Penalized Splines and Reproducing Kernel Methods

N. D. PEARCE and M. P. WAND

Two data analytic research areas—*penalized splines* and *reproducing kernel methods*—have become very vibrant since the mid-1990s. This article shows how the former can be embedded in the latter via theory for reproducing kernel Hilbert spaces. This connection facilitates cross-fertilization between the two bodies of research. In particular, connections between support vector machines and penalized splines are established. These allow for significant reductions in computational complexity, and easier incorporation of special structure such as additivity.

KEY WORDS: Bioinformatics; Classification; Data mining; Generalized additive models; Kernel machines; Machine learning; Mixed models; Reproducing kernel Hilbert spaces; Semiparametric regression; Statistical learning; Supervised learning; Support vector machines.

1. INTRODUCTION

In the mid-1990s two vibrant areas of data analytic research emerged. Each built on ideas that had accumulated over the previous decades, but were ignited by a few key papers and results. One of them was in the statistics literature, within the subject of nonparametric regression or "smoothing," and will be referred to here as *penalized splines*. The other was primarily in the computing science literature, within the subject of machine learning, and will be referred to here as *reproducing kernel methods*. This article elucidates the connection between these two sets of literature with the intention of promoting fruitful cross-fertilization.

The main catalyst for the flurry of penalized spline research was Eilers and Marx (1996). Another key reference is Hastie (1996), although the essential ideas have been around for much longer (e.g., Schoenberg 1969; Parker and Rice 1985; Wahba 1990, chap. 7). The thrust of this research is generalization of ordinary smoothing splines to knot sequences different from, and usually much fewer than, the observed predictor variables. Hastie (1996) and Marx and Eilers (1998) illustrated the benefits for additive models. Mixed model and Bayesian representations of penalized spline smoothers have allowed, for example, straightforward incorporation of longitudinal data into nonparametric regression (e.g., Verbyla, Cullis, Kenward, and Welham 1999). Fitting and inference can be accomplished via established software, such as `PROC MIXED` in SAS and `lme()` in R and WinBUGS, provided the number of basis functions is relatively low (e.g., Ngo and Wand 2003; Crainiceanu, Ruppert, and Wand 2005). There are now several packages in R, such as `mgcv` (Wood 2006), for fitting such models. Other developments include simpler incorporation of measurement error (Berry, Carroll, and Ruppert 2002) and geostatistical data (Kammann and Wand 2003). Much of the work on penalized splines up until about 2002 is summarized in the book by Ruppert, Wand, and Carroll (2003). There is also a large literature on nonpenalized splines for multivariate function estimation, such as Stone (1994) and Hansen and Kooperberg (2002).

The emergence of support vector machines, starting with Boser, Gyon, and Vapnik (1992), has blossomed into a huge literature since the mid-1990s, and is the main catalyst for what have become known as *reproducing kernel methods*, *kernel methods* or *kernel machines* in machine learning. The first label will be used here since it is somewhat more general, and avoids confusion with kernel smoothing methods in the nonparametric regression literature (e.g., Wand and Jones 1995). A comprehensive overview of reproducing kernel methods in machine learning research was provided by Schölkopf and Smola (2002). These authors also maintain a Web site, *www.kernel-machines.org*, that disseminates research on the topic. Other useful summaries were provided by Burges (1998), Evgeniou, Pontil, and Poggio (2000) and Cristianini and Shawe-Taylor (2000). Before the emergence of support vector machines, reproducing kernel methods were prominent in the nonparametric regression literature as a framework for smoothing spline methodology, as summarized by Wahba (1990). However, the adoption of these ideas by the machine learning community has widened the scope of reproducing kernel methods quite considerably.

This article shows how penalized splines are embedded in the class of reproducing kernel methods and helps connect these two bodies of research. It is envisaged that support vector machine and other kernel machine research has the most to gain from this connection. There the main objectives are classification and prediction; usually from large, complex, multidimensional datasets. The reduced knot aspect of penalized splines allows for significant savings in computational complexity, as we explain in Section 6. In addition, much of the support vector machine research is done within the machine learning discipline, and largely oblivious to many statistical principles such as interpretation, model building, diagnosis, low-dimensional structure, and proper accounting for data dependencies. Kernels based on penalized splines offer the opportunity to incorporate some of these principles more straightforwardly than commonly used kernels. Similar recent work has been done using the ideas of smoothing spline analysis of variance; see Lin and Zhang (2006), and Lee, Kim, Lee, and Koo (2006).

