ANNUAL REVIEWS

Annual Review of Control, Robotics, and Autonomous Systems

System Identification: A Machine Learning Perspective

A. Chiuso and G. Pillonetto

Department of Information Engineering, University of Padova, 35131 Padova, Italy; email: chiuso@dei.unipd.it

Annu. Rev. Control Robot. Auton. Syst. 2019. 2:281–304

First published as a Review in Advance on November 28, 2018

The *Annual Review of Control, Robotics, and Autonomous Systems* is online at control.annualreviews.org

https://doi.org/10.1146/annurev-control-053018-023744

Keywords

system identification, machine learning, kernel methods, Gaussian processes

Abstract

Estimation of functions from sparse and noisy data is a central theme in machine learning. In the last few years, many algorithms have been developed that exploit Tikhonov regularization theory and reproducing kernel Hilbert spaces. These are the so-called kernel-based methods, which include powerful approaches like regularization networks, support vector machines, and Gaussian regression. Recently, these techniques have also gained popularity in the system identification community. In both linear and nonlinear settings, kernels that incorporate information on dynamic systems, such as the smoothness and stability of the input–output map, can challenge consolidated approaches based on parametric model structures. In the classical parametric setting, the complexity of the model (the model order) needs to be chosen, typically from a finite family of alternatives, by trading bias and variance. This (discrete) model order selection step may be critical, especially when the true model does not belong to the model class. In regularization-based approaches, model complexity is controlled by tuning (continuous) regularization parameters, making the model selection step more robust. In this article, we review these new kernel-based system identification approaches and discuss extensions based on nuclear and ℓ_1 norms.

1. INTRODUCTION

Function estimation plays a central role in machine learning. The problem consists of discovering mathematical relationships between data. In the learning-from-examples scenario, we are given the so-called training set containing couples (x_i, y_i), where x_i is often called the input location and y_i is the corresponding output. The key task is to infer a function g that is able to well predict future data so that, for a new pair (x, y), the scalar $g(x)$ is close (in some sense) to y.

A first approach to solve this problem is to use parametric methods. We can introduce a finite-dimensional vector θ of size m to parameterize the unknown function, e.g., using a polynomial model $g_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{m-1} x^{m-1}$. Estimation of θ can then be performed through well-known paradigms taken from the statistical literature, such as least squares and maximum likelihood. An important issue is the selection of model complexity, which, in our simple example, is regulated by m. This selection can be accomplished by means of classical criteria such as the Akaike or Bayesian information criterion (1, 2) or cross-validation techniques (3).

A second approach exploits nonparametric estimation, where the function g is assumed to belong to an infinite-dimensional (or high-dimensional) space. Some form of regularization is then introduced to control the complexity of the solution. This paradigm has been widely adopted in the machine learning literature, where algorithms abound and hinge on, e.g., generalized additive models, trees, boosting, and nearest-neighbors techniques (see, e.g., 4). Many of these approaches can also be cast under the framework of artificial neural networks (5, 6), which have recently garnered renewed interest thanks to deep networks' success in classification and pattern recognition (7).

In this review, we focus on a particular nonparametric algorithmic class obtained by coupling Tikhonov regularization theory (8, 9) with reproducing kernel Hilbert spaces (RKHSs) (10, 11). RKHS applications in statistics, approximation theory, and computer vision are numerous and trace back to References 12–14, and RKHSs were introduced to the machine learning community in Reference 15. The popularity of RKHSs stems from many important and useful properties, such as their one-to-one correspondence with the class of positive definite kernels and the fundamental connections with Gaussian processes (16–19). Such features have allowed the derivation of a very rich class of learning machines, the so-called kernel-based methods (20, 21), which include smoothing splines and regularization networks (13, 14), support vector machines (22, 23), and Gaussian regression (24). In all cases, the unknown function is the minimizer of a suitable objective over an RKHS. More specifically, the estimator's structure is

$$\hat{g} = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{V}_i(y_i, L_i[f]) + \gamma \|f\|_{\mathcal{H}}^2, \qquad 1.$$

with the following definitions:

- $\mathcal{H}$ is an RKHS defined by the kernel $\mathcal{K}$, a symmetric and positive definite function, with a norm denoted by $\|\cdot\|_{\mathcal{H}}$. In machine learning, $\mathcal{K}$ is often selected just to include information on function smoothness. An example is the spline kernel (13), which leads to a penalty given by the squared energy of the derivative—i.e., $\|f\|_{\mathcal{H}}^2 = \int \dot{f}^2(x)\mathrm{d}x$.
- $\mathcal{V}_i$ is the so-called loss function, which must measure the adherence of f to experimental data. For classification, the hinge loss is often adopted (4). For regression, Vapnik's ϵ-insensitive loss, e.g., leads to support vector regression (25), while the classical squared loss $\mathcal{V}_i(y_i, L_i[f]) = (y_i - L_i[f])^2$ defines the so-called regularization networks.
- $L_i[f]$ is a linear operator. For example, one can think of $L_i[f] = f(x_i)$, when the unknown map must be inferred from direct measurements.

- γ is the regularization parameter that must balance two contrasting terms: the data fit (regulated by the loss) and the penalty term (given by the RKHS squared norm), which restores the well posedness. Its choice has a major effect on the quality of the estimate since it can well balance bias and variance, leading to an estimator with favorable mean squared error properties. For its tuning, the empirical Bayes approach (26–28) is often adopted.

The interesting features of Equation 1 also include the property of $\hat{g}$ to converge to the optimal predictor (also when $\mathcal{H}$ is infinite-dimensional) as the data set size grows to infinity; Reference 29 is a classic work, and recent developments can be found in References 30–34. This point is also related to Vapnik's concepts of generalization and consistency (23); see Reference 20 for connections among regularization in RKHS, statistical learning theory, and the concept of V_γ dimension (35, 36) as a measure of function class complexity (35, 36). The link between consistency and well posedness is discussed in References 33, 34 and 37.

In parallel with the study and successful application of Equation 1 for both regression and classification (21), the control community has designed many system identification techniques in the last few decades. Here, the problem is to reconstruct a dynamic system fed with a known input from a finite collection of noisy and output samples. In the time-invariant linear scenario, this task corresponds to inferring a particular function known as the system impulse response.

For impulse response estimation, the first regularized approaches trace back to References 38–40; for a thorough overview, see Reference 41. Other techniques where model error is described via a nonparametric structure can be found in References 42 and 43, while approaches relying on nuclear and atomic norms are described in References 44–49. In the nonlinear scenario, several nonparametric approaches have also been exploited for nonlinear system identification, using, e.g., Volterra theory (50); artificial neural networks and wavelets, polynomial functions, or orthonormal functions (51–54); and kernel-type estimators (55–57), with weight optimization used to control the mean squared error (58–60). Least squares support vector machines (61, 62) also represent important links between kernel-based regularization and nonlinear system identification; Gaussian regression for state space models is described in References 63 and 64.

Many of the approaches cited above exploit machine learning concepts, which is a consequence of the fact that predictor estimation is at the core of both the classical system identification paradigm (65) and the machine learning philosophy. However, there are important dynamic systems peculiarities that must be considered when adopting the estimator in Equation 1 in system identification. To obtain good models, the availability of large data sets is also insufficient. In fact, for big data and data science to play a role in system dynamics, it is essential to infer plausible theories and model structures from input–output measurements (66). Interestingly, for this purpose, RKHSs that are, e.g., suited to linear system identification have been proposed only recently, introducing stable-spline kernels, which embed information on impulse response regularity and stability (67–69), or, in computer vision, exploiting compound matrices built from system trajectories (70). The stability of an RKHS (i.e., its property of containing only stable systems or predictors) has been treated only recently in linear settings (71–73); some extensions to the nonlinear setting are discussed in Reference 74.

Our aim in this article is to provide an overview of recent advances connected with the use of Equation 1 in system identification. The main focus is regularization in RKHSs, but we also discuss some extensions where RKHS norms $\|\cdot\|_{\mathcal{H}}$ are replaced by more general penalties. The latter include sparsity-promoting regularizers induced by, e.g., nuclear and ℓ_1 norms. The article is organized as follows. Section 2 provides some background material on RKHSs and the related Bayesian formulation. Section 3 discusses linear system identification in RKHSs, focusing on the case of quadratic losses; we discuss bounded-input, bounded-output (BIBO) stability conditions

and provide examples of stable kernels useful for applications, and simple links with regularized finite impulse response (FIR) estimation also naturally arise. Section 4 describes nonlinear system identification. Section 5 reports generalizations of the regularized least squares estimators, discussing sparsity in system identification related to variable and structure selection. Section 6 concludes the review.

2. PRELIMINARIES

In this section, we discuss some preliminary material on RKHSs, their role in regularized inverse problems, and a Bayesian framework, which will be particularly useful later for tuning hyperparameters (i.e., parameters that describe the model space).

2.1. Regularization in Reproducing Kernel Hilbert Spaces

Let $\mathcal{X}$ denote a nonempty set whose generic elements are often denoted below by x or a. We then use $\mathcal{H}$ to denote a Hilbert space of real-valued functions, which will be the hypothesis space that the candidate function $g : \mathcal{X} \to \mathbb{R}$ belongs to. Since our aim is to introduce some smoothness information on g in our estimator in Equation 1, a first basic requirement is that all the functions in $\mathcal{H}$ be well defined pointwise for any $x \in \mathcal{X}$. An even stronger assumption is that all the pointwise evaluators $g \to g(x)$ are linear and bounded—i.e., $\forall x \in \mathcal{X}$, there exists $C_x < \infty$ so that

$$|g(x)| \leq C_x \|g\|_{\mathcal{H}}, \quad \forall g \in \mathcal{H}. \tag{2.}$$

Note that C_x can depend on x but not on g. This property already leads to the spaces of interest.

