Reproducing kernel Hilbert space based estimation of systems of ordinary differential equations ☆

Javier González [a,*], Ivan Vujačić [b], Ernst Wit [b]

[a] Sheffield Institute for Translational Neuroscience, University of Sheffield, Glossop Road S10 2HQ, Sheffield, UK
[b] Johann Bernoulli Institute of Mathematics and Computer Science, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands

ABSTRACT

Non-linear systems of differential equations have attracted the interest in fields like system biology, ecology or biochemistry, due to their flexibility and their ability to describe dynamical systems. Despite the importance of such models in many branches of science they have not been the focus of systematic statistical analysis until recently. In this work we propose a general approach to estimate the parameters of systems of differential equations measured with noise. Our methodology is based on the maximization of the penalized likelihood where the system of differential equations is used as a penalty. To do so, we use a reproducing kernel Hilbert space approach that allows us to formulate the estimation problem as an unconstrained numeric maximization problem easy to solve. The proposed method is tested with synthetically simulated data and it is used to estimate the unobserved transcription factor cdaR in *Streptomyces coelicolor* using gene expression data of the genes it regulates.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Despite the fact that differential equations are a common modelling tool within science and engineering, statistical methods for estimating such models have only received wide-spread attention during the last few years. The difficulty of solving differential equations in general has been a major stumbling block for efficient statistical procedures. Only in the last years, serious progress has been made on the estimation of parameters within systems of differential equations measured with noise. More importantly, solving the differential equation in these methods is not necessary for estimating the parameters of the differential equation. Ramsay et al. [19] introduced an approximate, two-stage maximum likelihood estimation procedure, where the solution of the differential equation was linked to a smoothed version of the data. In [5] a Bayesian method is proposed similar in spirit, whereby the solution of the ordinary differential equation (ODE) was given as a Gaussian process prior. Estimation was effectively again a two-stage process, where the *product of experts* provided a crucial link. In [21] a kernel method is developed for estimating 1-dimensional, periodic differential equations. In [6] a fully Bayesian inferential framework is developed to quantify uncertainty in ODE models. In this paper, we combine the frequentist set-up, such as in [19], with the kernel

approach of [21]. The main advantage is that we can define the estimation problem explicitly as a maximum likelihood problem, whereby the differential equation is interpreted as a constraint. By introducing a reproducing kernel Hilbert space (RKHS), we transform the constrained maximization problem into an unconstrained maximization problem. We detail this idea in Sections 3 and 4 after a revision of the main properties of RKHSs and penalized likelihood models in Section 2. In Section 5 we focus on the implementation of our methodology. Sections 6 and 7 illustrate the behaviour of the technique in simulated and real data scenarios, respectively. We conclude in Section 8 with a discussion of practical results of this work.

2. The RKHS framework: Green's functions, penalized likelihood models and Gaussian processes

Reproducing kernel Hilbert spaces [3,7] have played an important role in a number of successful applications in the last decades [17,23,15,10]. In this work, we use this theory for ODE estimation in the context of constrained likelihood models.

An RKHS is a Hilbert space of functions where all the linear evaluation functionals $\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$ such that $\mathcal{F}_t(x) = x(t)$, for $t \in T$, are bounded. By virtue of the Riesz representation theorem, for each $t \in T$ there exists $k_t \in \mathcal{H}$ such that for every $x \in \mathcal{H}$, $x(t) = \langle k_t, x \rangle$, where $\langle , \rangle$ denotes the inner product in $\mathcal{H}$. The RKHS $\mathcal{H}$ is uniquely characterized by a continuous, symmetric and positive definite function $K : X \times X \to \mathbb{R}$ named Mercer Kernel or reproducing kernel

for $\mathcal{H}$ [3]. All linear combinations $x(t) = \sum_i \alpha_i K(t, t_i)$ where $\alpha_i \in \mathbb{R}$ and $t_i \in T$ equipped with the adequate inner product form a space which is dense in $\mathcal{H}$.

Let T be a closed interval and consider a random sample $S = \{(y_i, t_i) \in \mathbb{R} \times T\}_{i=1}^n$ made up of n independent observations. Assume that the conditional distribution of y_i given t_i is Gaussian such that $y_i | t_i, x, \sigma^2 \sim \mathcal{N}(x(t_i), \sigma^2)$, where σ^2 is the variance of the model and $x : T \to \mathbb{R}$ is the target regression function to be estimated. To estimate x, one needs to restrict x to belong to a particular class of models, which can be done by penalizing the likelihood with a convex functional acting on x, generally a norm or seminorm in some Hilbert space $\mathcal{H}$. A widely studied approach is to use a differential operator P to impose smoothing conditions on x [9]. In the Gaussian case, assuming that σ^2 is known, the resulting penalized log-likelihood is

$$l_\lambda(x|S, \sigma^2) = -\frac{1}{2\sigma^2} \|x(\mathbf{t}) - \mathbf{y}\|^2 - \frac{\lambda}{2} \|Px\|_{L_2}^2, \tag{1}$$

where $\|\cdot\|_{L_2}$ is L_2 norm, $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\lambda > 0$ controls the balance between the fitting to the data and the smoothness of the model.

The maximization of (1) can be written as a regularization problem in an RKHS [4,18]. To do so, a way to connect the differential operator P with a kernel in a Hilbert space is required. A semi-constructive way to build such kernel is to use the concept of the Green's function. The Green's function of a linear operator P is a function $\mathcal{G} : T \times T \to \mathbb{R}$ that satisfies $P\mathcal{G}(t, z) = \delta(t - z)$ for δ the Dirac distribution. Roughly speaking $\mathcal{G}$ is the inverse of P [8]. Consider a differential operator P and let $\mathcal{K}$ be a Greens's function of P^*P where P^* is the adjoint operator of P. It can be shown [2] that $\|Px\|_{L^2}^2 = \|x\|_{\mathcal{H}_\mathcal{K}}^2$, for $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ the norm in the RKHS defined by $\mathcal{K}$.

