is written as $l(\bar{c}, \eta, \sigma^2) = Y_N^T(\Phi_N^T \bar{c}\bar{Z}(\eta)\Phi_N + I_M)^{-1} Y_N/\sigma^2 + \log|\Phi_N^T \bar{c}\bar{Z}(\eta)\Phi_N + I_M| + M\log\sigma^2$, which is minimized with respect to σ^2 at

$$\sigma^{2^*} = Y_N^T(\Phi_N^T \bar{c}\bar{Z}(\eta)\Phi_N + I_M)^{-1} Y_N/M \tag{22}$$

Replacing σ^2 with σ^{2^*} in $l(\bar{c}, \eta, \sigma^2)$ yields,

$$\begin{aligned} l_{\text{profile}}(\bar{c}, \eta) = {} & M\log Y_N^T(\Phi_N^T \bar{c}\bar{Z}(\eta)\Phi_N + I_M)^{-1} Y_N \\ & + \log|\Phi_N^T \bar{c}\bar{Z}(\eta)\Phi_N + I_M| + M - M\log M \end{aligned} \tag{23}$$

which is essentially the so-called profile log-likelihood or concentrated log-likelihood, Venzon and Moolgavkar (1988); Murphy and Van der Vaart (2000). The hyper-parameters $\bar{c}, \eta$ are then estimated by minimizing (23), i.e.,

$$\tilde{c}, \tilde{\eta} = \arg\min_{\bar{c},\eta} l_{\text{profile}}(\bar{c}, \eta) \tag{24}$$

With $\tilde{c}, \tilde{\eta}$, the regularized least squares estimate (4b) is $\hat{\theta}^R = (\Phi_N\Phi_N^T + (\tilde{c}\bar{K}(\tilde{\eta}))^{-1})\Phi_N Y_N$.

5. INITIALIZATION

As seen from (3), for $t = 0,\cdots,n-1$, the output $y(t)$ depends on the unknown initial values $u(t-n),\cdots,u(-1)$. In our previous works Chen et al. (2012, 2014) and also in Section 2.1, for simplicity, we chose not to use the first n outputs $y(t)$, $t = 0,\cdots,n-1$ and start from $y(n)$. This way of handling the unknown initial conditions is called "non-windowing", see e.g. Ljung (1999). When the data length N is small (compared to the FIR model order n), the non-windowing method may not work well. In what follows, two other methods are considered.

5.1 Pre-windowing

For $t = 0, \cdots, n-1$, this method simply replaces the unknown initial values $u(t-n), \cdots, u(-1)$ by zeros. In this case, the quantities M, Y_N, Φ_N and V_N in Sections 2.1 and 4 should be redefined as follows: $M = N$, $Y_N, V_N \in \mathbb{R}^N$ with the ith element of Y_N and V_N being $y(i)$ and $v(i)$ respectively, and $\Phi_N^T \in \mathbb{R}^{N\times n}$ with the ith row $[u(i-1) \ \cdots \ u(i-n)]$.

5.2 Estimating the transient as an additional regularized FIR with impulsive input

System (1) can be realized in state space form:

$$\begin{aligned} x(t+1) &= A_d x(t) + B_d u(t), \\ y(t) &= C_d x(t) + v(t), \ t \geq 0, \end{aligned} \tag{25}$$

which yields $y(t) = C_d A_d^t x(0) + \sum_{k=0}^{t-1} C_d A_d^{t-1-k} B_d u(k) + v(t)$. For $t = 1, \cdots, N$, $y(t)$ depends on the unknown initial condition $x(0)$. Then it is easy to see that the system

$$\begin{aligned} z(t+1) &= A_d z(t) + B_d u(t) + A_d(x(0) - z(0))\delta(t), \\ y(t) &= C_d z(t) + C_d(x(0) - z(0))\delta(t) + v(t), \ t \geq 0, \end{aligned} \tag{26}$$

where $\delta(t) = 1$ for $t = 0$ and $\delta(t) = 0$ for $t \neq 0$, has the same output $y(t)$ as (25). On the one hand, the role of $z(0)$ is to represent our guess on the unknown $x(0)$, which may not be right. On the other hand, the transient $C_d A_d(x(0) - z(0))$ in $y(t)$ due to the guess error of the initial value $x(0) - z(0)$ can then be captured by an impulse response from an additional input which is an impulse $\delta(t)$; see also Ljung (2004).