Definition 1 (reproducing kernel Hilbert space). An RKHS $\mathcal{H}$ over $\mathcal{X}$ is a Hilbert space of functions $g : \mathcal{X} \to \mathbb{R}$ where Equation 2 holds.

RKHSs are connected to the concept of a positive definite kernel, a particular function defined over $\mathcal{X} \times \mathcal{X}$.

Definition 2 (positive definite kernel and kernel section). A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if, for any integer p, one has

$$\sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j K(x_i, x_j) \geq 0, \quad \forall (x_k, c_k) \in (\mathcal{X}, \mathbb{R}), \quad k = 1, \ldots, p.$$

The kernel section K_x centered at x is the function from $\mathcal{X}$ to $\mathbb{R}$ defined by $K_x(a) = K(a, x)$, $\forall a \in \mathcal{X}$.

The following theorem shows that there exists a one-to-one correspondence between RKHSs and positive definite kernels (10).

Theorem 1 (Moore–Aronszajn and reproducing property). To every RKHS $\mathcal{H}$ there corresponds a unique positive definite kernel K such that the so-called reproducing property holds—i.e.,

$$\langle K_x, g \rangle = g(x), \quad \forall (x, g) \in (\mathcal{X}, \mathcal{H}). \tag{3.}$$

The opposite also holds: For any given positive definite kernel K, there exists a unique RKHS of real-valued functions defined over $\mathcal{X}$ where Equation 3 holds.

Hence, an RKHS $\mathcal{H}$ is completely defined by a kernel $\mathcal{K}$, namely, the reproducing kernel of $\mathcal{H}$. It can be also proved that the RKHS contains all of the finite linear combinations of kernel sections along with some particular infinite sums (72, part II). As a consequence, any $g \in \mathcal{H}$ inherits kernel properties. For example, if $\mathcal{K}$ is a Mercer kernel (i.e., it is also continuous), then all of the $g \in \mathcal{H}$ are continuous (75, p. 35).

We now illustrate the famous representer theorem, which provides a way to numerically compute the solution $\hat{g}$ in Equation 1. The key ingredient is the function $L_i[\mathcal{K}]$, which corresponds to the kernel filtered by the linear functionals entering in Equation 1. Note that it is irrelevant which argument L_i is applied to, as a consequence of the symmetry of $\mathcal{K}$. As an example, if $L_i[f] = f(x_i)$, then $L_i[\mathcal{K}] = \mathcal{K}_{x_i}$—i.e., we obtain the kernel section centered on x_i.

Theorem 2 (representer theorem). If each $L_i : \mathcal{H} \to \mathbb{R}$ is linear and bounded, and the solution of Equation 1 exists, then any solution admits the following expression:

$$\hat{g} = \sum_{i=1}^{N} \hat{c}_i L_i[\mathcal{K}], \qquad 4.$$

where $\hat{c}_i$ represents suitable scalars.

Thus, even if $\mathcal{H}$ is infinite-dimensional, the function estimate $\hat{g}$ belongs to the N-dimensional subspace spanned by the kernel sections $\mathcal{K}_{x_i}$ centered on the input data. One can then plug the expression in Equation 4 into Equation 1, reducing function estimation to a finite-dimensional optimization problem that returns the $\hat{c}_i$. For more general versions of the representer theorem, see Reference 76.

2.2. A Bayesian Framework

The solution $\hat{g}$ to the regularized minimization problem in Equation 1 can also be equivalently written in a Bayesian context under a suitable probabilistic framework where the unknown f is modeled as a stochastic process. This interpretation will be useful in Section 3.5.3 when discussing hyperparameter estimation strategies.

To avoid technical complications in this section, we assume that the loss function $\mathcal{V}_i(y_i, L_i[f])$ is quadratic—i.e., $\mathcal{V}_i(y_i, L_i[f]) = (y_i - L_i[f])^2$. In a probabilistic framework, this is equivalent to assuming a measurement model

$$y_i = L_i[f] + e_i,$$

where the errors e_i are independent and identically distributed zero-mean Gaussian with variance σ^2 and are independent of f. Under this assumption, the conditional distribution of y_i given f is a zero-mean Gaussian with variance σ^2. In addition, we can assume that f is a zero-mean Gaussian process with covariance function $\mathbb{E} f(x_i) f(x_j) = \lambda \mathcal{K}(x_i, x_j)$. Under the assumption that the functional L_i is linear and bounded, $L[f] := [L_1[f], ..L_N[f]]^\top$ is a Gaussian vector, and so is $Y := [y_1, .., y_N]$; in addition, the process (Y, f) is jointly Gaussian, and thus the minimum variance estimator of $f(t)$,

$$\hat{f}(t) = \arg\min_{\eta(Y)} \mathbb{E}[(f(t) - \eta(Y))^2],$$

where minimization is taken with respect to all measurable functions $\eta(\cdot)$ of Y, is the conditional mean—i.e.,

$$\hat{f}(t) = \mathbb{E}[f(t)|Y] = \mathbb{E} f(t) Y^\top \left(\mathbb{E} Y Y^\top\right)^{-1} Y. \qquad 5.$$

Observing that $\mathbb{E} f(t) y_i = L_i[\mathcal{K}(t,:)]$ and defining the coefficient vector

$$\alpha = [\alpha_1, .., \alpha_N]^\top := \left(\mathbb{E} Y Y^\top\right)^{-1} Y \in \mathbb{R}^N,$$

Equation 5 becomes

$$\hat{f}(t) = \mathbb{E}[f(t)|Y] = \sum_{i=1}^{N} \alpha_i L_i[\mathcal{K}(t,:)], \tag{6.}$$

which has the same form as Equation 4 evaluated at point t. It is possible to show that the coefficients α_i in Equation 6 coincide with the coefficients c_i in Equation 4 under the assumption that the scaling factor λ in $\mathbb{E} f(x_i) f(x_j) = \lambda \mathcal{K}(x_i, x_j)$ satisfies $\lambda = \sigma^2/\gamma$, where γ is the regularization parameter in Equation 1.

3. LINEAR SYSTEM IDENTIFICATION

3.1. Linear System Identification as Function Estimation

Let us consider a single-input, single-output dynamic system with input $u(t)$ and output $y(t)$, where $t \in \mathbb{R}$ (continuous-time systems) or $t \in \mathbb{Z}$ (discrete-time systems). We also assume that the system is causal, linear, and time invariant, with an impulse response denoted by g. The measurement model is

$$y_i := y(t_i) = (g \otimes u)(t_i) + e_i, \quad i = 1, \ldots, N, \tag{7.}$$

where the symbol $\otimes$ denotes convolution (in continuous or discrete time), u is the known (deterministic) input, and e_i is typically modeled as white noise. Our problem is to estimate g from the measurements y_i.

Our aim is to exploit the estimator in Equation 1. We then interpret the unknown function as the impulse response, with the linear functionals L_i defined through convolutions—e.g.,

$$L_i[f] = \int_0^{+\infty} u(t_i - s) f(s) \mathrm{d}s \tag{8.}$$

in continuous time. In view of system causality, the class of kernels must be restricted to the causal one, i.e., satisfying $\mathcal{K}(s,t) = 0$ for negative s or t. Then, according to Theorem 2, the basis functions defining our continuous-time impulse response estimate $\hat{g}$ are the causal functions

$$L_i[f](t) = \int_0^{+\infty} u(t_i - s) \mathcal{K}(s,t) \mathrm{d}s, \quad i = 1, \ldots, N. \tag{9.}$$

It is now important to select an RKHS suited for impulse response reconstruction. Below, we discuss how to incorporate information on system stability in $\mathcal{H}$.

3.2. Bounded-Input, Bounded-Output Stability in Reproducing Kernel Hilbert Spaces

A system is said to be BIBO stable if for every bounded input u the system output y is also bounded. It is well known that the necessary and sufficient condition for BIBO stability is that the impulse response g be in the space ℓ_1 of absolutely summable functions. It therefore becomes important to characterize the RKHSs contained in ℓ_1.

Definition 3 (stable reproducing kernel Hilbert space). Let $\mathcal{H}$ be the RKHS of functions (impulse responses) induced by a kernel $\mathcal{K}$. Then, $\mathcal{K}$ and $\mathcal{H}$ are said to be stable if $\mathcal{H} \subset \ell_1$ (72, 73).

The following theorem then provides the necessary and sufficient condition $\mathcal{H}$ to be stable. We state it in continuous time using ℓ_∞ to denote the space of essentially bounded functions. Its discrete-time version is obtained by simply replacing the integral with summation.

Theorem 3 (reproducing kernel Hilbert space stability). Let $\mathcal{H}$ be the RKHS induced by $\mathcal{K} : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$. It holds that (71)

$$\mathcal{H} \subset \ell_1 \iff \int_{\mathbb{R}^+} \left| \int_{\mathbb{R}^+} \mathcal{K}(x,a) f(a) \mathrm{d}a \right| \mathrm{d}x < \infty \ \ \forall f \in \ell_\infty. \qquad 10.$$

One consequence of the above result is that RKHS stability is implied by

$$\int_{\mathbb{R}_+ \times \mathbb{R}_+} |\mathcal{K}(t,\tau)| \mathrm{d}t \mathrm{d}\tau < \infty. \qquad 11.$$

Therefore, continuous and absolutely integrable kernels induce spaces of continuous and BIBO-stable impulse responses. The condition in Equation 11 also becomes necessary for nonnegative-valued kernels (72, section 13). The radial basis kernels $\mathcal{K}(t,s) = h(|s-t|)$, which include the popular Gaussian kernel

$$\mathcal{K}(t,s) = \exp\left(-\frac{(t-s)^2}{\zeta}\right), \quad t,s \geq 0, \qquad 12.$$

where ζ is the kernel width, are all unstable. Stability instead holds for the stable-spline kernel obtained in Reference 77 by applying an exponential change of coordinates to Green functions (14). In continuous time, the stable-spline kernel is defined by

$$\mathcal{K}(t,s) = e^{-\beta \max(t,s)}, \quad t,s \geq 0, \qquad 13.$$

where $\beta > 0$ is related to the impulse response decay rate. Reference 72 provides an overview of recently proposed kernels, and Reference 78 discusses the kernel properties in the frequency domain.