Expression (1) can be understood as a projection mechanism onto the finite dimensional space spanned by $\{\mathcal{K}(t, t_i)\}$ when the penalization term $\|Px\|_{L^2}^2$ is replaced by $\|x\|_\mathcal{K}^2$ [12,14]. In particular, the maximizer of (1) is the function $\hat{x}(t) = \sum_{i=1}^n \hat{\alpha}_i \mathcal{K}(t, t_i)$ with coefficients $\hat{\alpha}_i$, obtained by maximizing

$$l_\lambda(x|S, \sigma^2) = -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathcal{K}[\mathbf{t}]\boldsymbol{\alpha}\|^2 - \frac{\lambda}{2} \boldsymbol{\alpha}^T \mathcal{K}[\mathbf{t}]\boldsymbol{\alpha}, \tag{2}$$

with respect to $\boldsymbol{\alpha} = (\alpha_1 \ldots, \alpha_n)^T$, where $\mathcal{K}[\mathbf{t}]$ is the matrix with entries $(\mathcal{K}[\mathbf{t}])_{ij} = \mathcal{K}(t_i, t_j)$. By using standard methods of differential calculus it can be shown that the solution to the maximization problem in (2) is given by $\hat{\boldsymbol{\alpha}} = (\lambda\sigma^2 \mathbf{I}_n + \mathcal{K}[\mathbf{t}])^{-1}\mathbf{y}$, where $\mathbf{I}_n$ is the n-dimensional identity matrix [22,10].

3. Explicit ODE and regularization approach

3.1. Explicit ODE

In this work, we are interested in dynamical systems with m interacting elements evolving in some closed time interval T. We denote by $x_j : T \to \mathbb{R}$ for $j = 1, \ldots, m$, the functions describing the evolution of the elements of the system and by $u_j : T \to \mathbb{R}$ for $i = 1, \ldots, p$, the action of p external forces. In compact notation, we denote by $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))^T$ and $\mathbf{u}(t) = (u_1(t), \ldots, u_p(t))^T$ the vectors of state variables and external forces respectively. We assume that each state variable x_j satisfies

$$P_{\theta_j} x_j(t) = f_j(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\beta}), \quad j = 1, \ldots, m, \; t \in T, \tag{3}$$

where $P_{\theta_j} = \sum_{k=1}^d \theta_{jk} d^{k-1}/dt$ is the linear differential operator associated to the jth equation of the system, which is defined by parameters $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd})^T$, and f_j is a known function depending on t through $\mathbf{x}(t)$ and $\mathbf{u}(t)$ for a vector of parameters $\boldsymbol{\beta}$. In the sequel, we refer to the whole set parameters of the system by $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$.

Typically, a finite sample of the states $\mathbf{x}$ is observed at a grid of time points $\mathbf{t} = (t_1, \ldots, t_n)^T$, that for simplicity we assume equal for all the states. The sample $\mathbf{y}(\mathbf{t})$ is made up of noisy measurement of the states of the ODE. That is, $\mathbf{y}(\mathbf{t})$ satisfies

$$\mathbf{y}(\mathbf{t}) = \mathbf{x}(\mathbf{t}) + \boldsymbol{\epsilon}(\mathbf{t}),$$

where $\boldsymbol{\epsilon}(\mathbf{t})$ represents a noise process. We assume $\boldsymbol{\epsilon}(\mathbf{t})$ is independent multivariate zero-mean Gaussian noise with variance σ_j^2 for each jth state. Other noise models are possible though not trivial.

Let $\mathbf{y}_j$ indicate the available data for state j and let $x_j(\mathbf{t})$ indicate the vector of values corresponding to evaluated jth state at time points $\mathbf{t}$. The *log-likelihood* of the model given the sample $\mathbf{y}(\mathbf{t})$ is given by

$$l(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}|\mathbf{y}(\mathbf{t})) = -\sum_{j=1}^m \frac{1}{2\sigma_j^2} \|\mathbf{y}_j - x_j(\mathbf{t})\|^2 \text{ given that}$$

$$P_{\theta_j} x_j(t) = f_j(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\beta}), \tag{4}$$

where $\mathbf{x}$ depends on $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ although it is not explicitly specified. Indeed, for each set of parameters $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ a family of feasible solutions of the system of differential equation is available for different initial conditions $\mathbf{x}(0)$.

3.2. Regularization

The maximum likelihood estimators of $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$ and $\mathbf{x}$ require explicit solutions of the differential equation, which are generally unknown. Alternatively, computational ODE solvers result intractable in cases with a large number of parameters. Our goal is to reformulate (4) in order to obtain a computationally tractable solution that does not require an explicit solution of the differential equation. The key element of our approach is the penalized log-likelihood

$$l_\lambda(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}|\mathbf{y}(\mathbf{t})) = -\sum_{j=1}^m \frac{1}{2\sigma_j^2} \|\mathbf{y}_j - x_j(\mathbf{t})\|^2 - \frac{\lambda}{2} \sum_{j=1}^m \Omega(x_j), \tag{5}$$

where $\Omega(x_j)$ is a convex functional that adds to the probabilistic model of the data the information provided by the ODE and $\lambda > 0$ the regularization parameter. As we detail next, an RKHS representation of the ODE can be used to define $\Omega(x_j)$ avoiding the computational burden of numerical ODE solvers.

4. RKHS representation of ODE systems

If the system is homogeneous, each equation

$$P_{\theta_j} x_j = 0 \tag{6}$$

is independent of the rest. Consider the set of functions $\mathcal{H}_j = \{x : \|P_{\theta_j} x_j\|^2 = 0\}$. By the definition of the norm, $\mathcal{H}_j$ contains all the solutions of the differential equation $P_{\theta_j} x_j = 0$ for some fixed $\theta_j : \|P_{\theta_j} x_j\|^2 = 0$ if and only if $P_{\theta_j} x = 0$. Using the connection between differential operators and RKHS detailed in Section 2, we can express the set $\mathcal{H}_j$ as $\mathcal{H}_j = \{x : \|x\|_{\mathcal{K}_{\theta_j}}^2 = 0\}$ where $\mathcal{K}_{\theta_j}$ is the Green's function of $P_{\theta_j}^* P_{\theta_j}$ and $\|\cdot\|_{\mathcal{K}_{\theta_j}}$ the norm in the RKHS defined by $\mathcal{K}_\theta$. Therefore, it makes sense to define

$$\Omega(x_j) = \|P_{\theta_j} x_j\| = \|x_j\|_{\mathcal{K}_{\theta_j}}^2, \tag{7}$$

as penalty and proceed using the properties of penalized likelihood models in RKHS described in Section 2. Note that by using this penalty in (5) each x_j is assumed to be an element in $\mathcal{H}_{\theta_j}$ rather than a solution of the differential equation unless we force $\|x\|_{\mathcal{K}_{\theta_j}}^2$ to be zero, which can be imposed by taking $\lambda \to \infty$. In this case, approaches in (4) and (5) are equivalent.