An illustration of a support vector machine classifier which uses low-dimensional structure and is immediately interpretable is given in Figure 1. It arises from use of additive model penalized spline kernels (Sections 5.3 and 6) to build a classifier for

N. D. Pearce is a Doctoral Student, and M. P. Wand is Professor, Department of Statistics, School of Mathematics, University of New South Wales, Sydney 2052, Australia (E-mail addresses: *nathan@maths.unsw.edu.au* and *wand@maths.unsw.edu.au*). We are grateful for advice from Ian Doust, Inge Koch, and John Ormerod. This article also benefited from the high-quality feedback of an associate editor and two referees. Partial support has been provided from the Australian Research Council under its Centres of Excellence programme.

 DOI: 10.1198/000313006X124541

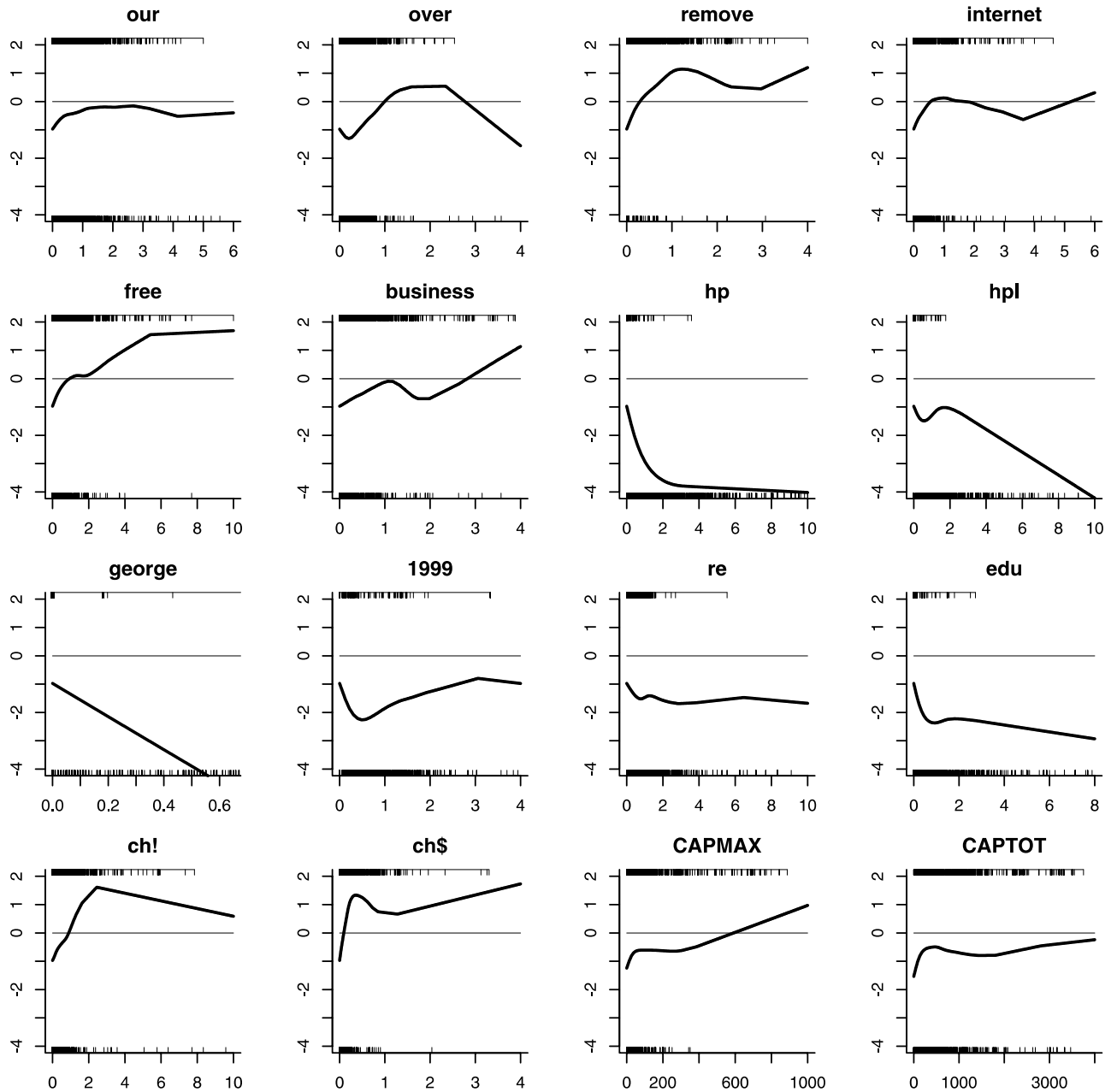

Figure 1. Visualization of a penalized spline support vector classifier for the "spam" data. Each panel shows the slice of the classifier with all other predictors set to their medians. The tick-marks show the predictor values: spam e-mail messages along the top, ordinary e-mail messages along the bottom.

the "spam" data, described by Hastie, Tibshirani, and Friedman (2001), with spam e-mail messages coded as $+1$ and ordinary messages coded as -1. Each panel shows the slice of the classification surface for the labeled predictor, with all other predictors set to their medians. It is seen, for example, that frequency of the word "free" has a monotonic effect on classification while frequency of exclamation marks (`ch!`) has a nonmonotonic effect.