3.3. Bounded-Input, Bounded-Output Stability Under the Bayesian Framework

We now frame the notion of stable models in the Bayesian setting introduced in Section 2.2, which is formalized in the following definition.

Definition 4 (stable prior). A stochastic process $\{g(t)\}$ over a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ is said to be almost surely (a.s.) stable if

$$\mathbb{P}[g \in \ell_1] = 1.$$

A sufficient condition for a.s. stability is provided by the following lemma (stated only in continuous time for ease of exposition).

Lemma 1. Assume that $\{g(t)\}_{t \in \mathbb{R}^+}$ is a continuous-time process with a zero-mean and covariance function $\mathcal{K}(t,s)$, $t,s \in \mathbb{R}^+$. If

$$\int_0^\infty \mathcal{K}(t,t)^{1/2} \mathrm{d}t = M < \infty, \qquad 14.$$

then

$$\|g\|_1 := \int_0^\infty |g(t)|\, \mathrm{d}t < \infty \qquad \text{a.s.} \quad \text{and} \quad \mathbb{E}\|g\|_1 < \infty.$$

That is, not only is g stable ($g \in \ell_1$ a.s.), but also the expectation of the ℓ_1 norm is finite.

The condition in Equation 14 has been used in Lemma 1 as a sufficient condition to guarantee that realizations from the prior are a.s. finite. The following lemma provides a link with the BIBO stability of functions in the RKHS $\mathcal{H}$.

Lemma 2. Under the condition in Equation 14, the RKHS $\mathcal{H}$ with kernel $K(t,s)$ is contained in ℓ_1.

Under the common condition in Equation 14, Lemmas 2 and 1 guarantee, respectively, that (*a*) $\mathcal{H} \subset \ell_1$ ($\mathcal{H}$ is stable) and (*b*) $\mathbb{P}\left[\{g(t)\}_{t\in\mathbb{R}^+} \in \ell_1\right] = 1$ ($\{g\}$ is stable). In addition, under the condition in Equation 14, the estimated system is also a.s. BIBO stable, as formally stated in the following lemma.

Lemma 3. Under the condition in Equation 14, the minimum variance estimator $\hat{g} := \mathbb{E}[g|Y]$ is a.s. in ℓ_1.

The proofs of Lemmas 1–3 are provided in the Appendix (Section 7).

3.4. Discrete-Time Linear Systems and Finite Impulse Response Models

A simple yet significant case arises when a discrete-time linear and stable system is identified that adopts a FIR of (possibly high) dimension m. In particular, we now assume that the measurement model in Equation 7 can be rewritten in matrix form as

$$Y = \Phi\theta + e, \tag{15.}$$

where the vector $Y \in \mathbb{R}^N$ contains the noisy outputs, the components of $\theta \in \mathbb{R}^m$ are the impulse response coefficients, Φ is a known regression matrix (independent of v) built with the system inputs, and e is the noise vector.

The simplest solution is to use least squares to infer θ from Y, but this approach can suffer from high variance due to ill conditioning. We can then resort to Equation 1, using an RKHS induced by an $m \times m$ symmetric and semidefinite positive matrix Σ, which, in place of functions, thus contains m-dimensional vectors. For instance, we can use the stable-spline kernel in Equation 13 to define Σ, setting its (i, j) entry to $e^{-\beta \max(i,j)}$. If the squared loss is adopted, the estimator in Equation 1 becomes regularized least squares, and assuming Σ is invertible (as happens in the stable-spline case), one has

$$\hat{\theta} = \arg\min_{f\in\mathbb{R}^m} \|Y - \Phi f\|^2 + \gamma f^T \Sigma^{-1} f \tag{16a.}$$

$$= (\Phi^T\Phi + \gamma\Sigma^{-1})^{-1}\Phi^T Y. \tag{16b.}$$

It is worth observing that, following the Bayesian approach in Section 2.2 under the assumptions that (*a*) $\theta \sim \mathcal{N}(0, \lambda\Sigma)$, (*b*) the noise e in Equation 15 satisfies $e \sim \mathcal{N}(0, \sigma^2 I)$, and (*c*) $\gamma = \sigma^2/\lambda$, we also have that $\hat{\theta}$ in Equation 16 can be written as

$$\hat{\theta} = \mathbb{E}[\theta|Y] = \lambda\Sigma\Phi\left(\lambda\Phi\Sigma\Phi^\top + \sigma^2 I\right)^{-1} Y = (\Phi^T\Phi + \gamma\Sigma^{-1})^{-1}\Phi^T Y, \tag{17.}$$

where the last equality follows from the matrix inversion lemma.

3.5. Hyperparameter Tuning Strategies

One important problem connected with the use of Equation 1 in real applications is that the regularization parameter γ is typically unknown and must be determined from data. Furthermore, kernel parameters could also be unknown, a relevant example being β entering in Equation 13, which regulates how fast the impulse response is expected to go to zero. We describe some of the most important tuning strategies below. For the sake of simplicity, in doing so we refer mainly to the estimator in Equation 16. In addition, we use η to denote the hyperparameter vector—e.g., $\eta = [\gamma\ \beta]$ in the stable-spline case.

3.5.1. Cross-validation, predicted residual sums of squares, and generalized cross-validation. In multistage cross-validation, data are divided into complementary subsets. During each stage, some subsets are exploited to train the model, while the remaining subsets are used to evaluate predictive performances. Such a procedure can then be repeated by rotating the choice of estimation and validation data sets. This allows the definition of a cross-validation score (the average of the scores relative to each stage), which corresponds to an estimate of the predictive capability of the model and must then be optimized with respect to η.

In k-fold cross-validation, the training set is partitioned in k disjoint subsets (folds) of approximately the same size, and k estimation rounds are performed. While a typical choice would be $k = 5$ or $k = 10$, leave-one-out validation is an extreme case where $k = N$, so that at each round only one measurement falls in the validation set. The leave-one-out score with a quadratic loss is known as predicted residual sums of squares (PRESS) (79, 80). Interestingly, for linear estimators and, in particular, when Equation 16 is adopted, PRESS evaluation reduces to a single model estimation (13, theorem 4.2.1). Indeed, the estimate of η is

$$\hat{\eta} = \arg\min_{\eta} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i(\eta)}{1 - a_{ii}(\eta)} \right)^2, \qquad 18.$$

where, defining the so-called hat matrix as $A(\eta) = \Phi(\Phi^T\Phi + \gamma\Sigma^{-1})^{-1}\Phi^T$, the $\hat{y}_i$ is the ith component of the (output prediction) vector $\hat{Y} = AY$, while a_{ii} is the ith diagonal element of A.

The trace of the hat matrix A is especially useful and goes under the name of degrees of freedom. We denote it with

$$d_f(\eta) = \mathrm{tr}\,(A(\eta))\,. \qquad 19.$$

Such notation is already useful if the diagonal elements a_{ii} of A in Equation 18 are replaced by their average. This leads to generalized cross-validation (GCV) (81, 82), which determines η as

$$\hat{\eta} = \arg\min_{\eta} \frac{1}{N} \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i(\eta))^2}{(1 - d_f(\eta)/N)^2}. \qquad 20.$$

GCV enjoys many important asymptotic properties. Also, as described in, e.g., Reference 81, for a finite data set size it is a good approximation of the output mean squared error, a performance index discussed in Section 3.5.2.

3.5.2. Stein's unbiased risk estimation. Let the measurement model be $y_i = c_i + e_i$, for $i = 1, \ldots, N$, with c_i, e.g., to represent the noiseless system output at instant t_i, while the noise e_i is uncorrelated with variance σ^2. If $\mathcal{E}$ denotes the expectation with respect to the measurement noise,

an indication of the performance of the predictors $\hat{y}_i(\eta)$ is then given by

$$\mathcal{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i(\eta)-c_i)^2\right], \tag{21.}$$

a quantity related to the concept of output mean squared error. Note that the output mean squared error depends on c_i, which, in our linear setting, corresponds to $L_i[g]$. Such a quantity is therefore not accessible to direct measurement, as it is a function of the unknown impulse response. However, when the noise variance is known, η can be obtained by minimizing an unbiased estimator of Equation 21. More specifically, if the e_i noise is white with variance σ^2, this leads to the Stein's unbiased risk estimation (SURE) estimator, given by

$$\hat{\eta} = \arg\min_{\eta} \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i(\eta))^2 + \frac{2d_{\mathrm{f}}(\eta)}{N}\sigma^2. \tag{22.}$$

Interestingly, the above expression coincides with the Akaike information criterion except that the model dimension is measured by the degrees of freedom $d_{\mathrm{f}}(\eta)$ in Equation 19 parameterized by η. One also has that $0 \le d_{\mathrm{f}}(\eta) \le N$, and if $H(\eta)$ is full rank, then $d_{\mathrm{f}}(\eta)$ goes from N to 0 as the regularization parameter γ varies from 0 to $+\infty$. Therefore, $d_{\mathrm{f}}(\eta)$ is a real number that describes the flexibility of the model, allowing one to change model complexity with continuity by tuning the regularization parameter.