The main drawback is that a closed-form expression for the Green's function of $P_{\theta_j}^* P_{\theta_j}$ is rarely available. However, note that

as shown in (2), to solve the regularization problem only the evaluation of $\mathcal{K}_{\theta_j}$ on the sampled time points is required. Consider the difference equation given by

$$\mathbf{P}_{\theta_j}\mathbf{x}_j = 0, \tag{8}$$

where $\mathbf{P}_{\theta_j} = \sum_{k=1}^{d} \theta_{jk}\mathbf{D}^{k-1}$ is a d-order polynomial difference operator (matrix) acting on elements defined on $\mathbf{t}$ and where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a first order difference operator such that $(\mathbf{Dx}_j)_1 = (t_2 - t_1)^{-1}(\mathbf{x}_{j,2} - \mathbf{x}_{j,1})$, $(\mathbf{Dx}_j)_i = (2(t_{i+1} - t_{i-1}))^{-1}(\mathbf{x}_{j,i+1} - \mathbf{x}_{j,i-1})$ and $(\mathbf{Dx}_j)_n = (t_n - t_{n-1})^{-1}(\mathbf{x}_{j,n} - \mathbf{x}_{j,n-1})$. Note that this difference operator is different to the one proposed in [21], which it is only valid for periodic differential equations. The following proposition holds.

Proposition 1. [21] *Denote by $\mathcal{H}$ the space of functions (vectors) $\mathbf{x}_j : \mathbf{t} \subset T \to \mathbb{R}$ equipped with the usual L_2 inner product $\langle \cdot, \cdot \rangle$. Then the Green's function of the difference operator $\mathbf{P}_{\theta_j}^* \mathbf{P}_{\theta_j}$, where $\mathbf{P}_{\theta_j}^*$ is the adjoint (transpose) operator of $\mathbf{P}_{\theta_j}$, is the $s \times s$ dimensional matrix*

$$\mathbf{K}_{\theta_j} = (\mathbf{P}_{\theta_j}^T \mathbf{P}_{\theta_j})^{-1}. \tag{9}$$

Roughly, the proof uses the fact that $\mathbb{R}^n$ is an RKHS whose reproducing kernel is the location function δ_i, whose inner product with $x_j(\mathbf{t})$ yields $\delta_i^T x_j(\mathbf{t}) = x_j(t_i)$ and therefore only the inverse of $\mathbf{P}_{\theta_j}^T \mathbf{P}_{\theta_j}$ is needed. See [21] for details.

Penalty $\Omega(x_j)$, although originally defined for the function x_j, affects the regularization process through the values of x_j evaluated at a finite dimensional grid, generally the one corresponding to the available sample. By replacing the differential equation in (6) by the difference equation (8) and using Proposition 1, we define the penalty of the log-likelihood as

$$\Omega(x_j) = \|\mathbf{P}_{\theta_j}\mathbf{x}_j\|^2 = \mathbf{x}_j^T \mathbf{P}_{\theta_j}^T \mathbf{P}_{\theta_j}\mathbf{x}_j = \boldsymbol{\alpha}_j^T \mathbf{K}_{\theta_j}\boldsymbol{\alpha}_j,$$

where $\boldsymbol{\alpha}_j = \mathbf{K}_{\theta_j}^{-1}\mathbf{x}_j$.

4.1. Generalizations to nonhomogeneous equations

In general, the interest in inferring parameters of ODE is for systems which are not homogeneous. In the same spirit of above, one might like to consider

$$\Omega(x_j) = \|P_{\theta_j}x_j - f_j(\mathbf{x}, \mathbf{u}, \boldsymbol{\beta})\|^2 \tag{10}$$

as a penalty. However (10) cannot be directly used as penalizer for two reasons. Expression (10) cannot be reformulated as a norm of x_j in an RKHS. Note that when $x_j = 0$ then $\|P_{\theta_j}x_j - f_j(\mathbf{x}, \mathbf{u}, \boldsymbol{\beta})\|^2$ is not necessarily zero. Also, in this case the equations of the ODE are not independent. In a general setting, each x_j is affected by $x_1, \ldots, x_n$.

To circumvent previous problem we follow an approach that reduces the nonhomogeneous system to a homogeneous one. The key aspect is to consider that each f_j is a function of $\boldsymbol{\beta}$ that does not depends directly on $\mathbf{x}$ but that still reflects the dynamics of the system. To do so, we replace $\mathbf{x}$ by some fixed surrogate, namely $\mathbf{x}'(t)$, which is independent of θ and $\boldsymbol{\beta}$ and that it is estimated using the data. In Section 5 we elaborate on the definition of an appropriate $\mathbf{x}'$.

In general, in order to find a RKHS representation of the ODE system we assume that a Green's function G_j of each P_{θ_j} exists. Consider

$$\tilde{x}_j(t) = x_j(t) - x_j^*(t), \tag{11}$$

where $x_j^*(t) = \int_T G_j(z,t)f_j(\mathbf{x}'(z), \mathbf{u}(z), \boldsymbol{\beta})dz$ is a collection of solutions of the differential equation. Since P_{θ_j} is a linear operator we obtain that, for all $\tilde{x}_j$,

$$P_{\theta_j}\tilde{x}_j(t) = P_{\theta_j}x_j(t) - P_{\theta_j}x_j^*(t) = P_{\theta_j}x_j(t) - f_j(\mathbf{x}'(z), \mathbf{u}(z), \boldsymbol{\beta}),$$

which includes the trivial solution $\tilde{x}_j = 0$. We can now write

$$\Omega(\tilde{x}_j) = \|P_{\theta_j}x_j(t) - P_{\theta_j}x_j^*(t)\|^2 = \|P_{\theta_j}\tilde{x}_j\|^2, \tag{12}$$

which can be studied and computed as a norm for $\tilde{x}_j$ in $\mathcal{H}_{\theta_j}$ equivalent to (7). Note that the surrogate $\mathbf{x}'$ is essential to linearize the system and to write $\Omega(\tilde{x}_j)$ as a norm. In practice, noise samples are obtained for x_j and not for $\tilde{x}_j$. Therefore to focus the inference problem on $\tilde{x}_j$ requires to transform the original data. Details are given in Section 5.

