MACHINE LEARNING METHODS FOR ESTIMATING OPERATOR EQUATIONS

Florian Steinke * **Bernhard Schölkopf** *

* *Max Planck Institute For Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany*

Abstract: We consider the problem of fitting a linear operator induced equation to point sampled data. In order to do so we systematically exploit the duality between minimizing a regularization functional derived from an operator and kernel regression methods. Standard machine learning model selection algorithms can then be interpreted as a search of the equation best fitting given data points. For many kernels this operator induced equation is a linear differential equation. Thus, we link a continuous-time system identification task with common machine learning methods.
The presented link opens up a wide variety of methods to be applied to this system identification problem. In a series of experiments we demonstrate an example algorithm working on non-uniformly spaced data, giving special focus to the problem of identifying one system from multiple data recordings.

Keywords: machine learning, model selection, differential equations, continuous-time modelling

1. INTRODUCTION

In recent years kernel machines have attracted a lot of attention in machine learning; for an overview cf. (Schölkopf and Smola, 2002). In many regression and classification tasks they are among the top performers.

Given some training input-output data pairs $\{(x_i, y_i)\}_{i=1,\dots,m} \subseteq \mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is the domain of the data, these methods estimate a function $f : \mathcal{X} \to \mathbb{R}$ which is then used for predictions at previously unseen points $x \in \mathcal{X}$. All kernel methods share the use of a similarity measure $R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the so-called *kernel*. The estimated functions are given in terms of these kernels centred at the training data points, i.e.

$$f(x) = \sum_i \alpha_i R(x, x_i). \tag{1}$$

To obtain f, kernel methods typically minimize a functional like

$$\alpha^T K \alpha + C \operatorname{Loss}\{(f(x_i), y_i)\}_i \tag{2}$$

over all $\alpha \in \mathbb{R}^m$. The term $\alpha^T K \alpha$ with $K_{ij} = R(x_i, x_j)$ enforces smoothness in a manner dependent on the kernel R, and the second term $\operatorname{Loss}\{(f(x_i), y_i)\}_i$ measures how closely the function fits the given data points. Common choices for the loss function are the quadratic loss $\frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2$ or a one norm like loss such as $\frac{1}{m}\sum_{i=1}^m |f(x_i) - y_i|$.

Model selection algorithms try to determine the optimal kernel and its parameters for a given problem. One also has to choose the regularization parameter C which measures the relative weight of a prior belief of a smooth function versus preferring a function that exactly matches the given data points (Wahba, 1990). Numerous methods have been proposed to attack this task, including most recently (Argyriou et al., 2005; Sonnenburg et al., 2005). A simple, yet very common approach

is cross validation using the leave-one-out (LOO) error to evaluate the generalization performance.

It is well known that kernels give rise to a linear regularization operator $P : \mathcal{H} \to L_2(\mathcal{X})$ such that minimizing (2) is equivalent to minimizing

$$\|Pf\|_{L_2}^2 + C \text{ Loss}\{(f(x_i), y_i)\}_i \tag{3}$$

over all f chosen from an appropriate function space $\mathcal{H}$ (Wahba, 1990). For the commonly used translation invariant, radial basis kernels on $\mathcal{X} \subseteq \mathbb{R}^D$ this operator turns out to be a linear combination of derivative operators. Conversely, one could be given P in advance and construct a kernel R from it such that minimizing (2) yields the same result as does (3).

One of the main points of this paper is that minimizing (3) can be interpreted as solving the operator induced equation

$$Pf = 0 \tag{4}$$

with the "boundary conditions" that the solution function f approximately interpolates the given data points. The objective (3) tries to minimize the residual error of (4) in an L_2-norm sense while at the same time approximating the data points. For radial basis kernels the implied equation (4) is a linear differential equation.

Given the correspondence between kernels and regularization operators, the second idea is that determining the kernel R which models the underlying data distribution best – as done by model selection algorithms – is equivalent to finding an operator P such that (4) is optimally fulfilled in an L_2-sense. We could thus use an established model selection algorithm to determine the structure and the parameters of such an operator.