Motivated by the above observations, system (1) is written as

$$y(t) = G_0(q)u(t) + G^*(q)\delta(t) + v(t), \ t = 1, \cdots, N \tag{27}$$

where $G^*(q)$ is the transient dynamics due to the guess error of the initial value $x(0) - z(0)$. When computing $G_0(q)u(t)$, we need to know $u(t)$ for $t < 0$, which is chosen by the users and actually reflects the users' guess $z(0)$ on the unknown $x(0)$.

Again high order FIR models are used to model $G_0(q)$ and $G^*(q)$: $G(q, \theta_1)$ for $G_0(q)$ and $G(q, \theta_2)$ for $G^*(q)$, where θ_1 and θ_2 are the FIR coefficient vectors of $G(q, \theta_1)$ and $G(q, \theta_2)$, respectively. Now consider (4a). Define $\theta^T = \begin{bmatrix} \theta_1^T & \theta_2^T \end{bmatrix} \in \mathbb{R}^{2n}$. Then the quantities M, Y_N, Φ_N and V_N in Sections 2.1 and 4 should be redefined as follows: $M = N$, $Y_N, V_N \in \mathbb{R}^N$ with the ith element of Y_N and V_N being $y(i)$ and $v(i)$ respectively, and $\Phi_N^T \in \mathbb{R}^{N\times 2n}$ with the ith row $[u(i-1) \ \cdots \ u(i-n) \ \nu_i]$ where for $i \leq n$ all elements of $\nu_i^T \in \mathbb{R}^n$ are zero except the ith one and for $i > n$, $\nu_i = 0$. In this paper, the kernel matrix Z in (4a) is assumed, for simplicity, to have a block diagonal structure, i.e., $Z = \text{diag}(Z_1, Z_2)$ (the regularization term $\theta^T Z \theta = \theta_1^T Z_1 \theta_1 + \theta_2^T Z_2 \theta_2$, where the parameterization of Z_i, $i = 1, 2$ should be done as what described in Sections 2.2 and 3. Solving the hyper-parameter estimation problem in Section 4 and then the regularized least squares problem in Section 2.1 yields the regularized FIR model estimate $\hat{\theta}_{1,N}^R$ of θ_1 and the corresponding FIR model estimate $G(q, \hat{\theta}_{1,N}^R)$ for $G_0(q)$.

6. NUMERICAL SIMULATION

6.1 Test data-bank

To examine the proposed stable kernels and also the proposed methods for hyper-parameter estimation and initialization, we regenerate the data-bank in Chen et al. (2012) and revisit the data bank in Chen et al. (2014):

- D1: 1000 fast systems, data sets with $N = 210$, SNR=10
- D2: 1000 fast systems, data sets with $N = 210$, SNR=1
- D3: 2500 fast systems, data sets with $N = 500$, SNR=10
- D4: 2500 slow systems, data sets with $N = 500$, SNR=10
- D5: 2500 fast systems, data sets with $N = 375$, SNR=1
- D6: 2500 slow systems, data sets with $N = 375$, SNR=1

All systems are randomly generated 30th order stable discrete-time systems. The fast systems have all poles inside the circle with center at the origin and radius 0.95 and the slow systems have at least one pole outside this circle. The signal to noise ratio (SNR) is defined as the ratio of the variance of the noise-free output over the variance of the white Gaussian noise. In all cases the input is Gaussian random signal with unit variance. Here D1 and D2 are the data collections in Chen et al. (2014). D2 to D6 are regenerated and correspond to data collections S1D1, S1D2, S2D1 and S2D2 in Chen et al. (2012), respectively. The difference is that all data in the regenerated D2 to D6 are collected after getting rid of initial conditions effect. More details can be found in Chen et al. (2012, 2014).