5. Approximate ODE inference

The goal of this section is to provide computational details to infer the set of parameters $\{\theta, \boldsymbol{\beta}\}$ using the approximate ODE representation described in Section 4. As detailed in Section 4.1, a definition of each x_j', a surrogate of x_j is required. In this work we express each x_j' in terms of a penalized splines basis expansion,

$$x_j'(t) = \sum_{k=1}^{q} c_{jk}\phi_k(t) = \mathbf{c}_j^T \boldsymbol{\phi}(t) \tag{13}$$

for $\boldsymbol{\phi} = (\phi_1(t), \ldots \phi_q(t))$ and where each $\phi_j(t)$ is a piecewise polynomial function of degree k whose brake points or knots are located at $t_1, \ldots, t_n$. The coefficients $\mathbf{c}_j$ can be explicitly estimated by $\mathbf{c}_j = (\Phi^T \mathbf{W}\Phi)^{-1}\Phi^T \mathbf{W}\mathbf{y}_j$ for Φ the design matrix and where $\mathbf{W}$ is a weight matrix which allows for possible covariance structure among the residuals. See [20] for details. The number of basis functions q is assumed to be large enough to capture the variation of the solutions of the ODE. Here we assume that coefficients $\mathbf{c}_j$ are fixed values obtained by smoothing the data but further generalizations are possible.

Definition 1. Consider the penalized likelihood model in (5). Consider the objects $\mathbf{t}$, $\mathbf{y}_j$, $\mathbf{x}_j$ and $\mathbf{P}_{\theta_j}$ as defined above. Let $\hat{x}_1', \ldots, \hat{x}_m'$ be estimates of (13) for $j = 1, \ldots, m$. We define the approximated pseudo-log-likelihood of the ODE model associated to (5) as

$$l_\lambda(\theta, \boldsymbol{\beta}|\mathbf{y}(\mathbf{t}), \hat{\mathbf{x}}') = -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\mathbf{y}_j - \mathbf{x}_j\|^2$$
$$-\frac{\lambda}{2}\sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\mathbf{x}_j - f_j(\hat{\mathbf{x}}'(\mathbf{t}), \mathbf{u}(\mathbf{t}), \boldsymbol{\beta})\|^2,$$

where $\lambda > 0$ and $\hat{\mathbf{x}}'(\mathbf{t}) = (\hat{x}_1(\mathbf{t}), \ldots, \hat{x}_m(\mathbf{t}))^T$.

Next, we show how to compute l_λ in practice.

Proposition 2. *Assume that $\mathbf{P}_{\theta_j}^{-1}$ exists and define,*

$$\tilde{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{P}_{\theta_j}^{-1}f_j(\mathbf{u}(\mathbf{t}), \hat{\mathbf{x}}'(\mathbf{t}), \boldsymbol{\beta}). \tag{14}$$

for $j = 1, \ldots, m$. Consider the function

$$g_\lambda(\theta, \boldsymbol{\beta}|\mathbf{y}(\mathbf{t})) = \sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\tilde{\mathbf{y}}_j^T\left[\mathbf{I} - (\mathbf{I} + \sigma^2\lambda\mathbf{K}_{\theta_j}^{-1})^{-1}\right]\tilde{\mathbf{y}}_j. \tag{15}$$

It holds that

$$\{\hat{\theta}, \hat{\boldsymbol{\beta}}\} = arg\max_{\theta, \boldsymbol{\beta}} l_\lambda(\theta, \boldsymbol{\beta}, \mathbf{y}(\mathbf{t})) = arg\max_{\theta, \boldsymbol{\beta}} g_\lambda(\theta, \boldsymbol{\beta}|\mathbf{y}(\mathbf{t})).$$

Remark that we do not need the explicit computation of $\mathbf{K}_{\theta_j}$. Only its inverse is needed, which can be obtained directly from (9). Notice that (14) performs the data transformation equivalent to (11) that is needed to obtain a RKHS representation of the ODE in general cases. Optimization of (2) with a conjugate gradient method produces estimates of $\{\theta, \boldsymbol{\beta}\}$. If the set of parameters of the systems is separable by equations, independent optimization can be done for those, which helps to avoid local minima and speed up the procedure. Finally, estimates for each $\hat{\mathbf{x}}_j$ are available by means of

$$\hat{\mathbf{x}}_j = \mathbf{K}_{\hat{\theta}_j}\hat{\boldsymbol{\alpha}}_j + \mathbf{P}_{\hat{\theta}_j}^{-1} f_j(\mathbf{u}(\mathbf{t}), \hat{\mathbf{x}}'(\mathbf{t}), \hat{\boldsymbol{\beta}}), \tag{16}$$

where $\hat{\boldsymbol{\alpha}}_j = (\mathbf{K}_{\hat{\theta}_j} + \lambda\sigma_j^2\mathbf{I})^{-1}\tilde{\mathbf{y}}_j$.

5.1. Model selection

The effective number of parameters for each equations is defined as $df_j = Tr(\mathbf{K}_{\hat{\theta}_j}(\mathbf{K}_{\hat{\theta}_j} + \lambda\sigma_j^2\mathbf{I}))^{-1}$ where $Tr(\cdot)$ represents the trace operator. We propose as model selection criterion for λ the Akaike information criterion [1], which is defined in our context as

$$AIC(\lambda) = -2l(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y}(\mathbf{t}), \hat{\mathbf{x}}') + 2\sum_{j=1}^{m} df_j. \tag{17}$$

In practice, the minimum AIC for a grid of penalization parameters $\lambda_1, \ldots, \lambda_k$ should be used to select the optimal model. Note, however, that one can simply select a large value of λ to force the regularization approach to find an exact solution of the ODE model.