Finally, note that, when $d_{\mathrm{f}} \ll N$, we have

$$\frac{1}{(1-d_{\mathrm{f}}/N)^2} \approx 1 + 2d_{\mathrm{f}}/N,$$

and therefore GCV tends to a version of SURE, where the variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2.$$

3.5.3. Marginal likelihood. In the regularized estimator in Equation 1, the unknown impulse response θ (or the function g in the more general setting) is modeled as a deterministic vector. As discussed in Section 2.2 for the case of quadratic losses, the same estimator can be given a Bayesian interpretation, which is also useful for hyperparameter tuning. We refer readers to Reference 83 for the theoretical details of RKHSs, but the correspondence with Equation 16 can be easily described as follows. Let us assume that θ and the noise e are independent zero-mean normal vectors, as specified as follows:

$$\theta \sim \mathcal{N}(0, \lambda\Sigma), \quad e \sim \mathcal{N}(0, \sigma^2 I_N). \tag{23.}$$

Applying the Bayes rule, one obtains that the posterior distribution of θ given Y is

$$\theta|Y \sim \mathcal{N}\left(\hat{\theta}, \left(\frac{\Phi^T\Phi}{\sigma^2} + \frac{\Sigma^{-1}}{\lambda^2}\right)^{-1}\right), \tag{24.}$$

with the minimum variance estimate $\hat{\theta}$ indeed given by Equation 17 if $\gamma = \sigma^2/\lambda$.

Exploiting such a connection, one way to estimate η is to optimize the so-called marginal likelihood $\mathbf{p}(Y|\eta)$, i.e., the joint density $\mathbf{p}(Y|\theta,\eta)\mathbf{p}(\theta|\eta)$ where the dependence on θ is integrated out.

Letting $\Sigma_y(\eta) = \lambda \Phi \Sigma \Phi^T + \sigma^2 I_N$, one has

$$\hat{\eta} = \arg\min_{\eta} \; Y^T \Sigma_y(\eta)^{-1} Y + \log\det(\Sigma_y(\eta)). \qquad 25.$$

This tuning method relies on the concept of Bayesian evidence and includes the Occam's razor principle, automatically penalizing unnecessarily complex models. References 84–86 discuss connections between marginal likelihood and degrees of freedom $d_f(\eta)$, and Reference 87 discusses links with mean squared error minimization via the concept of excess degrees of freedom.

3.6. Numerical Example

To illustrate the use of the estimator in Equation 16, let us consider the reconstruction of the impulse response whose z-transform is the rational transfer function

$$G(z) = \frac{(z+1)^2}{z(z-0.8)(z-0.6)}.$$

The function is virtually equal to zero after 50 samples. The identification data consist of 1,000 input–output pairs and are shown in **Figure 1**. In particular, at $t = 0$, the system was fed white noise low-pass filtered by $z/(z-0.99)$, while the measurement noise is white and Gaussian with variance equal to that of the noiseless output divided by 20.

The adoption of least squares in conjunction with a FIR of length $m = 50$ would lead in this case to a large reconstruction error. The reason is that the problem is ill conditioned (high condition

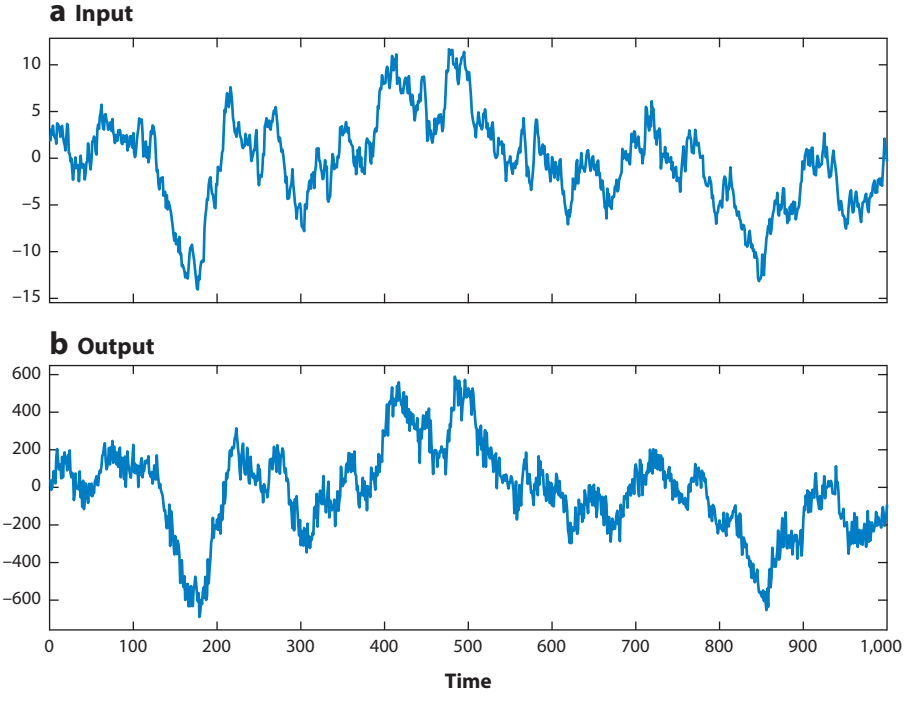

Figure 1

Identification data used for the numerical example in Section 3.6, consisting of 1,000 input–output pairs.

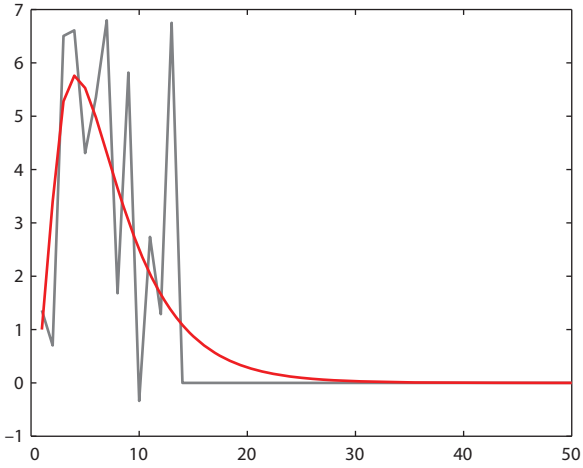

Figure 2

True impulse response (*red line*) and least squares estimate with optimal finite impulse response order m chosen by the oracle (*gray line*).

number of Φ) due to poor excitation of the input, so that a (relatively) high-order FIR suffers from high variance. It is essential to control model complexity to reduce the variance component, which introduces some bias (i.e., decreasing model flexibility). In the classical paradigm, model complexity can be controlled by selecting a discrete model order, such as the length m of the FIR. Let us adopt an oracle that knows the true θ and minimizes $\|\theta - \hat{\theta}(m)\|$ over $m \in [1, \ldots, 50]$. One obtains $m = 18$, and the corresponding impulse response estimate is shown in **Figure 2**. Even if this tuning procedure is ideal, the estimate is far from satisfactory.

An alternative is to resort to the estimator in Equation 16 equipped with the stable-spline kernel in Equation 13. In contrast to the previous approach, the parameter θ dimension does not vary, instead being fixed to $m = 50$. Model complexity is now tuned in a continuous manner by varying $\eta = [\lambda\ \beta]$ (noise variance is assumed known). To estimate the hyperparameter vector, the marginal likelihood optimization in Equation 25 is adopted. **Figure 3** displays three reconstructions

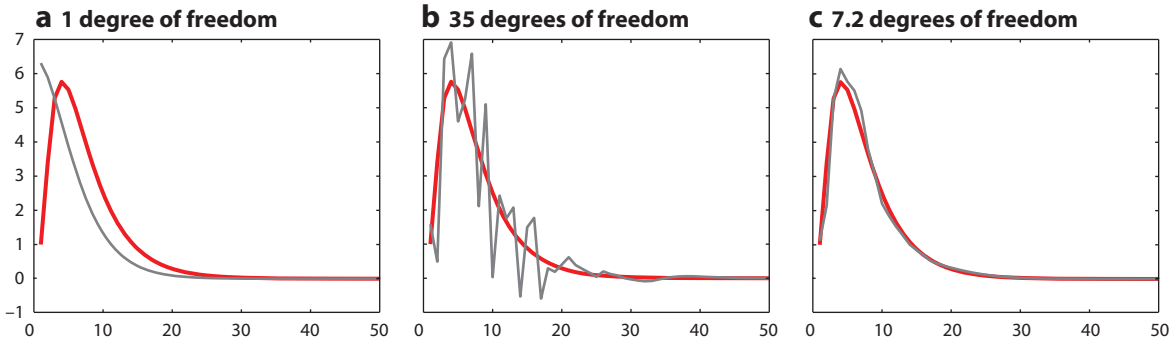

Figure 3

True impulse responses (*red lines*) and stable-spline estimates (*gray lines*) for (*a*) 1 degree of freedom, (*b*) 35 degrees of freedom, and (*c*) 7.2 degrees of freedom. Panel *c* uses marginal likelihood optimization.

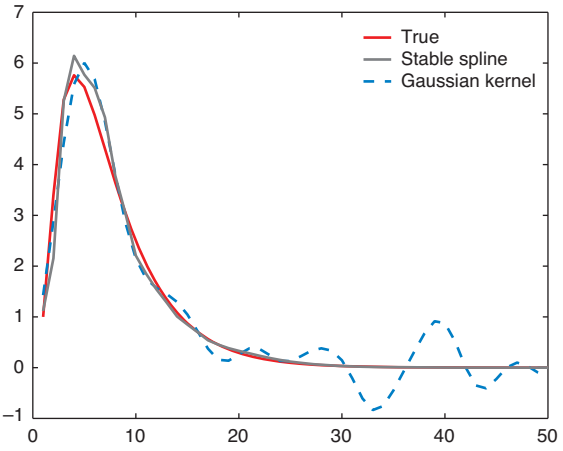

Figure 4

True impulse response (*solid red line*), stable spline (*solid gray line*), and Gaussian kernel estimate (*dashed blue line*), with hyperparameters tuned via marginal likelihood optimization.

obtained by setting the decay rate β to its marginal likelihood estimate. In the first case (**Figure 3*a***), λ is manually set in order to obtain a value of the degrees of freedom $d_f = 1$. This leads to oversmoothing, i.e., a too simple model unable to describe the experimental data. In the second case (**Figure 3*b***), a value of $d_f = 35$ is imposed. The resulting estimate is ill conditioned since the model is too complex. The third case (**Figure 3*c***) exploits the maximum likelihood estimate of both β and λ. This corresponds to $d_f = 7.2$, a value leading to an impulse response estimate close to the truth.