The next section provides a brief description of the simplest version of penalized splines. Section 3 describes the basics of reproducing kernel methods. The link between these two concepts is laid out in Section 4. Various extensions are treated in Section 5. Section 6 is devoted to the special case of support vector machines and advantages of the penalized spline approach are explained. We close with some summary remarks in Section 7.

2. PENALIZED SPLINES

For the moment we will consider only the "scatterplot smoothing" situation where the observed data are $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $1 \leq i \leq n$, and both variables are continuous. The simplest penalized spline model is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ + \varepsilon_i. \qquad (1)$$

Here $x_+ = \max(0, x)$, $\kappa_1, \ldots, \kappa_K$ are a dense set of knots over the range of the x_i's and the ε_i are independent zero mean

random variables with common variance σ_ε^2. Typically the knots are chosen to mimic the distribution of the x_i's, such as their K-tiles (e.g., Ruppert 2002) and usually $K \ll n$. Fitting is typically performed via penalized least squares:

$$\min_{\boldsymbol{\beta},\mathbf{u}} \left[\sum_{i=1}^{n} \left\{ y_i - \beta_0 - \beta_1 x_i - \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ \right\}^2 + \lambda \sum_{k=1}^{K} u_k^2 \right], \quad (2)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and $\mathbf{u} = (u_1, \ldots, u_K)^T$. The quadratic penalty $\lambda \sum_{k=1}^{K} u_k^2$ is a simple way to restrict the influence of the spline terms $(x_i - \kappa_k)_+$ and avoid overfitting. The smoothing parameter $\lambda > 0$ controls the trade-off between overfitting and bias. A matrix formulation of (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (3)$$

where

$$\mathbf{X} = [1 \;\; x_i]_{1 \le i \le n}, \qquad \mathbf{Z} = \underset{1 \le k \le K}{[(x_i - \kappa_k)_+]_{1 \le i \le n}},$$

and $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ contain the respective subscripted variables. Then (2) becomes

$$\min_{\boldsymbol{\beta},\mathbf{u}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \lambda \|\mathbf{u}\|^2 \right), \quad (4)$$

where $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$ denotes the norm of the vector $\mathbf{v}$. The solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}, \quad \widehat{\mathbf{u}} = \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \quad (5)$$

with $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I}$. The notation of (3) suggests a linear mixed model and (5) corresponds exactly to best linear unbiased prediction if $\mathbf{u}$ is treated as a random effects vector with covariance matrix $(\sigma_\varepsilon^2/\lambda)\mathbf{I}$ (e.g., Brumback, Ruppert, and Wand 1999).

Although we use the term "penalized splines" throughout this article, it should be pointed out that there are several alternative names for what is essentially the same general approach. These include low-rank splines, P-splines, pseudosplines, and reduced knot splines.

3. REVIEW OF REPRODUCING KERNEL METHODS

This section provides a brief review of reproducing kernel methods. This facilitates the reproducing kernel representation of penalized splines in the next section.

Reproducing kernel methods are performed within the functional analytic structure known as a reproducing kernel Hilbert space (RKHS). An early RKHS reference is Aronszajn (1950) and contemporary summaries include Wahba (1999) and Evgeniou, Pontil, and Poggio (2000). Of particular relevance to penalized splines are penalizations over subspaces, based on projection operators. Relevant background material on Hilbert space projection theory may be found in, for example, Simmons (1983) and Rudin (1991).

A RKHS on $\mathbb{R}^d$ is a Hilbert space of real-valued functions that is generated by a bivariate symmetric, positive definite function $\mathcal{K}(\mathbf{s},\mathbf{t})$, $\mathbf{s},\mathbf{t} \in \mathbb{R}^d$, called the kernel. The steps for RKHS construction from $\mathcal{K}$ are:

1. Determine the eigen-decomposition of $\mathcal{K}$: $\mathcal{K}(\mathbf{s},\mathbf{t}) = \sum_{j=0}^{\infty} \lambda_j \phi_j(\mathbf{s})\phi_j(\mathbf{t})$. This series is assumed to be well-defined (e.g., uniformly convergent).