5.2. Variance of the parameter estimates and confidence intervals

Confidence intervals and standard errors of the parameter estimates can be obtained from the Hessian matrix $\mathbf{H}_{l_\lambda}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$, which is available as output of the Conjugate Gradient algorithm [16] used to optimize l_λ. The covariance matrix of the parameter estimates is therefore obtained as $\hat{\Sigma} = -\mathbf{H}_{l_\lambda}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})^{-1}$ and its diagonal (variances of the parameters) used to estimate calculate the Wald confidence intervals.

6. Examples using synthetically generated data

6.1. Explicit ODE versus regularization approach

In this section we use a toy example to illustrate the advantages of using a regularization approach to estimate the parameters of a dynamical system. We consider the differential equation $dx/dt - \theta x = 0$ where $D = d/dt$. For fixed θ and initial condition $x(0)$ the solution of the differential equation is given by $x(t) = x(0)\exp\{\theta t\}$. We fix $\theta = -2$, $x(0) = -1$ and we generate 500 samples of 10 equally spaced points in the interval $[0, 2]$ using Gaussian noise with $\sigma = 0.25$. For each sample we calculate the maximum likelihood estimator (MLE) of θ and our RKHS based

estimator for λ selected by means of the Akaike Information criterion (AIC) (see Section 5.1). The averaged absolute deviance to the true parameter of the MLE is 0.73 with a standard deviation of 1.03 whereas the averaged error for the penalized approach is 0.53 with a standard deviation of 0.38. In Fig. 1 we show the results for one run of the experiment. Fig. 1(a) shows the negative log likelihood and the penalized log-likelihood of the model for one of the generated data sets. The penalized approach results in an 'improvement' of the original likelihood for parameter estimation with a minimum closer to the true value of the parameter. Also note that the original negative log likelihood becomes extremely flat for small values of the parameters, which can produce computational problems in the optimization step. Fig. 1(b) shows the MLE and RKHS solutions together with the true function $x(t) = -\exp\{2t\}$ for $t \in [0, 2]$. Penalizing the likelihood improves the estimator in this example. The true function x is better approximated using the penalized approach due to the finite sample bias of the MLE. Also the estimate of the parameter is closer to the true value of θ in this particular realization ($\hat{\theta}_{MLE} = -2.55$ vs. $\hat{\theta}_{RKHS} = -2.05$).

6.2. Lotka Volterra equations

In this experiment we work with the Lotka–Volterra system of differential equations originally proposed in the theory of auto-catalytic chemical reaction [13]. The formulation of the system is given by

$$\frac{dx_1}{dt} = x_1(\theta_1 - \beta_1 x_2), \quad \frac{dx_2}{dt} = -x_2(\theta_1 - \beta_2 x_2), \tag{18}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ are the parameters.

Our aim is to evaluate the accuracy and speed of our RKHS penalized approach in comparison with the classical MLE approach. To do so, we run a simulation study for fixed $\theta_1 = 0.2$, $\beta_1 = 0.35$ $\theta_2 = 0.7$, $\beta_2 = 0.40$ and initial conditions $x_1(0) = 1$ and $x_2(0) = 2$. We generate samples made up of n fixed and equally spaced independent noisy observations of the state variables x_1 and x_2 in the interval $T = [0, 30]$ that we perturb with zero mean Gaussian noise with standard deviation σ. We generate data for the samples sizes $n = 35, 70, 100$ and two noise scenarios $\sigma = 0.1, 0.25$. In Fig. 2(a) we show the true solutions of the model
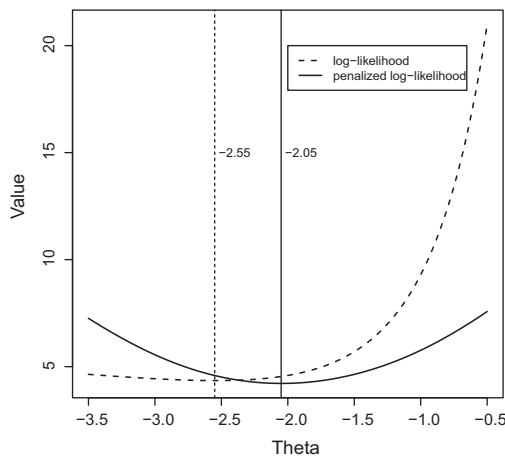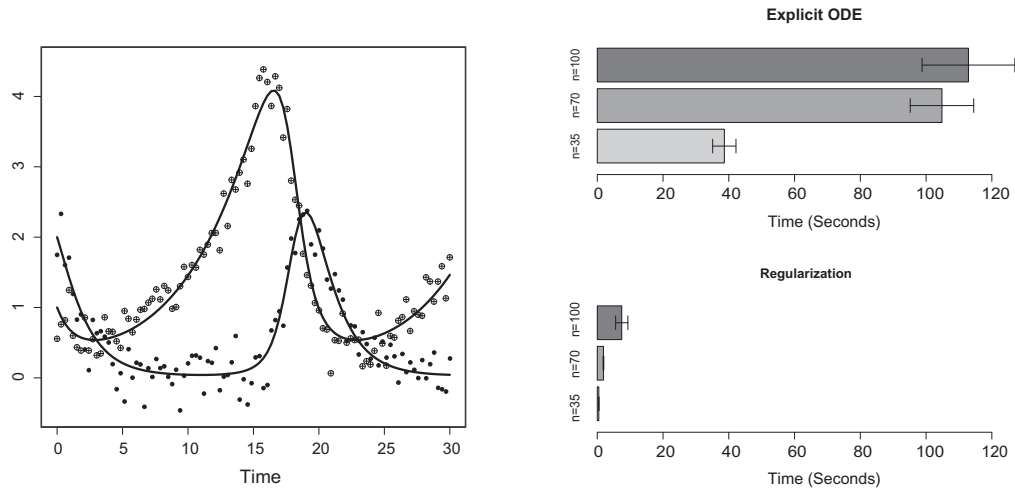

(a) Negative log likelihood and the penalized log-likelihood of the model for one of the generated data sets.
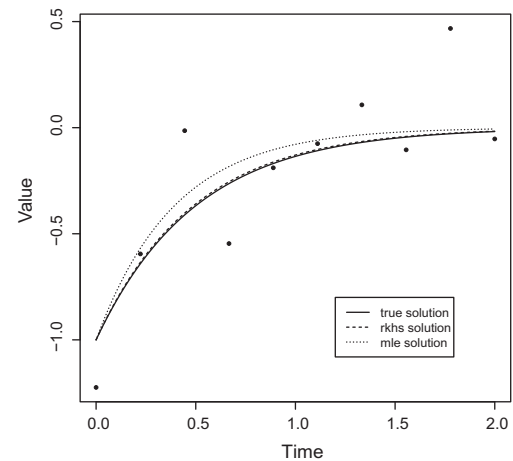
(b) Simulated data, true solution, and the two estimated solutions using the MLE and RKHS approaches.