In the system identification literature there are many methods proposed to estimate a continuous-time differential equation from point sampled data sets (cf. the reviews of Ljung (2004) or Müller and Timmer (2002)). Most methods use two independent steps. For Fourier space methods a Fourier transform of the given signal is obtained first. If the signal is only given at discrete randomly spaced points this requires estimating an approximation function, e.g. piecewise linear, as an initial step. The identification of the differential equation then builds on the Fourier spectrum of the resulting function. This may not be a very good estimate of the true data structure as the function estimation step introduces new information into the process which renders the results dependent on a somewhat arbitrary smoothing step. Similarly, direct methods make use of estimated derivatives. These are computed using some smoothed model of a function, e.g. a spline estimate, where the smoothing model is chosen independently of the equation to be estimated. This problem has also been described in (Moussaoui et al., 2003).

The method we propose uses again a two step process, however, these steps are not independent. The smoothness assumption used in the first part, e.g. the estimation of an approximating function, is based on the current guess of the underlying equation. If this guess is close to the true structure we are not introducing arbitrary information in this step. Thus, we may be able to get better results than what is possible if one uses a fixed smoothness assumption. In the second step, we evaluate the predictive quality of our estimated smoothing function and adapt our current guess of the operator P accordingly. If a low prediction error can be achieved, the proposed equation models the given data set well.

In Section 2 we will take a closer look at the correspondence between a kernel R and its matching regularization operator P in order to be able to compute one given the other. Section 3 will gather some intuitive ideas of how to estimate an operator induced equation from data, and then show how this exactly transfers to the model selection problem if one applies the correspondence described above. In Section 4 an explicit example will demonstrate how a parametric differential equation leads to a parametric kernel, and how its parameters can be determined using machine learning techniques. We will also show how multiple datasets can be combined intuitively into a single estimate of the necessary parameters. Section 5 will summarize the findings and point at some extensions of our method.

2. THE REPRESENTER THEOREM REVISITED

The duality between (3) and (2) is a well-known tool in the machine learning community (Wahba, 1990; Smola et al., 1998; Schaback, 2000). We will restate the basic theorems here giving special focus on how one can construct a kernel R from a regularization operator P and vice versa.

For ease of notation we adopt the Dirac notation commonly used in physics. $|.\rangle$ denotes a vector of $L_2(\mathcal{X})$ and $\langle .|$ the corresponding dual element. Then, $\langle .|.\rangle$ stands for a dot product in the L_2-sense and for $x \in \mathcal{X}$ a delta function centred at the point x is denoted by $|x\rangle$. In a strict mathematical sense, the delta functions $|x\rangle$ are not part of $L_2(\mathcal{X})$. However, this notation introduced by Dirac has been proven to be very useful.

Theorem 1. (Representer theorem). *Let P be a linear operator with a finite-dimensional null space spanned by $\{\phi_\nu\}_{\nu=1,\ldots,M}$. Then the minimizer of (3) is a function f of the form*

$$f(x) = \sum_i \alpha_i R(x, x_i) + \sum_\nu \beta_\nu \phi_\nu(x_i),$$

with real parameters α_i, β_i. The kernel R is given as

$$R(x, y) = \langle x | (P^* P)^\dagger | y \rangle \quad (5)$$

Here, P^* is the adjoint operator of P and $(P^*P)^\dagger$ denotes the Moore-Penrose pseudo-inverse of (P^*P). Thus, $R(x, y)$ is the Green's function of the operator P^*P.

Theorem 2. (Inverse representer theorem). *Given a symmetric, bounded, positive definite*[1] *kernel R let us define an integral operator K as $(Kf)(y) = \int R(x, y) f(x) dx$. If the self-adjoint operator K is also positive definite then minimizing (3) with*

$$P = \left(K^{\frac{1}{2}}\right)^\dagger$$

yields the same result as minimizing (2).

A complete proof of the representer theorem is given in (Wahba, 1990). For the inverse, see (Schaback, 2000) or note that if we plug the P into the forward theorem, then we recover the kernel given. The correspondence is not entirely one-to-one as operators P that just differ by the sign of some eigenvalues give rise to the same kernel.