2. Define the space of real-valued functions on $\mathbb{R}^d$:

$$\mathcal{H}_\mathcal{K} = \left\{ f : f = \sum_{j=0}^{\infty} a_j \phi_j, \quad \text{such that} \quad \sum_{j=0}^{\infty} a_j^2/\lambda_j < \infty \right\}.$$

3. Endow $\mathcal{H}_\mathcal{K}$ with the inner product

$$\left\langle \sum_{j=0}^{\infty} a_j\phi_j, \sum_{j=0}^{\infty} a'_j\phi_j \right\rangle_{\mathcal{H}_\mathcal{K}} = \sum_{j=0}^{\infty} a_j a'_j/\lambda_j.$$

From this last condition is follows that the norm of $f = \sum_{j=0}^{\infty} a_j\phi_j$ in $\mathcal{H}_\mathcal{K}$ is

$$\|f\|^2_{\mathcal{H}_\mathcal{K}} = \sum_{j=0}^{\infty} a_j^2/\lambda_j.$$

The adjective "reproducing" arises from the important result

$$\langle \mathcal{K}(\mathbf{s},\cdot), \mathcal{K}(\mathbf{t},\cdot) \rangle_{\mathcal{H}_\mathcal{K}} = \mathcal{K}(\mathbf{s},\mathbf{t}).$$

This has important implications, and gives rise to the "kernel trick" that we discuss shortly.

Sufficient conditions for the kernel $\mathcal{K}$ to admit the above RKHS construction are given by Mercer's Theorem (e.g., Cristianini and Shawe-Taylor 2000). Popular kernels, particularly in machine learning contexts, include the pth degree polynomial kernel, $\mathcal{K}(\mathbf{s},\mathbf{t}) = (1 + \mathbf{s}^T\mathbf{t})^p$, and the radial basis kernel, $\mathcal{K}(\mathbf{s},\mathbf{t}) = \exp\left(-\gamma \|\mathbf{s} - \mathbf{t}\|^2\right)$, for some $\gamma > 0$. Smoothing and thin plate splines correspond to (conditional) kernels such as $\mathcal{K}(s,t) = |s-t|^3$ $(d = 1)$ and $\mathcal{K}(\mathbf{s},\mathbf{t}) = \|\mathbf{s} - \mathbf{t}\|^2 \log(\|\mathbf{s} - \mathbf{t}\|)$ $(d = 2)$ (Girosi, Jones, and Poggio 1995).

Let $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le n$ be a dataset, $\mathcal{L}(\cdot,\cdot)$ be a loss function and $\lambda > 0$ be a smoothing parameter. The fit $\widehat{f}$ within $\mathcal{H}_\mathcal{K}$, with respect to $\mathcal{L}$ and λ, is the solution to

$$\min_{f \in \mathcal{H}_\mathcal{K}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(\mathbf{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_\mathcal{K}} \right\}. \quad (6)$$

For continuous y_i, examples of loss functions are

$$\mathcal{L}(a,b) = \begin{cases} (a-b)^2 & \text{(squared error loss)} \\ (|a-b| - \varepsilon)_+ & (\varepsilon\text{-insensitive loss, for some } \varepsilon > 0). \end{cases}$$

For $y_i \in \{-1, 1\}$, as arises in two-category classification, examples are

$$\mathcal{L}(a,b) = \begin{cases} \log(1 + e^{-ab}) & \text{(Bernoulli log-likelihood)} \\ (1 - ab)_+ & \text{(hinge loss).} \end{cases}$$

The latter loss functions result in what are respectively known as support vector regression and classification, and collectively known as support vector machines. Wahba (1999) described the infinite dimensional RKHS theory for support vector machines.

Subsequent work in this area includes Lin, Lee, and Wahba (2002), and Lin, Wahba, Zhang, and Lee (2002).

The solution to (6) can be shown to be of the form $\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{c}_i \mathcal{K}(\mathbf{x}_i, \mathbf{x})$ for some $\widehat{c}_i$, $1 \le i \le n$, depending on $\mathcal{L}$ and the data, and is known as the dual form of the solution. The "kernel trick" is that we do not need to calculate the eigenfunctions, $\phi_0, \phi_1, \ldots$, in order to find the dual form. The kernel trick is a popular theme in kernel methods, and has allowed linear algorithms to be easily converted into nonlinear algorithms. The corresponding primal form of the solution is a linear combination of the eigenfunctions. When the eigenfunctions are easily calculated the primal form may offer a simpler and more intuitive form of the solution.