6.2 Simulation setup

For data collections D1 and D2, we estimate FIR model (2) with $n = 100$ and for D3 to D6, we estimate (2) with $n = 125$.

For illustrations, we examine the following three stable kernels:

- (8b) with $\eta = \lambda$ and is denoted by 'TC';
- (19a) with $\eta = \beta$ and $\beta = 2\alpha = 2/\gamma = C^2$, and is denoted by 'SSMiS1';
- (19b) with $\eta = \beta$ and $\gamma = 0$ and is denoted by 'SSMiS2'.

Two methods for the hyper-parameter estimation are examined here: the marginal likelihood maximization (21) and the profile marginal likelihood maximization (24). The implementation details can be found in Chen and Ljung (2013).

Three kinds of initialization scheme are compared here:

- Non-windowing method: it is denoted by 'NW'.
- Pre-windowing method: it is denoted by 'PW';
- Estimation of the unknown transient by a regularized FIR with an impulsive input: it is denoted by 'EST'.

For all estimated FIR models, the model fit is defined as

$$W = 100\left(1 - \left[\frac{\sum_{k=1}^n |g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^n |g_k^0 - \bar{g}^0|^2}\right]^{1/2}\right), \quad \bar{g}^0 = \frac{1}{n}\sum_{k=1}^n g_k^0$$

where for $k = 1, \cdots, n$, g_k^0 and $\hat{g}_k$ represent the true and the regularized impulse response estimate, respectively.

6.3 Simulation results

The average model fits are reported in the following two tables.

Marginal likelihood maximization

	D1	D2	D3	D4	D5	D6	OD
TC + NW	81.5	56.9	90.3	72.2	66.3	42.5	68.3
TC + PW	85.2	68.4	91.1	76.0	68.9	**54.8**	74.1
TC + EST	84.4	65.4	91.1	75.7	68.1	49.3	72.3
SSMiS1 + NW	81.6	56.4	90.3	72.3	66.8	44.3	68.6
SSMiS1 + PW	85.1	68.4	91.1	76.0	**69.3**	54.2	74.0
SSMiS1 + EST	84.4	65.5	91.1	75.9	68.7	50.5	72.7
SSMiS2 + NW	83.8	60.4	91.6	74.7	65.9	45.7	70.3
SSMiS2 + PW	86.8	**71.0**	92.2	78.2	67.8	51.7	**74.6**
SSMiS2 + EST	**86.9**	69.2	**92.5**	**78.3**	67.9	47.3	73.7
OKM	84.4	64.6	91.3	75.5	67.7	48.9	72.1

Profile marginal likelihood maximization

	D1	D2	D3	D4	D5	D6	OD
TC + NW	82.5	60.3	90.3	72.2	64.9	41.7	68.6
TC + PW	85.3	68.5	91.1	76.0	67.9	51.7	73.4
TC + EST	86.1	68.5	91.2	75.9	67.2	48.9	73.0
SSMiS1 + NW	82.5	60.1	90.3	72.3	66.4	43.8	69.2
SSMiS1 + PW	85.3	68.5	91.1	76.0	**69.2**	**53.8**	**74.0**
SSMiS1 + EST	86.2	68.4	91.2	76.0	68.2	49.7	73.3
SSMiS2 + NW	84.6	62.8	91.6	74.6	63.9	40.4	69.7
SSMiS2 + PW	86.9	71.0	92.2	78.2	66.5	49.4	**74.0**
SSMiS2 + EST	**88.2**	**71.3**	**92.5**	**78.3**	67.4	46.1	**74.0**
OKM	85.3	66.6	91.3	75.5	66.8	47.3	72.2

In the above tables, the column "OD" shows the average model fits over the 6 data collections and the row "OKM" shows the average model fits over all tested kernels and methods for initialization, and the largest average model fit for each column in the table is written in bold.