3. APPROACHES TO IDENTIFY AN OPERATOR INDUCED EQUATION

Suppose we have acquired some data and assume that the data generating function f follows an equation of form (4) for a fixed operator P.

If we assume that the data can be modelled by an exact solution f of (4), then a simple approach to find P is to compute solutions of the equation (4) for several guesses of P, and select the one that best fits the data. E.g. given some parametric structure of $P = P(\theta)$ we could attempt compute the parametric solution space of (4) and use parametric fitting methods to determine the optimal parameter set θ.

However, if the data cannot be modelled by an exact solution of (4) for any fixed operator P, the above parametric approach does not make sense. This situation may occur for example if our system at hand is intrinsically modelled correctly by the equation $Pf = 0$ but external influences disturb its behaviour from time to time. These external influences may be hard to model separately, e.g. think of a tennis ball moving freely which is hit by two players or where some wind is slightly changing its path.

To determine P, we may again propose several operators P, possibly in a parametric form $P = P(\theta)$. But now we assume that the function f is only an approximate solution to (4), i.e. the L_2 residual error is small. Of course, f should at the same time be able to approximate the given data points well. Thus, it is desirable to see whether $\|Pf\|^2 + C \text{ Loss}\{(f(x_i), y_i)\}_i$ can be effectively minimised for some f. Using the correspondence between operators and kernels introduced in the previous section, we observe that this is exactly the objective of kernel machines, and the problem of determining P corresponds directly to the model selection task of selecting the optimal kernel.

For comparing differently structured operators P or different parameter sets θ one could use the value of the objective function above. There are some model selection algorithms that work directly with it (Argyriou et al., 2005). However, if the norm of P changes, results are not directly comparable and have to be normalised. Many algorithms therefore use a different approach: For a given operator P it is easier to focus on the predictive properties of the estimated function on unseen data points. If we have estimated the correct interdependency in the data, i.e. a good structure of the operator P as well as the right parameters θ, then the prediction properties of our function are superior to other cases. In our experiments we will apply this approach using a simple cross validation scheme.

4. A DEMO EXAMPLE

We will demonstrate the above ideas in a simple example in one dimension. We assume that the given data can be modelled well by the approximate solutions of a differential equation which takes the parametric form

$$P_w f = (\nabla^2 + w^2) f = 0. \quad (6)$$

Here, we just need to determine one parameter, namely the frequency w. More complex situations can be handled in an analogous manner.

4.1 The wave kernel

The null space of the operator P_w is spanned by the plane waves, i.e. $M = 2$ and $\phi_1(x) = \sin(wx), \phi_2(x) = \cos(wx)$. $P_w^* P_w$ has the eigenfunctions $|k\rangle = e^{ik\cdot}$, thus it can be written in spectral notation as

$$P_w^* P_w = \int dk \, |k\rangle \, (-k^2 + w^2)^2 \, \langle k|$$

The kernel R then is given by

$$R_w(x, y) = \langle y | (P_w^* P_w)^\dagger | x \rangle$$

[1] i.e. for all finite sets $\{(\alpha_i, x_i)\}_i \subseteq \mathbb{R} \times \mathcal{X}$, $\sum_{i,j} \alpha_i \alpha_j R(x_i, x_j) \geq 0$.

$$= \int dk \, \langle y|k\rangle \, \frac{1}{(-k^2+w^2)^2} \, \langle k|x\rangle$$
$$= \int dk \frac{1}{(-k^2+w^2)^2} e^{ik(x-y)}$$

This is the Fourier transform of $\frac{1}{(-k^2+w^2)^2}$, which is

$$R_w(x,y) = \frac{2}{\pi w^3}(-r\cos(r)+\sin(r)) \qquad (7)$$

with r denoting $w\,|x-y|$. A plot of this kernel is depicted in Figure 1.

This derivation is similar to that of the well-known thin-plate spline. The resulting kernel R is only conditionally positive definite,[2] as the Fourier spectrum of R contains singularities and $\mathcal{X} = \mathbb{R}$ is not compact. However, the theory presented in Section 2 still applies with slight modifications (Schaback, 2000).