It is often desirable that certain functions in $\mathcal{H}_\mathcal{K}$ are unpenalized. Let $\mathcal{H}_0 = \text{span}\{\psi_0, \ldots, \psi_p\}$ be such a subspace of $\mathcal{H}_\mathcal{K}$ for which penalization is not desired. Mathematically, this means that fits over $\mathcal{H}_0$ are found by simply minimizing $\sum_{i=1}^{n} \mathcal{L}(y_i, f(\mathbf{x}_i))$. Let $\mathcal{H}_1 = \mathcal{H}_0^\perp$ be the orthogonal complement of $\mathcal{H}_0$ and P_1 denote the linear operator corresponding to projection onto $\mathcal{H}_1$. Then $\mathcal{H}_\mathcal{K} = \mathcal{H}_0 \oplus \mathcal{H}_1$ with $\mathcal{H}_0$ being the null space of P_1. With respect to the null space $\mathcal{H}_0$, smoothing parameter λ and loss function $\mathcal{L}$, we define fits $\widehat{f}$ according to

$$\min_{f \in \mathcal{H}_\mathcal{K}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(\mathbf{x}_i)) + \lambda \|P_1 f\|^2_{\mathcal{H}_\mathcal{K}} \right\}. \tag{7}$$

It can also be shown (Aronszajn 1950) that $\mathcal{H}_0$ and $\mathcal{H}_1$ are reproducing kernel Hilbert spaces in their own right, with kernels $\mathcal{K}_0$ and $\mathcal{K}_1$ such that $\mathcal{K}_0 + \mathcal{K}_1 = \mathcal{K}$.

Before closing this section, we note that only the so-called 2-norm penalty is being discussed in this article. There has been a considerable amount of research on alternative penalties such as the 1-norm penalty (e.g., Cristianini and Shawe-Taylor 2000).

4. REPRODUCING KERNEL REPRESENTATION OF PENALIZED SPLINES

We now show how penalized splines are a special case of reproducing kernel methods. In particular, penalized splines correspond to finite dimensional RKHS as covered in Part I, Section 3 of Aronszajn (1950). However, explicitly laying out the reproducing kernel representation of penalized splines with its terminology and notation is, in our view, very worthwhile. For example, it allows researchers familiar with the penalized splines literature to see how certain principles (e.g., additive modeling) can be extended to other settings such as support vector machines.

Consider the setting of Section 2 with prespecified knots $\kappa_1, \ldots, \kappa_K$. The kernel that allows penalized splines to be couched in a RKHS framework is

$$\mathcal{K}(s,t) = 1 + st + \sum_{k=1}^{K} (s - \kappa_k)_+ (t - \kappa_k)_+.$$

The eigenfunctions are, trivially,

$$\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_{k+1}(x) = (x - \kappa_k)_+, \quad 1 \le k \le K$$

with eigenvalues $\gamma_0 = \gamma_1 = \cdots = \gamma_{K+1} = 1$ ($\phi_i = 0$ for $i > K+1$). The RKHS is

$$\mathcal{H}_\mathcal{K} = \left\{ f : f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \right\}$$

with inner product

$$\left\langle \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+, \beta_0' + \beta_1' x + \sum_{k=1}^{K} u_k' (x - \kappa_k)_+ \right\rangle_{\mathcal{H}_\mathcal{K}} = \beta_0 \beta_0' + \beta_1 \beta_1' + \sum_{k=1}^{K} u_k u_k'.$$

In particular,

$$\|f\|^2_{\mathcal{H}_\mathcal{K}} = \|\boldsymbol{\beta}\|^2 + \|\mathbf{u}\|^2.$$

The penalized spline RKHS is a particularly simple Hilbert space in that it is finite-dimensional and isomorphic to $\mathbb{R}^{K+2}$. This means that projections in $\mathcal{H}_\mathcal{K}$ correspond to familiar Euclidean projections of the coefficients, as illustrated in the next paragraph.