Finally, **Figure 4** complements **Figure 3*c*** by reporting the estimate obtained using the Gaussian kernel in Equation 12. The regularization parameter γ and the kernel width ζ are estimated via marginal likelihood optimization. The stable spline outperforms the Gaussian kernel, which represents the classical choice in machine learning. Notice that, in the FIR setting, any kernel is stable since it is obtained by truncation. However, if $\mathcal{K}$ is derived by truncating an unstable kernel, the model does not include information on the fact that the impulse response decays to zero as time progresses. This explains the undue oscillations now affecting the estimate.

4. NONLINEAR SYSTEM IDENTIFICATION

Many physical systems are too complex to be captured by linear relationships, requiring the development of nonlinear system identification techniques. In this section, we describe some nonparametric approaches that are still built around the estimator in Equation 1. We consider two situations. In the first, we assume that our prior information is limited to smoothness and some form of stability on the input–output map. In the second, we review some approaches that exploit more structured information on the problem, such as Wiener/Hammerstein and hybrid systems.

4.1. Regularized Nonlinear Black-Box System Identification in Reproducing Kernel Hilbert Spaces

The connection between nonlinear system identification and the estimator in Equation 1 can be obtained by suitably defining the input space, i.e., the domain of the unknown map. First,

let each linear operator L_i be an evaluation functional associated with an input location x_i—i.e., $L_i[g] = g(x_i)$. From Theorem 2, one then has that the basis functions defining $\hat{g}$ become the kernel sections centered on x_i:

$$L_i[f](x) = K_{x_i}(x), \quad i = 1, \ldots, N. \tag{26.}$$

Second, in the data (x_i, y_i), we can now think of y_i as the output at instant t_i, while x_i encodes past input and output values. For the sake of simplicity, only the discrete-time setting is considered, and no autoregressive part is included in the model. Hence, when using the so-called nonlinear FIR models, one has

$$x_i = [u_{t_i} \;\; u_{t_i-1} \;\; \ldots \;\; u_{t_i-m+1}]^T, \tag{27.}$$

where m is the system memory. The case

$$x_t = [u_t \;\; u_{t-1} \;\; u_{t-2} \;\; \ldots]^T, \tag{28.}$$

where any input location is a sequence (an infinite-dimensional column vector), accounts for infinite-memory systems.

As in the linear case, the quadratic loss $\mathcal{V}_i(y_i, L_i[f]) = (y_i - L_i[f])^2$ is considered. We can now focus on the choice of an RKHS that includes some possible information on the nonlinear system. To simplify the notation, in what follows Equation 27 is considered using x or a to denote a generic input location.

Smoothness information on the input–output map g is often available and can be easily encoded using Mercer (continuous) kernels. As mentioned above, in machine learning, radial basis kernels $K(x, a) = h(\|x - a\|)$ are widely adopted; an important example is the Gaussian kernel:

$$K(x, a) = \exp\left(-\frac{\|x - a\|^2}{\zeta}\right), \quad \eta > 0. \tag{29.}$$

Even if successfully adopted in many real applications, some drawbacks affect Equation 29. Beyond the choice of the regularization parameter γ and kernel width ζ, the estimation of the input space dimension m is an issue. From a computational point of view, its discrete nature precludes the use of gradient methods for tuning. From a modeling point of view, the fact that the diagonal element $K(x, x)$ is constant also implies that radial basis kernels do not embed information that output energy is likely to augment if input energy increases. Finally, no information is given about the fact that, in dynamic systems, inputs $u_{i-\tau}$ are expected to have less influence on y_i as the positive lag τ increases. As is done in the linear case, one would rather replace model order m with a hyperparameter connected (in some sense) with system exponential stability. For some recent insights on the concept of a stable RKHS in the nonlinear case (and related consistency results tailored for dynamic systems), see Reference 74.

To face the above issues, the stable-spline kernel in Equation 13 can also be useful in this nonlinear context (56, 74). It can be exploited to measure the distance between different input trajectories, leading to the class of nonlinear stable-spline (NSS) kernels. To describe them, let us again use Σ to denote the $m \times m$ stable-spline kernel matrix with (i, j) entry $\Sigma(i, j) = e^{-\beta \max(i,j)}$. As is clear in what follows, the choice of m is now much less critical since it now only needs to represent an upper bound on the (expected) system memory size. The parameter β will then be in charge of establishing the effective input space dimension (and will be determined from data).

As a first example, we can modify Equation 29 as follows:

$$\mathcal{K}(x,a) = \exp\left(-\frac{(x-a)^T \Sigma (x-a)}{\eta}\right) \quad (\mathrm{NSS}_1). \tag{30.}$$

Thus, note that the presence of Σ inside the exponential is to give the information that past inputs' influence decays exponentially to zero. A refinement is achieved that specifies that more output variability is expected if input energy increases. For this purpose, one can use the kernel

$$\mathcal{K}(x,a) = \left(x^T \Sigma a\right) \times \exp\left(-\frac{(x-a)^T \Sigma (x-a)}{\eta}\right) \quad (\mathrm{NSS}_2), \tag{31.}$$

which is no longer constant over the diagonal. Another example is obtained using sums of kernels,

$$\mathcal{K}(x,a) = \left(x^T \Sigma a\right) + \exp\left(-\frac{(x-a)^T \Sigma (x-a)}{\eta}\right) \quad (\mathrm{NSS}_3) \tag{32.}$$

or

$$\mathcal{K}(x,a) = \left(x^T \Sigma a\right) + \left(x^T \Sigma a\right) \times \exp\left(-\frac{(x-a)^T \Sigma (x-a)}{\eta}\right) \quad (\mathrm{NSS}_4), \tag{33.}$$

which model the system as the sum of a linear and nonlinear component.

4.2. Hyperparameter Tuning Strategies

We now briefly discuss estimation of the hyperparameter vector η, pointing out some natural connections with the techniques illustrated in the linear scenario in Section 3.5.

When a quadratic loss is adopted, even in the nonlinear setting, the estimator in Equation 1 leads to linear predictors. In fact, let $\mathbf{K}$ be the $N \times N$ matrix with (i, j) entry $\mathcal{K}(x_i, x_j)$. A regularization network then has the structure

$$\hat{g}(x) = \sum_{i=1}^{N} \hat{c}_i K_{x_i}(x),$$

where the weights vector $\hat{c} = [\hat{c}_1, \ldots, \hat{c}_N]^T$ is given by $\hat{c} = (\mathbf{K}(\eta) + \gamma I_N)^{-1} Y$. Hence, techniques like PRESS, GCV, and SURE, as reported in Equations 18, 20, and 22, respectively, can be immediately applied: It is sufficient to set the predictions $\hat{y}_i$ to the components of the vector $\mathbf{K}\hat{c}$. The marginal likelihood is also easy to obtain. In fact, we can now think of the unknown map as a zero-mean Gaussian random field with covariance $\mathcal{K}$, independent of the white Gaussian noise of variance σ^2 (24). Then, Y is a zero-mean Gaussian vector with covariance $Z(\eta) = \lambda \mathbf{K}(\eta) + \sigma^2 I_N$, and the marginal likelihood estimate of η is

$$\hat{\eta} = \arg\min_{\eta} \; Y^T Z(\eta)^{-1} Y + \log\det(Z(\eta)). \tag{34.}$$

When losses other than quadratic are adopted, one can still resort to the cross-validation techniques described at the beginning of Section 3.5, but the use of the marginal likelihood is much more difficult. Since the marginal likelihood is no longer available in closed form, the options are to use the expectation–maximization algorithm or stochastic simulation techniques like Markov chain Monte Carlo (see, e.g., 88–91). The nonlinearity of the predictor in Equation 1 also makes it more difficult to compute the other scores. However, for some popular losses, such as ℓ_1 and Vapnik's penalty, the degrees of freedom can still be computed, allowing, e.g., the use of SURE (for details, see 92–94).

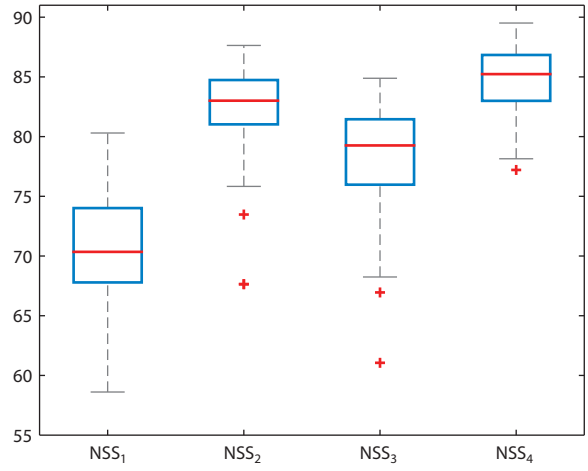

Figure 5

Box plots of the 100 test set fits achieved by the nonlinear stable-spline (NSS) estimators equipped with the kernels in Equations 30–33.

4.3. Numerical Experiment Using Nonlinear Stable-Spline Kernels

We test the new kernels in Equations 30–33 using the following nonlinear system from Reference 56:

$$y_i = u_i + 0.6u_{i-1} + 0.35\,(u_{i-2} + u_{i-4}) + 0.2(u(i-5) + u(i-6)) - 0.25u_{i-3}^2 + 0.9u_{i-3} + 0.25u_iu_{i-1} + 0.75u_{i-2}^3 - u_{i-1}u_{i-2} + 0.5\left(u_i^2 + u_iu_{i-2} + u_{i-1}u_{i-3}\right) + e_i,$$

where u_i and e_i represent independent white Gaussian noise of unit variance. The identification data are given by 2,000 input–output pairs. The estimator in Equation 1 equipped with the kernels in Equations 30–33 is used to identify the system, adopting the quadratic loss and setting the input space dimension to $m = 100$. Marginal likelihood optimization is used to tune the regularization parameter γ and the kernel hyperparameters (β, η). The performance is measured by the percentage fit on a test set of size 2,000.