4.2 The data

The true distribution of the dataset is chosen to be piecewise sinusoidal: We construct a sine function of frequency $w = 10$ and amplitude 1. At random locations which occurred on average every 6 periods the phase of the sine jumps by an amount drawn from a uniform distribution over the interval $[0, 2\pi]$. This corresponds to externally enforced, new initial conditions at the jump sites.

The data points, i.e. the measurements accessible to our algorithm, were sampled from this partial sine at random locations with average density of 5 points per period yielding moderate sparsity. Gaussian acquisition noise of standard deviation $\sigma = 0.1$ was added to the measurements.

4.3 Frequency from one dataset

A simple kernel regression method using the quadratic loss function is kernel ridge regression. The expansion parameters α, β can be found by solving the following linear system (Wahba, 1990):

$$\begin{pmatrix} M & T \\ T' & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \qquad (8)$$

where $M_{i,j} = R(x_i, x_j) + \frac{m}{C}\delta_{ij}$ and $T_{i,\nu} = T'_{\nu,i} = \phi_\nu(x_i)$.

In Figure 2 a reconstruction using different methods and kernels is shown. The parametric fit using only sine and cosine functions that exactly fulfil the assumed differential equation (6) is not able to fit the true data generating function even if the true frequency w is known. Therefore one cannot expect to draw valid conclusions about the parameter w from that approach.

[2] i.e. for all finite sets $\{(\alpha_i, x_i)\}_i \subseteq \mathbb{R} \times \mathcal{X}$ with $\sum_i \alpha_i \phi_\nu(x_i) = 0 \, \forall \nu = 1, \ldots, M$, $\sum_{i,j} \alpha_i \alpha_j R(x_i, x_j) \geq 0$.

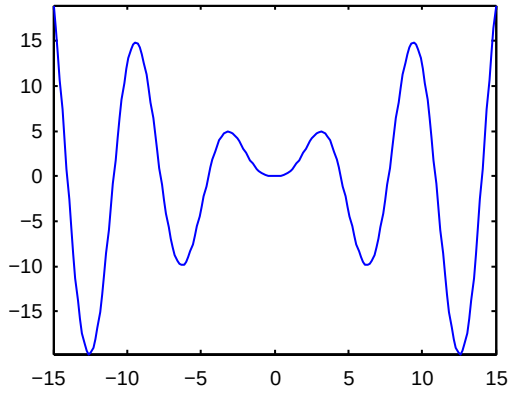

Fig. 1. The wave kernel centred at 0.

In addition, a non-parametric fit using the standard Gaussian kernel $R(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ does not yield good results. Using this kernel corresponds to assuming the differential form (Girosi et al., 1993)

$$P_\sigma = \sum_{n=0}^{\infty} (-1)^n \frac{\sigma^{2n}}{n! 8^n} \nabla^{2n}. \qquad (9)$$

The fit is especially poor in regions of low data density.

The fit using ridge regression with the wave kernel (7) and its null space, however, yields a good result: It extrapolates well into regions of sparse data points, as the correct differential model (6) is implicitly used.

As demonstrated by this initial experiment selecting a differential equation by evaluating the generalisation properties of the corresponding kernel machine seems promising.

One simple yet effective way to estimate the true generalization error from only a finite sample from the underlying data distribution is the leave-one-out (LOO) scheme: one trains on a subset of $m-1$ data points and then evaluates on the remaining one. If one iterates this procedure for all possible combinations and averages the error, one gets an almost unbiased estimator of the true expected error. Even though this method sounds very expensive to compute, there is an analytic expression of the LOO-error which requires just one training of the regression algorithm on the whole dataset, as well as some additional work of the same time complexity (Wahba, 1990). We will use this simple cross validation scheme for all our experiments below, both for computing the optimal regularization parameter C – which is determined for each proposed kernel parameter separately – as well as for comparing different kernels and their parameters.