For penalized splines the subspace of unpenalized functions is the linear component

$$\mathcal{H}_0 = \{ f : f(x) = \beta_0 + \beta_1\, x \}$$

and the orthogonal complement

$$\mathcal{H}_1 = \mathcal{H}_0^\perp = \left\{ f : f(x) = \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \right\}$$

is the spline basis function component. The projection of $f \in \mathcal{H}_\mathcal{K}$ onto $\mathcal{H}_1$ is given by

$$P_1 \left(\beta_0 + \beta_1\, x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \right) = \sum_{k=1}^{K} u_k (x - \kappa_k)_+$$

and, hence, $\|P_1 f\|^2_{\mathcal{H}_\mathcal{K}} = \|\mathbf{u}\|^2$. Therefore (7) is equivalent to (4) for squared error loss. For more general loss, (7) reduces to

$$\min_{\boldsymbol{\beta}, \mathbf{u}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i) + \lambda \|\mathbf{u}\|^2 \right\}. \tag{8}$$

Define $\mathbf{X}_x = [1\ x]$ and $\mathbf{Z}_x = [(x - \kappa_1)_+ \ldots (x - \kappa_K)_+]$ and let $\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}}$ denote the solution to (8). Then the primal form of the solution is

$$\widehat{f}(x) = \mathbf{X}_x \widehat{\boldsymbol{\beta}} + \mathbf{Z}_x \widehat{\mathbf{u}} = \widehat{\beta}_0 + \widehat{\beta}_1 x + \sum_{k=1}^{K} \widehat{u}_k (x - \kappa_k)_+.$$

The dual form is

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \sum_{i=1}^{n} \widehat{c}_i \mathcal{K}_1(x, x_i)$$

for suitable $\widehat{c}_i$, $1 \le i \le n$ and where $\mathcal{K}_1(s,t) = \mathbf{Z}_s \mathbf{Z}_t^T$ is the kernel for $\mathcal{H}_1$. As an example consider squared error loss, $\mathcal{L}(a,b) = (a-b)^2$. Then $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ are given by (5) while the $\widehat{c}_i$ are the entries of

$$\widehat{\mathbf{c}} = \mathbf{K}_1 (\mathbf{K}_1 + \lambda \mathbf{I})^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

where $\mathbf{K}_1 = \mathbf{Z}\mathbf{Z}^T = [\mathcal{K}_1(x_i, x_j)]_{1\le i,j\le n}$, known as the *Gram matrix* of $\mathcal{K}_1$.

5. EXTENSIONS

Sections 2 and 4 only considered penalized splines for scalar predictors and truncated line basis functions. However, as shown in this section, the reproducing kernel representations apply for general penalized spline models such as those involving other spline basis functions, higher dimensional predictors and additive structure.

5.1 Other Spline Basis Functions

For $x \in \mathbb{R}$, general penalized spline models can be written as

$$f(x) = \mathbf{X}_x\boldsymbol{\beta} + \mathbf{Z}_x\mathbf{u}, \tag{9}$$

where $\mathbf{X}_x = [1\ x\ \ldots\ x^p]$ for some $p \ge 0$ and $\mathbf{Z}_x$ is a set of spline basis functions. Without loss of generality, the penalty on $\mathbf{u}$ can be taken to be $\|\mathbf{u}\|^2$ by appropriate transformation of the functions in $\mathbf{Z}_x$. Beyond the truncated line model (2.) the simplest basis is

$$\mathbf{Z}_x = \underset{1\le k\le K}{[(x-\kappa_k)_+^p]},$$

corresponding to truncated polynomials of degree p. For numerical stability reasons, it is usually advantageous to linearly transform the truncated polynomial basis functions to, say, B-spline basis functions (e.g., Eilers and Marx 1996). A suitable adjustment needs to be made to the penalization component.

Another family of bases is that corresponding to thin plate splines (French, Kammann, and Wand 2001) and takes the form $\mathbf{X}_x = [1\ x\ \ldots\ x^{m-1}]$ and

$$\mathbf{Z}_x = \underset{1\le k\le K}{[|x-\kappa_k|^{2m-1}]}\boldsymbol{\Omega}^{-1/2}, \quad \boldsymbol{\Omega} = \underset{1\le k,k'\le K}{[|\kappa_k-\kappa_{k'}|^{2m-1}]}.$$

These have an advantage of simple extension to higher dimensional x (Section 5.2).

At this level of generality, the appropriate kernel is

$$\mathcal{K}(s,t) = \mathbf{X}_s\mathbf{X}_t^T + \mathbf{Z}_s\mathbf{Z}_t^T,$$

and the RKHS representation of (9) ensues.