Figure 5 displays the box plots of the 100 fits obtained by the four stable-spline-based estimators after a Monte Carlo study of 100 runs. The obtained results suggest that the stable-spline metric can be useful for nonlinear system identification. Comparison of the fits achieved by, e.g., NSS_1 and NSS_2 also shows that the use of kernel products can significantly improve performance.

5. SPARSITY IN SYSTEM IDENTIFICATION: VARIABLE AND STRUCTURE SELECTION

The use of regularization in inverse problems to find sparse solutions has received significant attention in diverse areas of machine learning and signal processing, including dictionary learning and matrix factorization problems (95), compressive sensing (96), and selective shrinkage (97).

In system identification, many problems can be framed in the context of sparse estimation, but we believe it is fair to say that they belong to one of two categories, variable selection or structure selection, which are discussed in the following sections. Many machine learning methods have been developed for sparse estimation, which can be categorized mainly as those using ℓ_1 (or more

generally ℓ_p, $p \leq 1$) norms as penalty functions, such as the lasso (least absolute shrinkage and selection operator) and its variations (97–100), and those based on a probabilistic framework in which hierarchical prior models are developed to encode the presence or absence of one variable or group. The variations are the so-called automatic relevance determination (101, 102), sparse Bayesian learning (103), spike and slab priors (104), and reweighed schemes such as reweighed ℓ_1/ℓ_2 methods (105).

As shown below, the structure of dynamical systems can be used to inform the construction of suitable kernels or priors, again showing how the tight interplay between dynamical systems theory and machine learning plays a crucial role.

5.1. Variable Selection

When modeling multi-input, multi-output dynamical systems, it is of interest to detect whether one specific variable (considered as a time series) influences another variable. This dependency structure is often encoded using a graph where nodes are variables and (directed) edges encode the (causal) conditional dependencies among variables, leading to so-called dynamic network models (106–108). In the linear framework, these networks encode Granger causality (109) relations expressed in terms of conditional correlation, of which nonlinear extensions are possible using the notions of conditional independence and conditional mutual information. The main observation is that the presence or absence of one variable in a model can be framed as a group-sparse problem. For the sake of exposition, let us consider the multi-input, single-output linear case with m inputs:

$$y_i := y(t_i) = \sum_{j=1}^{m} (g_j \otimes u_j)(t_i) + e_i, \quad i = 1, \ldots, N. \tag{35.}$$

In this linear case, the impulse response $\{g_j(t)\}$ from input u_j to the output y can be regarded as a group, as was done in Reference 107, and a group-lasso algorithm to estimate the g_i's can be formulated as

$$\hat{g}_1, .., \hat{g}_m = \arg \min_{g_1, .., g_m \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{m} (g_j \otimes u_j)(t_i) \right)^2 + \gamma \sum_{j=1}^{m} \|g_i\|_{\mathcal{H}}. \tag{36.}$$

One well-known drawback of the lasso and group-lasso methods is that sparsity in the solution (i.e., some of the $\hat{g}_i$'s are exactly zero), which can be regulated by properly tuning the regularization parameter γ, is obtained at a price of significant shrinkage (i.e., bias towards zero) of the estimated nonzero impulse responses. An alternative, which has a much better shrinkage-versus-sparsity trade-off (28, 107), is found via the so-called sparse Bayesian learning formulation, which is the solution to

$$\hat{g}_1, .., \hat{g}_m = \arg \min_{g_1, .., g_m \in \mathcal{H}} \frac{1}{N\sigma^2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{m} (g_j \otimes u_j)(t_i) \right)^2 + \sum_{j=1}^{m} \lambda_i^{-1} \|g_i\|_{\mathcal{H}}^2, \tag{37.}$$

where the coefficients λ_i are tuned via marginal likelihood optimization as described in Section 3.5.3. Indeed, it can be proved that the optimal marginal likelihood solution for $\lambda_1, \ldots, \lambda_m$ yields, for nonzero probability sets in the data space, to some of the λ_i's exactly equal to zero, which in turn implies that $\hat{g}_i = 0$.

In general nonlinear models, which in the multi-input, single-output case can be written as

$$y_i := y(t_i) = f(x_{1,t}, x_{2,t}, .., x_{m,t}) + e_i, \quad i = 1, \ldots, N, \tag{38.}$$

where $x_{i,t}$ is the input location associated with the ith input u_i at time t as defined in Equation 28, variable selection amounts to detecting which variables the function $f(x_{1,t}, x_{2,t}, \ldots, x_{m,t})$ depends on. Considering the kernels in Equations 30–33, this can be done by imposing a certain block diagonal structure in the kernel matrix Σ and estimating (e.g., via marginal likelihood) suitable scaling factors that, when set to zero, imply that the estimated function will not depend on the corresponding input. Owing to space limitations, we do not pursue this avenue further here; for a more in-depth treatment, see References 56, 110, and 111.

5.2. Structure Selection

Prior knowledge concerning the specific structure that the dynamical system under investigation belongs to can often be incorporated into the structure of the underlying RKHS (or prior distribution), which contains candidate models. Estimation often entails selecting among a discrete family of alternatives, which arises, for instance, for finite-order linear systems (112, 113), linear dynamical systems with low-dimensional latent components (114, 115), hybrid system identification (116), and nonlinear systems where, for instance, the presence of interaction between specific variables may or may not be present (see, e.g., 56) or block structures are considered (Wiener, Hammerstein, or combinations thereof) (117–121).

Sparse methods may come in handy in these cases, as they avoid the need to perform multiple tests (possibly a combinatorial number), which is computationally unfeasible and may result in poor performance when the number of alternatives is very large—a well-known problem with multiple testing.

The simplest situation is when the system under analysis is linear time invariant and possibly of low order (McMillan degree), or when for control design purposes it is of interest to obtain a model that has a low McMillan degree. It is well known in system theory (122) that a linear system with impulse response $g(t)$ has a McMillan degree equal to n if and only if all (suitably large) finite Hankel matrices have a rank equal to n. This fact was used in References 112 and 123 to frame a finite-order system approximation as a constraint optimization problem, which in turn can be formulated as a penalized regression problem of the form

$$\hat{g} = \arg\min_g \frac{1}{N} \sum_{i=1}^{N} \left((y_i - g \otimes u)(t_i) \right)^2 + \gamma \, \|H(g)\|_*, \qquad 39.$$

where

$$\|\mathcal{H}(g)\|_* := \operatorname{trace} \sqrt{H(g) H^\top(g)}$$

denotes the nuclear norm of the Hankel matrix (of suitable size) H formed with the impulse response g. However, as thoroughly discussed in Reference 49, the optimization problem in Equation 39 can be interpreted as a maximum a posteriori estimator with a prior on g that does not include system stability (see 49, figure 1). In Reference 113, using maximum entropy arguments, a prior for the impulse response g is built that includes stability and favors the estimated impulse response $\hat{g}$ being close to a low order (in terms of McMillan degree).

6. CONCLUSIONS

We have provided a bird's-eye view on the role of machine learning tools in system identification. Functional analysis (RKHSs) and Bayesian statistics provide convenient frameworks under which dynamical model classes can be described, embedding structural properties such as stability, smoothness, and complexity. The Bayesian framework provides robust tools to perform model

selection, adapting model complexity in a continuous manner using a few hyperparameters but also exploiting sparsity-promoting priors to perform structure selection. As we have argued, these tools can be useful in both linear and nonlinear scenarios. While linear scenarios have been extensively studied in the past 10 years or so, we believe it is fair to say that the study of nonlinear scenarios is still in its infancy, and much remains to be done.

7. APPENDIX

7.1. Proof of Lemma 1

Proof. The probability that the ℓ_1 norm $\int_0^\infty |g(t)|\,\mathrm{d}t$ exceeds a threshold T_{ℓ_1} can be bounded using a Markov inequality as follows:[1]

$$\mathbb{P}\left[\int_0^\infty |g(t)|\,\mathrm{d}t \geq T\right] \leq \frac{1}{T^2}\mathbb{E}\left(\int_0^\infty |g(t)|\,\mathrm{d}t\right)^2 = \frac{1}{T^2}\int_0^\infty\int_0^\infty \mathbb{E}|g(t)||g(\tau)|\,\mathrm{d}t\mathrm{d}\tau$$
$$\leq \frac{1}{T^2}\int_0^\infty\int_0^\infty \mathcal{K}(t,t)^{1/2}\mathcal{K}(\tau,\tau)^{1/2}\,\mathrm{d}t\mathrm{d}\tau = \frac{M^2}{T^2}.$$

Using the assumption in Equation 14, we have that $\mathbb{P}\left[\int_0^\infty |g(t)|\,\mathrm{d}t \geq T\right] \leq \frac{M^2}{T^2}$, and therefore $\mathbb{P}\left[\int_0^\infty |g(t)|\,\mathrm{d}t < T\right] \geq 1 - \frac{M^2}{T^2}$. Taking the limit as $T \to +\infty$, we have $\mathbb{P}\left[\int_0^\infty |g(t)|\,\mathrm{d}t < +\infty\right] = 1$, which concludes the first part of the proof.