We used the test dataset of Figure 2 and tried to determine the underlying differential equation (see Figure 3). The Gaussian kernel and its corresponding differential equation induced by (9)

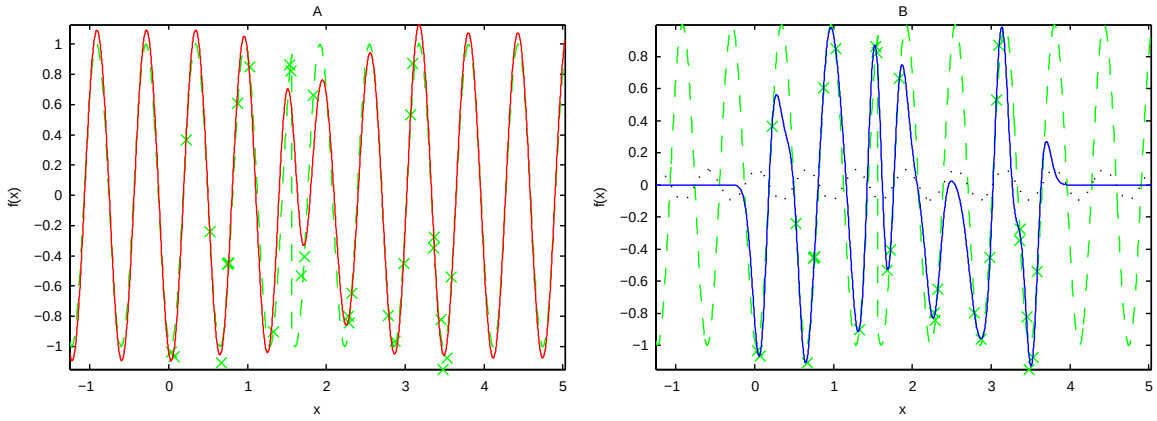

Fig. 2. In both figures, the dashed line depicts the data generating function (phase jump at $x = 1.6$), the crosses are the noisy measurements ($m = 30$). The other lines show reconstructions using different methods : **(solid,A)** ridge regression with the proposed wave kernel (7) and its null space at the correct frequency, **(solid,B)** ridge regression with Gaussian kernel and optimal parameter σ, **(dotted,B)** parametric fit of a plane wave at the correct frequency (The estimated value of the amplitude is incorrect due to the phase jump in the middle of the dataset). The regularization parameter C was always chosen to minimize the leave-one-out (LOO) error.

do not match the data well. The wave kernel at frequency $w = 10.25$ shows the best generalisation performance. The found frequency is close to the true underlying frequency of $w = 10$. Note that the algorithm used just 30 noisy points as its input.

If we increase the number of data points given to the algorithm, the frequency determination should become more exact. For different input sizes we repeatedly computed the optimal frequency (50 iterations). We used different initial conditions and different phase jump locations to produce independent datasets. In the table below, the standard deviation Δw of the results is given. The mean of the determined frequencies was always within $\frac{\Delta w}{5}$ of the true frequency $w = 10$.

m	10	50	100	200	500	1000
Δw	0.67	0.24	0.24	0.17	0.07	0.04

Thus, the uncertainty of determining the frequency decreases and a higher accuracy can be estimated reliably if more data is available.

4.4 Frequency estimation from multiple datasets

It is possible to naturally integrate additional information contained in several independent data recordings into a single identification process. This may be useful if only a short acquisition period for a signal is possible at each point in time, but we are able to observe the same system repeatedly. The data is then sampled from the same differential equation but with different initial conditions.

Given a candidate frequency w we compute the LOO-error for each data recording separately using the corresponding kernel. Then we average

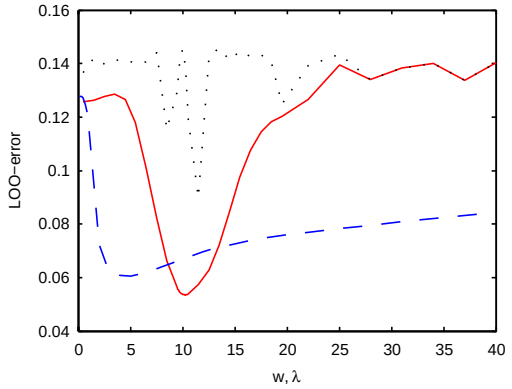

Fig. 3. The LOO-error plotted for different parameter values and different methods (in brackets the minimal attained value): **(dotted)** a parametric model using the plane waves at different frequencies w (0.09), **(dashed)** ridge regression using the Gaussian kernel with different $\lambda = \frac{1}{2\sigma^2}$ (0.06), **(solid)** ridge regression using the wave kernel and its null space at different frequencies w (0.05). The wave kernel shows the overall smallest LOO-error and the corresponding frequency $w = 10.25$ is close to the correct frequency $w = 10$,