5.2 Higher Dimensional Predictors

There are a number of ways by which spline basis functions can be extended to accommodate higher dimensional predictors. For example, an extension of the thin plate spline bases for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is

$$f(\mathbf{x}) = \mathbf{X}_\mathbf{x}\boldsymbol{\beta} + \mathbf{Z}_\mathbf{x}\mathbf{u}$$

where the columns of $\mathbf{X}_\mathbf{x}$ consist of all d-dimensional polynomials in $x_1, \ldots, x_d$ with degree less than m and

$$\mathbf{Z}_\mathbf{x} = \underset{1\le k\le K}{[r_{md}(\|\mathbf{x}-\kappa_k\|)]}\boldsymbol{\Omega}^{-1/2}, \quad \boldsymbol{\Omega} = \underset{1\le k,k'\le K}{[r_{md}(\|\kappa_k-\kappa_{k'}\|)]}$$

with

$$r_{md}(x) = \begin{cases} x^{2m-d} & , \quad d \text{ odd} \\ x^{2m-d}\log(x) & , \quad d \text{ even} \end{cases}$$

(e.g., Green and Silverman 1994). For $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ the appropriate kernel is

$$\mathcal{K}(\mathbf{s},\mathbf{t}) = \mathbf{X}_\mathbf{s}\mathbf{X}_\mathbf{t}^T + \mathbf{Z}_\mathbf{s}\mathbf{Z}_\mathbf{t}^T.$$

5.3 Additive Models

For two predictors x_1 and x_2 the linear penalized spline model is of the form

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i,$$

where

$$f(x_1,x_2) = \beta_0 + \beta_1 x_1 + \sum_{k=1}^{K_1} u_{1k}(x_1-\kappa_{1k})_+ + \beta_2 x_2 + \sum_{k=1}^{K_2} u_{2k}(x_2-\kappa_{2k})_+ \tag{10}$$

and, for $j = 1, 2$, the κ_{jk} denote a set of K_j knots for variable x_j. The fitting criterion is

$$\min_{\boldsymbol{\beta},\mathbf{u}} \left[\sum_{i=1}^n \{y_i - f(x_{1i},x_{2i})\}^2 + \lambda_1\sum_{k=1}^{K_1} u_{1k}^2 + \lambda_2\sum_{k=1}^{K_2} u_{2k}^2\right], \tag{11}$$

where λ_1 and λ_2 are, respectively, smoothing parameters for variables x_1 and x_2 and $\boldsymbol{\beta}$ and $\mathbf{u}$ contain the respective subscripted variables.

Let $\mathcal{H}_0$, $\mathcal{H}_1$, and $\mathcal{H}_2$ denote, respectively, the reproducing kernel Hilbert spaces generated by the kernels

$$\begin{aligned} \mathcal{K}_0(\mathbf{s},\mathbf{t}) &= 1 + \mathbf{s}^T\mathbf{t}, \\ \mathcal{K}_1(\mathbf{s},\mathbf{t}) &= \sum_{k=1}^{K_1}(s_1-\kappa_{1k})_+(t_1-\kappa_{1k})_+, \end{aligned}$$

and

$$\mathcal{K}_2(\mathbf{s},\mathbf{t}) = \sum_{k=1}^{K_2}(s_2-\kappa_{2k})_+(t_2-\kappa_{2k})_+,$$

where $\mathbf{s} = [s_1\ s_2]^T$ and $\mathbf{t} = [t_1\ t_2]^T$. Then

$$\mathcal{H}_\mathcal{K} = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \mathcal{H}_2$$

is the RKHS generated by $\mathcal{K} = \mathcal{K}_0 + \mathcal{K}_1 + \mathcal{K}_2$, where $\mathcal{H}_0$, $\mathcal{H}_1$ and $\mathcal{H}_2$ are mutually orthogonal subspaces of $\mathcal{H}_\mathcal{K}$. For $f \in \mathcal{H}_\mathcal{K}$ let $P_1 f$ denote the projection of f onto $\mathcal{H}_1$. Then, using the notation of (10),

$$P_1 f(x_1,x_2) = \sum_{k=1}^{K_1} u_{1k}(x_1-\kappa_{1k})_+$$

and

$$\|P_1 f\|_{\mathcal{H}_\mathcal{K}}^2 = \sum_{k=1}^{K_1} u_{1k}^2.$$

The projection operator P_2 is defined analogously and (10) may be written as

$$\min_{f\in\mathcal{H}_\mathcal{K}} \left[\sum_{i=1}^n \{y_i - f(x_{1i},x_{2i})\}^2 + \lambda_1\|P_1 f\|_{\mathcal{H}_\mathcal{K}}^2 + \lambda_2\|P_2 f\|_{\mathcal{H}_\mathcal{K}}^2\right].$$

For general loss functions the criterion is

$$\min_{f\in\mathcal{H}_\mathcal{K}}\left[\sum_{i=1}^{n}\mathcal{L}(y_i,f(x_{1i},x_{2i}))+\lambda_1\|P_1f\|^2_{\mathcal{H}_\mathcal{K}}+\lambda_2\|P_2f\|^2_{\mathcal{H}_\mathcal{K}}\right].$$

The extension to other basis functions, several predictors as well as to higher dimensional components (e.g., Kammann and Wand 2003) is straightforward.