Concerning the finiteness of the expected one norm, let us define the nonnegative random variable $z := \|g\|_1$ and denote with $F_z(a)$ its distribution—i.e., $F_z(a) := \mathbb{P}[z \leq a]$. Recall that $\mathbb{E}z = \int_0^\infty [1 - F_z(a)]\,\mathrm{d}a$. Using the fact that $1 - F_z(T) = \mathbb{P}[\|g\|_1 > T] \leq \frac{M^2}{T^2}$, we have $\mathbb{E}z = \int_0^1 [1 - F_z(a)]\,\mathrm{d}a + \int_1^\infty [1 - F_z(a)]\,\mathrm{d}a \leq \int_0^1 [1 - F_z(a)]\,\mathrm{d}a + \int_1^\infty \frac{M^2}{a^2}\,\mathrm{d}a < +\infty$, which concludes the proof. □

7.2. Proof of Lemma 2

Proof. According to Equation 11, it is sufficient to prove that $\int_0^\infty\int_0^\infty |K(t,s)|\,\mathrm{d}t\mathrm{d}s < \infty$. Using the Cauchy–Schwarz inequality $\frac{|K(t,s)|}{\sqrt{K(t,t)}\sqrt{K(s,s)}} \leq 1$, we have that

$$\int_0^\infty\int_0^\infty |K(t,s)|\,\mathrm{d}t\mathrm{d}s \leq \int_0^\infty\int_0^\infty \sqrt{K(t,t)}\sqrt{K(s,s)}\,\mathrm{d}t\mathrm{d}s = M^2 < \infty,$$

where the last inequality follows from the condition in Equation 14, thus completing the proof. □

7.3. Proof of Lemma 3

Proof. The proof goes by contradiction: If $\mathbb{P}[\hat{g} \notin \ell_1] = p$ were strictly positive (i.e., $p > 0$), it would follow that $\mathbb{E}\|\hat{g}\|_1 = \infty$. However, the following chain of inequalities holds:

$$\mathbb{E}\|\hat{g}\|_1 = \mathbb{E}\int_0^\infty |\mathbb{E}[g(t)|Y]|\,\mathrm{d}t$$
$$\leq \mathbb{E}\int_0^\infty \mathbb{E}[|g(t)||Y]\,\mathrm{d}t = \mathbb{E}\left[\mathbb{E}\left[\int_0^\infty |g(t)|\,\mathrm{d}t \mid Y\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\|g\|_1 \mid Y\right]\right] = \mathbb{E}\|g\|_1;$$

[1]The expectation and summation can be interchanged for nonnegative integrands thanks to Tonelli's theorem (see 124).

i.e., $\mathbb{E}\|\hat{g}\|_1 \leq \mathbb{E}\|g\|_1 < \infty$, where the last inequality follows from Lemma 2. Therefore, $\mathbb{E}\|\hat{g}\|_1 < \infty$, implying that $\mathbb{P}[\hat{g} \notin \ell_1] = 0$ must hold. □

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partially supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) under the project "Learning meets time: a computational approach to learning in dynamical systems" (RBFR12M3AC) and the project BIRD162411 funded by the University of Padova.

LITERATURE CITED

1. Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19:716–23
2. Schwarz G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–64
3. Arlot S, Celisse A. 2014. A survey of cross-validation procedures for model selection. *Statist. Surv.* 4:40–79
4. Hastie TJ, Tibshirani RJ, Friedman J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer
5. Bishops C. 1996. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford Univ. Press
6. Haykin S. 2009. *Neural Networks and Learning Machines*. Upper Saddle River, NJ: Pearson Educ.
7. Lecun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44
8. Tikhonov A, Arsenin V. 1977. *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley
9. Bertero M. 1989. Linear inverse and ill-posed problems. *Adv. Electron. Electron Phys.* 75:1–120
10. Aronszajn N. 1950. Theory of reproducing kernels. *Trans. Am. Math. Soc.* 68:337–404
11. Bergman S. 1950. *The Kernel Function and Conformal Mapping*. Providence, RI: Am. Math. Soc.
12. Bertero M, Poggio T, Torre V. 1988. Ill-posed problems in early vision. *Proc. IEEE* 76:869–89
13. Wahba G. 1990. *Spline Models for Observational Data*. Philadelphia: Soc. Ind. Appl. Math.
14. Poggio T, Girosi F. 1990. Networks for approximation and learning. *Proc. IEEE* 78:1481–97
15. Girosi F. 1997. *An equivalence between sparse approximation and support vector machines*. AI Memo 1606, CBCL Pap. 147, Mass. Inst. Technol., Cambridge, MA
16. Kimeldorf G, Wahba G. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* 41:495–502
17. Lukic M, Beder J. 2001. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Am. Math. Soc.* 353:3945–69
18. Bell B, Pillonetto G. 2004. Estimating parameters and stochastic functions of one variable using nonlinear measurement models. *Inverse Probl.* 20:627
19. Aravkin A, Bell B, Burke J, Pillonetto G. 2015. The connection between Bayesian estimation of a Gaussian random field and RKHS. *IEEE Trans. Neural Netw. Learn. Syst.* 26:1518–24
20. Evgeniou T, Pontil M, Poggio T. 2000. Regularization networks and support vector machines. *Adv. Comput. Math.* 13:1–50
21. Schölkopf B, Smola AJ. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press
22. Drucker H, Burges C, Kaufman L, Smola A, Vapnik V. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, ed. MC Mozer, MI Jordan, T Petsche, pp. 155–61. Cambridge, MA: MIT Press

23. Vapnik V. 1998. *Statistical Learning Theory*. New York: Wiley
24. Rasmussen C, Williams C. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press
25. Collobert R, Bengio S. 2001. SVMTorch: support vector machines for large-scale regression problems. *J. Mach. Learn. Res.* 1:143–60
26. Maritz JS, Lwin T. 1989. *Empirical Bayes Methods*. London: Chapman and Hall. 2nd ed.
27. Aravkin A, Burke J, Chiuso A, Pillonetto G. 2012. On the estimation of hyperparameters for empirical Bayes estimators: maximum marginal likelihood versus minimum MSE. *IFAC Proc. Vol.* 45(16):125–30
28. Aravkin A, Burke J, Chiuso A, Pillonetto G. 2014. Convex versus nonconvex estimators for regression and sparse estimation: the mean squared error properties of ARD and GLasso. *J. Mach. Learn. Res.* 15:217–52
29. Wahba G. 1977. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* 14:651–67
30. Smale S, Zhou D. 2007. Learning theory estimates via integral operators and their approximations. *Constr. Approx.* 26:153–72
31. Yuan M, Cai TT. 2010. A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Stat.* 38:3412–44
32. Wu Q, Ying Y, Zhou D. 2006. Learning rates of least-square regularized regression. *Found. Comput. Math.* 6:171–92
33. Mukherjee S, Niyogi P, Poggio T, Rifkin R. 2006. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.* 25:161–93
34. Poggio T, Rifkin R, Mukherjee S, Niyogi P. 2004. General conditions for predictivity in learning theory. *Nature* 428:419–22
35. Alon N, Ben-David S, Cesa-Bianchi N, Haussler D. 1997. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM* 44:615–31
36. Evgeniou T, Pontil M. 1999. On the V_γ dimension for regression in reproducing kernel Hilbert spaces. In *Algorithmic Learning Theory: 10th International Conference, ALT '99, Tokyo, Japan, December 1999*, ed. O Watanabe, T Yokomori, pp. 106–17. Berlin: Springer
37. Bousquet O, Elisseeff A. 2002. Stability and generalization. *J. Mach. Learn. Res.* 2:499–526
38. Schiller R. 1979. A distributed lag estimator derived from smoothness priors. *Econ. Lett.* 2:219–23
39. Akaike H. 1979. *Smoothness priors and the distributed lag estimator*. Tech. Rep. 40, Dep. Stat., Stanford Univ., Stanford, CA
40. Kitagawa G, Gersch W. 1996. *Smoothness Priors Analysis of Time Series*. New York: Springer
41. Chiuso A. 2016. Regularization and Bayesian learning in dynamical systems: past, present and future. *Annu. Rev. Control* 41:24–38
42. Goodwin G, Gevers M, Ninness B. 1992. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. Autom. Control* 37:913–28
43. Ljung L, Goodwin G, Agüero JC. 2014. Stochastic embedding revisited: a modern interpretation. In *53rd IEEE Conference on Decision and Control*. pp. 3340–45. New York: IEEE
44. Chandrasekaran V, Recht B, Parrilo P, Willsky A. 2012. The convex geometry of linear inverse problems. *Found. Comput. Math.* 12:805–49
45. Liu Z, Vandenberghe L. 2009. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* 31:1235–56
46. Grossmann C, Jones C, Morari M. 2009. System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Proceedings of the 48th IEEE Conference on Decision and Control*, pp. 4692–97. New York: IEEE
47. Mohan K, Fazel M. 2010. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, pp. 2953–59. New York: IEEE
48. Rojas C, Toth R, Hjalmarsson H. 2014. Sparse estimation of polynomial and rational dynamical models. *IEEE Trans. Autom. Control* 59:2962–77
49. Pillonetto G, Chen T, Chiuso A, De Nicolao G, Ljung L. 2016. Regularized linear system identification using atomic, nuclear and kernel-based norms: the role of the stability constraint. *Automatica* 69:137–49