Fig. 1. Results obtained for the differential equation $dx/dt - \theta x = 0$.

(a) True solutions and the data in the Lotka-Volterra experiment for $n = 100$ and $\sigma = 0.25$.

(b) Computational cost comparison between the explicit MLE approach and the proposed penalized RKHS based approach.

Fig. 2. Results obtained for the Lotka–Volterra equations.

for the above mentioned parameters together with the data of one of the simulations.

In order to apply the proposed approach we obtain the functions $\hat{x}'_1$ and $\hat{x}'_2$ using penalized splines and we fix $\lambda = 100$. To perform the MLE estimation we use a conjugate gradient algorithm with 10 different initial values of the parameters (randomly generated in the interval $[0, 1]$) and we use the likelihood value to select the best candidate.

In Fig. 2(b) we show a comparison of computational times for the averaged running times. The RKHS-based is 120.08, 24.06 and 14.41 times faster than the explicit ODE approach for $n = 35, 70$ and 100 respectively. In Table 1 we show the mean square errors of the estimates with respect to the true parameters for 100 runs of the experiment. For $n = 35$ the penalized RKHS approach performs significantly better than the explicit ODE estimates, which is explained by the empirical bias suffered by the MLE approach illustrated in Section 6.1. For $n = 75, 100$ both methods work similarly in terms of precision. Notice that the noise in the data is reflected in the precision of the estimates for both techniques; the errors are larger in all cases for $\sigma = 0.25$.

7. Real example: reconstruction of transcription factor activities in *Streptomyces coelicolor*

A gene regulatory network consists of a gene encoding a transcription factor (TF) together with the genes it regulates (genes whose activity can be activated or repressed by binding to the DNA). In the absence of reliable technology to measure the activity of the TF (number of TF-protein molecules in the cell), the problem is to reconstruct it from gene expression data of its target genes.

In this experiment we work with a data set of genes expression levels in the *S. coelicolor* bacterium. The goal is to reconstruct the activity of the transcription factor (TF) *cdaR* using 10-points time-series gene expression data of 17 genes. For each gene, two different series corresponding to a wild type and mutant type bacterium (for which a transcriptional regulator cdaR has been knocked out) are available. Measurements are available at time points (in mins.) $\mathbf{t} = \{16, 18, 20, 21, 22, 23, 24, 25, 39, 67\}$. The importance of understanding the behaviour of the cdaR relies on the fact that it is partially responsible for the production of a particular type of antibiotic.

Table 1
Mean square error for the inferred parameters in the Lotka–Volterra model. Standard deviations shown in parenthesis. The true value of the parameters are fixed to $\theta_1 = 0.2$, $\beta_1 = 0.35$ $\theta_2 = 0.7$, $\beta_2 = 0.40$.

σ	n	Method	Lotka–Volterra ODE model			
			$|\theta_1 - \hat{\theta}_1|$	$|\beta_1 - \hat{\beta}_1|$	$|\theta_2 - \hat{\theta}_2|$	$|\beta_2 - \hat{\beta}_2|$
0.1	35	RKHS	0.0002 (0.0003)	0.0007 (0.0007)	0.0031 (0.0036)	0.0014 (0.0014)
		MLE	0.0016 (0.0088)	0.0063 (0.0425)	0.0422 (0.1809)	0.0227 (0.1064)
	70	RKHS	0.0001 (0.0001)	0.0002 (0.0002)	0.0009 (0.0011)	0.0003 (0.0004)
		MLE	0.0000 (0.0001)	0.0001 (0.0006)	0.0017 (0.0034)	0.0005 (0.0010)
	100	RKHS	0.0001 (0.0001)	0.0001 (0.0002)	0.0005 (0.0006)	0.0002 (0.0002)
		MLE	0.0000 (0.0001)	0.0002 (0.0010)	0.0013 (0.0023)	0.0004 (0.0008)
0.25	35	RKHS	0.0010 (0.0013)	0.0017 (0.0024)	0.0111 (0.0205)	0.0038 (0.0059)
		MLE	0.0029 (0.0180)	0.0081 (0.0392)	0.0173 (0.0487)	0.0078 (0.0359)
	70	RKHS	0.0004 (0.0006)	0.0008 (0.0009)	0.0042 (0.0047)	0.0015 (0.0019)
		MLE	0.0007 (0.0025)	0.0030 (0.0115)	0.0151 (0.0474)	0.0062 (0.0301)
	100	RKHS	0.0003 (0.0004)	0.0005 (0.0006)	0.0034 (0.0043)	0.0011 (0.0016)
		MLE	0.0008 (0.0032)	0.0028 (0.0116)	0.0174 (0.0603)	0.0083 (0.0387)