the error of all datasets in order to evaluate the predictive qualities of a proposed differential operator/kernel. The correct differential equation is taken to be the one which minimises this criterion. We determine the regularization parameter C separately for each dataset and each kernel parameter.

To test our method we constructed N datasets ($m = 100$) with independent initial conditions as well as independent jump locations but with the same frequency from our data model. As above, we repeated the whole identification process 50

times with independent datasets and measured the standard deviation Δw of the determined frequencies. These are shown in the table below. The difference between the mean of the detected frequencies and the true frequency $w = 10$ was always within a fraction of the standard deviation Δw.

N	1	2	5	10	50
Δw	0.241	0.171	0.125	0.084	0.048

5. CONCLUSIONS

We have presented a framework to estimate linear operator induced equations – particularly differential equations – which can model given observations. The initial step of estimating a function is interlinked with the system identification part.

Using the link between model selection and operator estimation opens up a large variety of model selection methods to be applied in this system identification task. In our experiments we have just shown the most simple algorithm. Common model selection methods are often based on generalization error bounds that depend on the proposed kernel (Chapelle et al., 2002). Some of these bounds allow an analytic derivative with respect to kernel parameters rendering the problem amenable to gradient descent methods. Such approaches could speed up our example method based on the LOO-error.

In many standard cases where a dense, equally spaced data sampling is given, we do not expect our method to outperform other approaches. However, it is very flexible when dealing with special situations which may be difficult for standard algorithms. We are able work with non-uniformly sampled, sparse datasets and to integrate information from multiple, independent recordings.

Empirically, kernel methods have been shown to work well in a number of high dimensional applications. Even though we have only demonstrated our method in a one dimensional toy problem, the theory and the algorithms equally transfer to the multidimensional case where many traditional differential equation estimation techniques are problematic.

In theory it is possible to over-fit the operator parameters θ or to select a too complex structure for P if one just evaluates the empirical prediction error on a finite sample. This is particularly true when the number of free kernel parameters θ or the dimension of the null space of P increases towards the number of given data points. One way to avoid over-fitting is to include an additional regularization into the second stage optimisation over the kernel parameters (Ong and Smola, 2003). However, this would require having an explicit non-trivial prior belief regarding the correct structure and parameter range of the equation.

Although we consider the proposed approach theoretically intriguing, the experimental evaluation is still preliminary. Future work will experimentally test the method for higher dimensional problems and will explore ways to handle inhomogeneous differential equations.

ACKNOWLEDGEMENTS

For initial discussions the authors like to thank Jan Eichhorn, for proofreading Jan Eichhorn and Arthur Gretton. Matthias Hein and Alex Smola gave comments which helped us to recognise some of the problems as well as possible solutions.

REFERENCES

A. Argyriou, C.A. Micchelli, and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. In *Proc. Conf. on Learning Theory (COLT'05)*, 2005.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46 (1):131–159, 2002.

F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, MIT, 1993.

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

S. Moussaoui, D. Brie, and A. Richard. On the interpretation of a continuous-time model identification method in terms of regularization. In *Proc. SYSID 2003*, 2003.

T.G. Müller and J. Timmer. Fitting parameters in partial differential equations from partially observed noisy data. *Physica D*, 171:1–7, 2002.

C. S. Ong and A. J. Smola. Machine learning using hyperkernels. In *Proc. International Conference on Machine Learning*, 2003.

R. Schaback. A unified theory of radial basis functions. *J. of Comp. and Appl. Math.*, 121: 165–177, 2000.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

S. Sonnenburg, G. Rätsch, and C. Schäfer. Learning interpretable SVMs for biological sequence classification. *Research in Computational Molecular Biology*, pages 389–407, 2005.

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