50. Franz M, Schölkopf B. 2006. A unifying view of Wiener and Volterra theory and polynomial kernel regression. *Neural Comput.* 18:3097–118
51. Lin T, Horne B, Tino P, Giles C. 1996. Learning long-term dependencies in NARX recurrent neural networks. *IEEE Trans. Neural Netw.* 7:1329–38
52. Shun-Feng S, Yang F. 2002. On the dynamical modeling with neural fuzzy networks. *IEEE Trans. Neural Netw.* 13:1548–53
53. Fan J, Gijbels I. 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall
54. Billings S, Hua-Liang W. 2005. A new class of wavelet networks for nonlinear system identification. *IEEE Trans. Neural Netw.* 16:862–74
55. Leithead WE, Solak E, Leith DJ. 2003. Direct identification of nonlinear structure using Gaussian process prior models. In *2003 European Control Conference*, pp. 2565–70. New York: IEEE
56. Pillonetto G, Chiuso A, Quang MH. 2011. A new kernel-based approach for nonlinear system identification. *IEEE Trans. Autom. Control* 56:2825–40
57. Zhao W, Chen H, Bai E, Li K. 2015. Kernel-based local order estimation of nonlinear nonparametric systems. *Automatica* 51:243–54
58. Roll J, Nazin A, Ljung L. 2005. Nonlinear system identification via direct weight optimization. *Automatica* 41:475–90
59. Bai E, Liu Y. 2007. Recursive direct weight optimization in nonlinear system identification: a minimal probability approach. *IEEE Trans. Autom. Control* 52:1218–31
60. Bai EW. 2010. Non-parametric nonlinear system identification: an asymptotic minimum mean squared error estimator. *IEEE Trans. Autom. Control* 55:1615–26
61. Suykens J, Gestel TV, Brabanter JD, Moor BD, Vandewalle J. 2002. *Least Squares Support Vector Machines*. Singapore: World Sci.
62. Suykens J, Alzate C, Pelckmans K. 2010. Primal and dual model representations in kernel-based learning. *Stat. Surv.* 4:148–83
63. Frigola R, Lindsten F, Schön T, Rasmussen C. 2013. Bayesian inference and learning in Gaussian process state-space models with particle MCMC. In *Advances in Neural Information Processing Systems 26*, ed. CJC Burges, L Bottou, M Welling, Z Ghahramani, KQ Weinberger, pp. 3156–64. Red Hook, NY: Curran
64. Frigola R, Rasmussen C. 2013. Integrated preprocessing for Bayesian nonlinear system identification with Gaussian processes. In *52nd Annual Conference on Decision and Control*, pp. 5371–76. New York: IEEE
65. Ljung L. 1999. *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice Hall. 2nd ed.
66. Pruyt E, Cunningham S, Kwakkel J, de Bruijn J. 2014. From data-poor to data-rich: system dynamics in the era of big data. In *Proceedings of the 2014 International Conference of the System Dynamics Society*, pap. 1390. Albany, NY: Syst. Dyn. Soc.
67. Pillonetto G, De Nicolao G. 2010. A new kernel-based approach for linear system identification. *Automatica* 46:81–93
68. Pillonetto G, Chiuso A, De Nicolao G. 2011. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica* 47:291–305
69. Chen T, Ohlsson H, Ljung L. 2012. On the estimation of transfer functions, regularizations and Gaussian processes—revisited. *Automatica* 48:1525–35
70. Vishwanathan SVN, Smola AJ, Vidal R. 2007. Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *Int. J. Comput. Vis.* 73:95–119
71. Carmeli C, Vito ED, Toigo A. 2006. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl.* 4:377–408
72. Pillonetto G, Dinuzzo F, Chen T, Nicolao GD, Ljung L. 2014. Kernel methods in system identification, machine learning and function estimation: a survey. *Automatica* 50:657–82
73. Dinuzzo F. 2015. Kernels for linear time invariant system identification. *SIAM J. Control Optim.* 53:3299–317
74. Pillonetto G. 2018. System identification using kernel-based regularization: new insights on stability and consistency issues. *Automatica* 93:321–32
75. Cucker F, Smale S. 2001. On the mathematical foundations of learning. *Bull. Am. Math. Soc.* 39:1–49

76. Argyriou A, Dinuzzo F. 2014. A unifying view of representer theorems. In *Proceedings of the 31st International Conference on Machine Learning*, ed. EP Xing, T Jebara, pp. 748–56. Proc. Mach. Learn. Res. 32(2). N.p.: PMLR
77. Pillonetto G, Chiuso A, De Nicolao G. 2010. Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proceedings of the 2010 American Control Conference*, pp. 498–503. New York: IEEE
78. Zorzi M, Chiuso A. 2018. The harmonic analysis of kernel functions. *Automatica* 94:125–37
79. Allen DM. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16:125–27
80. Wang L, Cluett W. 1996. Use of PRESS residuals in dynamic system identification. *Automatica* 32:781–84
81. Craven P, Wahba G. 1979. Smoothing noisy data with spline functions. *Numer. Math.* 31:377–403
82. Golub G, Heath M, Wahba G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21:215–23
83. Aravkin A, Bell B, Burke J, Pillonetto G. 2015. The connection between Bayesian estimation of a Gaussian random field and RKHS. *IEEE Trans. Neural Netw. Learn. Syst.* 26:1518–24
84. Cox R. 1946. Probability, frequency, and reasonable expectation. *Am. J. Phys.* 14:1–13
85. MacKay D. 1992. Bayesian interpolation. *Neural Comput.* 4:415–47
86. De Nicolao G, Sparacino G, Cobelli C. 1997. Nonparametric input estimation in physiological systems: problems, methods and case studies. *Automatica* 33:851–70
87. Pillonetto G, Chiuso A. 2015. Tuning complexity in regularized kernel-based regression and linear system identification: the robustness of the marginal likelihood estimator. *Automatica* 58:106–17
88. Gilks W, Richardson S, Spiegelhalter D. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall
89. Ninness B, Henriksen S. 2010. Bayesian system identification via MCMC techniques. *Automatica* 46:40–51
90. Andrieu C, Doucet A, Holenstein R. 2010. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. B* 72:269–342
91. Bottegal G, Aravkin A, Hjalmarsson H, Pillonetto G. 2016. Robust EM kernel-based methods for linear system identification. *Automatica* 67:114–26
92. Gunter L, Zhu J. 2007. Efficient computation and model selection for the support vector regression. *Neural Comput.* 19:1633–55
93. Dinuzzo F, Neve M, De Nicolao G, Gianazza U. 2007. On the representer theorem and equivalent degrees of freedom of SVR. *J. Mach. Learn. Res.* 8:2467–95
94. Dinuzzo F, De Nicolao G. 2009. An algebraic characterization of the optimum of regularized kernel methods. *Mach. Learn.* 74:315–45
95. Mairal J, Bach F, Ponce J, Sapiro G. 2010. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.* 11:19–60
96. Donoho D. 2006. Compressed sensing. *IEEE Trans. Inf. Theory* 52:1289–306
97. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58:267–88
98. Efron B, Hastie T, Johnstone L, Tibshirani R. 2004. Least angle regression. *Ann. Stat.* 32:407–99
99. Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68:49–67
100. Zou H. 2006. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* 101:1418–29
101. Mackay D. 1994. Bayesian non-linear modelling for the prediction competition. *ASHRAE Trans.* 100:3704–16
102. Wipf D, Nagarajan S. 2007. A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems 20*, ed. JC Platt, D Koller, Y Singer, ST Roweis, pp. 1625–32. Red Hook, NY: Curran
103. Tipping M. 2001. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 1:211–44

104. Mitchell TJ, Beauchamp JJ. 1988. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* 83:1023–32
105. Wipf D, Nagarajan S. 2010. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE J. Sel. Top. Signal Process.* 4:317–29
106. Materassi D, Innocenti G. 2010. Topological identification in networks of dynamical systems. *IEEE Trans. Autom. Control* 55:1860–71
107. Chiuso A, Pillonetto G. 2012. A Bayesian approach to sparse dynamic network identification. *Automatica* 48:1553–65
108. Bottegal G, Chiuso A, van den Hof P. 2018. On dynamic network modeling of stationary multivariate processes. *IFAC-PapersOnLine* 51(15):850–55
109. Granger C. 1963. Economic processes involving feedback. *Inf. Control* 6:28–48
110. Lind I, Ljung L. 2008. Regressor and structure selection in NARX models using a structured ANOVA approach. *Automatica* 44:383–95
111. Hong X, Mitchell RJ, Chen S, Harris CJ, Li K, Irwin GW. 2008. Model selection approaches for non-linear system identification: a review. *Int. J. Syst. Sci.* 39:925–46
112. Fazel M, Kei PT, Sun D, Tseng P. 2013. Hankel matrix rank minimization with applications to system identification and realization. *SIAM J. Matrix Anal. Appl.* 34:946–77
113. Prando G, Chiuso A, Pillonetto G. 2017. Maximum entropy vector kernels for MIMO system identification. *Automatica* 79:326–39
114. Zorzi M, Sepulchre R. 2016. AR identification of latent-variable graphical models. *IEEE Trans. Autom. Control* 61:2327–40
115. Zorzi M, Chiuso A. 2017. Sparse plus low rank network identification: a nonparametric approach. *Automatica* 76:355–66
116. Pillonetto G. 2016. A new kernel-based approach to hybrid system identification. *Automatica* 70:21–31
117. Goethals I, Pelckmans K, Suykens J, De Moor B. 2005. Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica* 41:1263–72
118. Falck T, Pelckmans K, Suykens J, De Moor B. 2009. Identification of Wiener-Hammerstein systems using LS-SVMs. *IFAC Proc. Vol.* 42(10):820–25
119. Goethals I, Pelckmans K, Falck T, Suykens J, De Moor B. 2010. NARX identification of Hammerstein systems using least-squares support vector machines. In *Block-Oriented Nonlinear System Identification*, ed. F Giri, EW Bai, pp. 241–58. London: Springer
120. Falck T, Dreesen P, Brabanter KD, Pelckmans K, Moor BD, Suykens J. 2012. Least-squares support vector machines for the identification of Wiener-Hammerstein systems. *Control Eng. Pract.* 20:1165–74
121. Lindsten F, Schön T, Jordan M. 2012. A semiparametric Bayesian approach to Wiener system identification. *IFAC Proc. Vol.* 45(16):1137–42
122. Tether A. 1970. Construction of minimal linear state-variable models from finite input-output data. *IEEE Trans. Autom. Control* 15:427–36
123. Fazel M, Hindi H, Boyd S. 2001. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, Vol. 6, pp. 4734–39. New York: IEEE
124. Rudin W. 1987. *Real and Complex Analysis*. Singapore: McGraw-Hill