Gene SCO3235 inferred profile

cdaR TF inferred activity
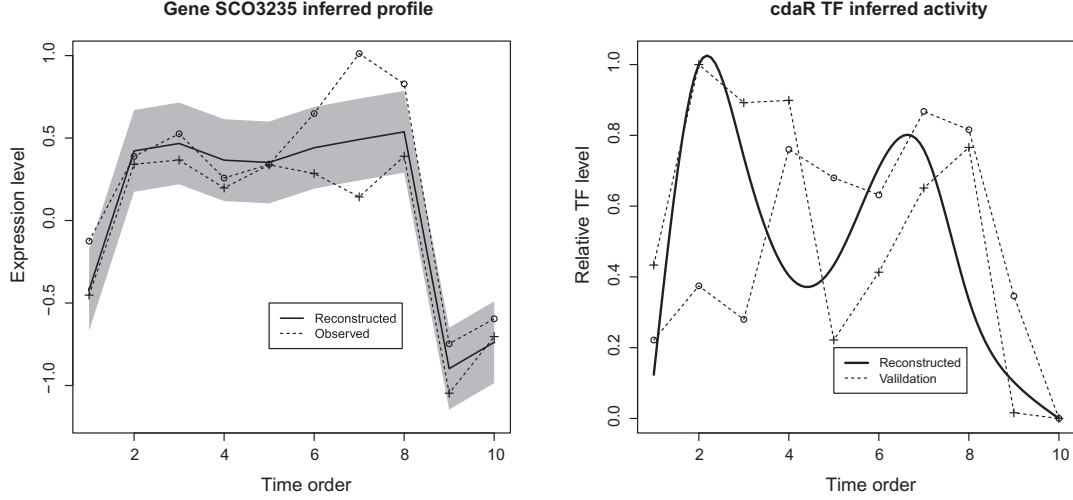
(a) Gene SCO3235, reconstructed profile. Circles and crosses represent the observed data and lines the obtained profiles. The estimated variance and initial conditions are $\hat{\sigma}^2 = 0.016$ and $\hat{x}(0) = -0.39$. The estimated parameters for this gene are $\beta_1 = 0.65$ (stdev= 0.57), $\beta_2 = 1.07(0.51)$, $\beta_3 = 2.09$ (0.26) and θ=1.06 (0.21).

(b) Reconstructed activity of the master activator cdaR scaled between 0 an 1. Circles and crosses represent the data obtained in two independent experiments not used in the estimation process.

Fig. 3. Reconstructed genes profiles and master activator cdaR.

Following Khanin et al. [11] we assume that changes in the expression levels of the genes are caused by changes in the cdaR protein and the mRNA degradation. We denote by $\eta(t)$ the activity profile of the regulator cdaR at time t and by $x_j(t)$ the expression level of each gene j in time t. This regulatory system is modelled by

$$\frac{d}{dt}x_j(t) + \theta_j x_j(t) = \beta_{1j} + \beta_{2j}\frac{\eta(t)}{\beta_{3j} + \eta(t)}, \qquad (19)$$

where θ_j is the rate of mRNA degradation, β_{2j} and β_{3j} are gene-specific kinetic parameters for the gene j and β_{1j} is an additive constant that accounts for the basal level of transcription and the nuisance effects from micro-arrays. The goal is to use the available sample to reconstruct the levels of the activator $\eta(t)$, which is unobserved, and the gene profiles via the estimation of the parameters in (19). To do so we apply the procedure described in Sections 3 and 4. We assume that the variance is equal for all the genes. For each gene, we work with the average of the two available time series. We model the activator η using a basis of cubic splines with equally spaced nodes, that is $\eta(t) = \sum_{j=1}^{15} a_j\phi_j$ where the ϕ_j's are elements of the basis and $a_1, \ldots, a_{15}$ parameters to estimate. We apply the procedure described in Section 5. We select the optimal penalization parameter λ by using the AIC as model selection criterion.

Fig. 3(a) shows the estimated profile for the gene SCO3235, which fits well the observed data. The reconstructed genes profiles exhibit a similar fit for the remaining genes. The reconstructed cdaR activator is shown in Fig. 3(b) together two independent replicates profiles obtained from a different experiment and that were not used in the estimation process. The values are normalized between 0 and 1 since the activity of the cdaR protein is expressed in arbitrary units and can be interpreted as relative levels. The estimated profile fits the observed data showing two hills around time points 4 and 9 similarly to the genes profiles. This agrees with the fact that cdaR is an activator of the genes activity. These result shows the ability of the proposed approach to identify correctly unknown elements of the ODE systems through a proper estimation of the parameters of the model. The estimation of the

parameters of this system takes around 15 mins in a personal laptop. Further, baseline maximum likelihood estimators are available in [11].

8. Conclusions and discussion

We have proposed a new method to estimate general systems of ordinary differential equations measured with noise. Our proposal is based on the penalization of the likelihood of the problem by means of the ODE. A reproducing kernel Hilbert space approach has provided the theoretical framework to make this idea feasible. The concept of Green's function and the connection between linear differential operators and Mercer kernels have been used to rewrite the penalized likelihood of the problem in a particular manner easy to handle in practice.

The main merit of the method is its ability to perform in a single step the estimation problem without solving the differential equation. In practice, our proposal is specially competitive in small sample scenarios as it is shown via simulation. A real example in system biology has been used to illustrate the utility of the method in scenarios with hidden components.

In the future, we aim to focus on further Bayesian extensions of this work, the estimation of the regularization parameter λ and on the analysis of the theoretical properties of the proposed method.

Appendix A

Proof of Proposition 2. We prove the proposition by showing that $l_\lambda(\theta, \beta|\mathbf{y}(\mathbf{t}), \hat{\mathbf{x}}')$ and $g_\lambda(\theta, \beta|\mathbf{y}(\mathbf{t}), \hat{\mathbf{x}}')$ are the same function. Consider the function

$$l_\lambda(\theta, \beta|\mathbf{y}(\mathbf{t}), \hat{\mathbf{x}}') = -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\mathbf{y}_j - \mathbf{x}_j\|^2 - \frac{\lambda}{2}\sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\mathbf{x}_j$$
$$- f_j(\hat{\mathbf{x}}'(\mathbf{t}), \mathbf{u}(\mathbf{t}), \beta)\|^2.$$

Denote by $f_j = f_j(\mathbf{u}(\mathbf{t}), \hat{\mathbf{x}}'(\mathbf{t}), \beta)$. We can rewrite the first term as

$$\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\mathbf{y}_j-\mathbf{x}_j\|^2 = \sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\mathbf{y}_j-\mathbf{P}_{\theta_j}^{-1}f_j+\mathbf{P}_{\theta_j}^{-1}f_j-\mathbf{x}_j\|^2$$

$$= \sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\tilde{\mathbf{y}}_j-\tilde{\mathbf{x}}_j\|^2,$$

where $\tilde{\mathbf{y}}_j=\mathbf{y}_j-\mathbf{P}_{\theta_j}^{-1}f_j$ and $\tilde{\mathbf{x}}_j=\mathbf{x}_j-\mathbf{P}_{\theta_j}^{-1}f_j$. Regarding the second term, we have that

$$\sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\mathbf{x}_j-f_j\|^2 = \sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\mathbf{x}_j-\mathbf{P}_{\theta_j}\mathbf{P}_{\theta_j}^{-1}f_j\|^2 = \sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}(\mathbf{x}_j-\mathbf{P}_{\theta_j}^{-1}f_j)\|^2$$

$$= \sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\tilde{\mathbf{x}}_j\|^2.$$