6.4 Findings

We have the following empirical findings. First, SSMiS1 and SSMiS2 kernels perform on the average a bit better than TC kernel. SSMiS2 kernel is the best kernel over all data collections except for slow systems (D5 and D6), for which TC and SSMiS1 kernels are better choices. Second, the two methods pre-windowing and estimation of the transient yield much better performance than the non-windowing method for all data collections. In particular, the pre-windowing method is a better choice on the average. Third, In contrast with the marginal likelihood method, the profile marginal likelihood maximization method works better for fast systems and very short data (D1 and D2), works comparably for fast systems and long data (D3 and D4), works worse for slow systems and moderately short data (D5 and D6). However, the worse performance in the last case may be because of too low FIR model order.

7. CONCLUSION AND FUTURE WORKS

In this contribution, we introduced a new fundamental kernel structure induced by stochastic state space models and its several preliminary 1st order and 2nd order instances. An outstanding topic in regularized system identification is whether it is beneficial to make use of the correlation between different dynamics associated with different inputs or outputs for multi-input and/or multi-output system identification. The reason why this topic has not been resolved so far is to a large extent because there is no cheap way (in terms of the number of hyper-parameters) to design the correlation between the different dynamics so that the positive semi-definiteness of the overall kernel is guaranteed. This obstacle is however swept away, with the state space model induced stable kernel proposed in this contribution. Indeed, it is straightforward to choose $A, B(t)$ and C with multiple rows, so that the corresponding kernel is not only positive semi-definite but also stable, provided that the conditions of Proposition 3.1 and Corollary 3.1 are satisfied. Now it remains to answer if there exist suitable $A, B(t)$ and C such that making use of the correlation between different dynamics is beneficial, which is under investigation. Another topic is to study regularized marginally stable or even unstable system identification by a careful design of $A, B(t)$ and C.

REFERENCES

Aravkin, A., Burke, J., Chiuso, A., and Pillonetto, G. (2012). On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum mse. In *IFAC symposium on system identification*.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3), 337–404.

Bottegal, G. and Pillonetto, G. (2013). Regularized spectrum estimation using stable spline kernels. *Automatica*, 49(11), 3199–3209.

Carli, F., Chen, T., Chiuso, A., Ljung, L., and Pillonetto, G. (2012). On the estimation of hyperparameters for bayesian system identification with exponentially decaying kernels. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, 5260–5265.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04), 377–408.

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

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

Chen, T., Chiuso, A., Pillonetto, G., and Ljung, L. (2013). Rank-1 kernels for regularized system identification. In *IEEE Conference on Decision and Control*, 5162–5167. Florence, Italy.

Chen, T. and Ljung, L. (2013). Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7), 2213–2220.

Dinuzzo, F. (2012). Kernels for linear time invariant system identification. *CoRR*, abs/1203.4930.

Ljung, L. (1999). *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition.

Ljung, L. (2004). State of the art in linear system identification: Time and frequency domain methods. In *Proc. American Control Conference*. Boston, MA.

MacKay, D.J.C. (1992). Bayesian interpolation. *Neural Computation*, 4(3), 415–447.

Murphy, S.A. and Van der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450), 449–465.

Pillonetto, G., Chiuso, A., and De Nicolao, G. (2011). Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2), 291–305.

Pillonetto, G. and De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*.

Pintelon, R. and Schoukens, J. (2012). *System identification: a frequency domain approach*. John Wiley & Sons, 2nd edition.

Shepp, L. (1966). Radon-nikodym derivatives of gaussian measures. *The Annals of Mathematical Statistics*, 37(2), 321–354.

Söderström, T. and Stoica, P. (1989). *System Identification*. Prentice-Hall Int., London.

Venzon, D. and Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 87–94.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.