Therefore, we can rewrite $l_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}')$ as

$$l_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}') = -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\tilde{\mathbf{y}}_j-\tilde{\mathbf{x}}_j\|^2 - \frac{\lambda}{2}\sum_{j=1}^{m}\|\mathbf{P}_{\theta_j}\tilde{\mathbf{x}}_j\|^2.$$

Using the properties of RKHSs detailed in Section 4, we can write

$$l_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}') = -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\|\tilde{\mathbf{y}}_j-\mathbf{K}_{\theta_j}\alpha_j\|^2 - \frac{\lambda}{2}\sum_{j=1}^{m}\alpha_j^T\mathbf{K}_{\theta_j}\alpha_j$$

and expanding the first terms we obtain

$$-\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\left[\tilde{\mathbf{y}}_j^T\tilde{\mathbf{y}}_j - 2\tilde{\mathbf{y}}_j^T\mathbf{K}_{\theta_j}\alpha_j + \alpha_j^T\mathbf{K}_{\theta_j}^T\mathbf{K}_{\theta_j}\alpha_j + \sigma^2\lambda\alpha_j^T\mathbf{K}_{\theta_j}\alpha_j\right].$$

For fixed $\{\theta,\beta\}$ and σ_j^2 the maximum of $l_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}')$ is given for the vectors $\alpha_j=(\mathbf{K}_{\theta_j}+\sigma_j^2\lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}_j$. We set the derivative to zero to find the point of maximum. Substituting each α_j and simplifying we obtain that

$$l_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}') = -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\left[\tilde{\mathbf{y}}_j^T\tilde{\mathbf{y}}_j - 2\tilde{\mathbf{y}}_j^T\mathbf{K}_{\theta_j}(\mathbf{K}_{\theta_j}+\sigma_j^2\lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}_j\right.$$

$$\left.+\tilde{\mathbf{y}}_j^T(\mathbf{K}_{\theta_j}+\sigma_j^2\lambda\mathbf{I})^{-1}(\mathbf{K}_{\theta_j}+\sigma_j^2\lambda\mathbf{I})\mathbf{K}_{\theta_j}(\mathbf{K}_{\theta_j}+\sigma^2\lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}_j\right]$$

$$= -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\left[\tilde{\mathbf{y}}_j^T\tilde{\mathbf{y}}_j - \tilde{\mathbf{y}}_j^T\mathbf{K}_{\theta_j}(\mathbf{K}_{\theta_j}+\sigma_j\lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}_j\right]$$

$$= -\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\tilde{\mathbf{y}}_j^T\left[\mathbf{I}-(\mathbf{I}+\sigma_j^2\lambda\mathbf{K}_{\theta_j}^{-1})^{-1}\right]\tilde{\mathbf{y}}_j = g_\lambda(\theta,\beta|\mathbf{y}(\mathbf{t}),\hat{\mathbf{x}}')$$

as we aimed to prove. □

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (6) (1974) 716–723.
[2] N. Aronszajn, Green's functions and reproducing kernels, in: Proceedings of the Simposium on Spectral Theory and Differential Problems, Oklahoma, 1951, pp. 355–411.
[3] N. Aroszajn, Theory of reproducing kernels, Trans. Am. Math. Soc. 68 (3) (1950) 337–404.
[4] A. Berlinet, C. Thomas-Agnan, Reproducing kernel Hilbert spaces in probability and statistics, Springer, 2005.
[5] B. Calderhead, M. Girolami, N. Lawrence, Accelerating bayesian inference over nonlinear differential equations with gaussian processes, in: Neural Information Processing Systems, vol. 22, 2008.
[6] O. Chkrebtii, D.A. Campbell, M.A. Girolami, B. Calderhead, Bayesian Uncertainty Quantification for Differential Equations, 2013 (ArXiv e-prints).
[7] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Am. Math. Soc. 39 (1) (2001) 1–49.
[8] D.G. Duffy, Green's Functions with Applications, Chapman and Hall/CRC, 2001.
[9] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, vol. 58, Chapman & Hall, 1994.
[10] T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning, Ann. Stat. 36 (2008) 1171–1220.
[11] R. Khanin, V. Vinciotti, V. Mersinias, C. Smith, E. Wit, Statistical reconstruction of transcription factor activity using michaelismenten kinetics 63 (3) (2007) 816–823.
[12] G.S. Kimeldorf, G. Wahba, A correspondence between bayesian estimation on stochastic processes and smoothing by splines, Ann. Math. Stat. 41 (2) (1970) 495–502.
[13] A.J. Lotka, Contribution to the theory of periodic reaction, J. Phys. Chem. 14 (3) (1910) 271–274.
[14] X. Ma, B. Dai, R. Klein, B.E.K. Klein, K.E. Lee, G. Wahba, Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate data, Statistics 17 (1159) (2010) 46.
[15] J.M. Moguerza, A. Muñoz, Support vector machines with applications, Stat. Sci. 21 (4) (2006) 322–336.
[16] J. Nocedal, S.J. Wright, Numerical Optimization, second ed., Springer, New York, 2006.
[17] E. Parzen, Statistical inference on time series by RKHS methods, in: R. Pyke (Ed.), Proceedings 12th Biennial Seminar, Canadian Mathematical Congress, Montreal, 1970, pp. 1–37.
[18] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (9) (1990) 1481–1497.
[19] J.O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: a generalized smoothing approach, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 69 (5) (2007) 741–796.
[20] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., Springer, New York, 2006.
[21] F. Steinke, B. Scholkopf, Kernels, regularization and differential equations, Pattern Recogn. 41 (11) (2008) 3271–3286.
[22] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[23] G. Wahba, Spline models for observational data, Series in Applied Mathematics, SIAM, Philadelphia, 1990.