5.4 Semiparametric Regression Models

General semiparametric regression models contain both smooth functional (nonparametric) and ordinary linear (parametric) components. The simplest is

$$y_i=\beta_0+\beta_1\,x_{1i}+f(x_{2i})+\varepsilon_i$$

which is often referred to as a partially linear model. If f has representation (9), then the appropriate kernel is $\mathcal{K}=\mathcal{K}_0+\mathcal{K}_1$ where

$$\begin{aligned}\mathcal{K}_0(\mathbf{s},\mathbf{t}) &= 1+s_1t_1+s_2t_2,\\ \mathcal{K}_1(\mathbf{s},\mathbf{t}) &= \sum_{k=1}^{K_2}(s_2-\kappa_{2k})_+(t_2-\kappa_{2k})_+,\end{aligned}$$

$\mathbf{s}=[s_1\ s_2]^T$ and $\mathbf{t}=[t_1\ t_2]^T$. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be the reproducing kernel Hilbert spaces generated by $\mathcal{K}_0$ and $\mathcal{K}_1$, respectively. Then $\mathcal{H}_\mathcal{K}=\mathcal{H}_0\oplus\mathcal{H}_1$ and the problem takes the same form as (7) with null space $\mathcal{H}_0$.

6. SUPPORT VECTOR MACHINE CLASSIFICATION

As mentioned in Section 1, most support vector machine classification research is within the discipline of machine learning. This section shows how support vector machine classification arises as a special case of penalized splines, and therefore allow for the incorporation of the constructions described in the previous section. In addition, we discuss how low-rank kernels of penalized splines offer significant computational savings. We will focus on the situation where the sample size n is much larger than the dimension of the predictors d. The reverse situation, sometimes called high dimension/low sample size, has been the subject of a great deal of attention in the recent literature; especially due to the advent of microarray gene expression data (e.g., Dudoit, Fridlyand, and Speed 2002). Penalized splines are more advantageous for the classical $n\gg d$ situation since, as discussed later in this section, they have low-rank kernels which afford faster computation for large n.

Two-category support vector machine classification corresponds to setting the loss function to be hinge loss: $\mathcal{L}(a,b)=(1-ab)_+$. However, the nonsmoothness of $\mathcal{L}$ in this case means that fitting is different from that of squared error and likelihood-based losses so some details are in order. Consider the generalization of the two-component additive model described in Section 5.3 corresponding to $\mathbf{x}_i\in\mathbb{R}^d$:

$$f(\mathbf{x}_i)=(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i=\left(\mathbf{X}\boldsymbol{\beta}+\sum_{\ell=1}^{L}\mathbf{Z}_\ell\mathbf{u}_\ell\right)_i \tag{12}$$

for design matrices $\mathbf{X},\mathbf{Z}=[\mathbf{Z}_1,\ldots,\mathbf{Z}_L]$, where each sub-vector $\mathbf{u}_\ell$ has its own smoothing parameter. The criterion to minimize is then

$$\sum_{i=1}^{n}(1-y_if(\mathbf{x}_i))_+ +\sum_{\ell=1}^{L}\lambda_\ell\|\mathbf{u}_\ell\|^2, \tag{13}$$

where $y_i\in\{-1,1\}$ codes the two categories. Note that (10) and (10) correspond to the situation where $d=L=2$,

$$\mathbf{X} = [1\ x_{1i}\ x_{2i}]_{1\le i\le n}$$

and

$$\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]=[\underset{1\le k\le K_1}{(x_{1i}-\kappa_{1k})_+}\ \underset{1\le k\le K_2}{(x_{2i}-\kappa_{2k})_+}]_{1\le i\le n}.$$

Although this example involves two univariate smooths, it should be noted that higher-dimensional smooths can also be accommodated by (12) and (13) (e.g., Kammann and Wand 2003).

Unlike least squares loss and Bernoulli log-likelihood loss, hinge loss is usually handled via Lagrangian optimization methods. A summary was provided by Cristianini and Shawe-Taylor (2000, chap. 5). See also Hastie et al. (2001, secs. 12.2 and 12.3). Minimization of (13) is equivalent to the constrained optimization problem

$$\min_{\boldsymbol{\beta},\mathbf{u}}\left(\sum_{\ell=1}^{L}\lambda_\ell\|\mathbf{u}_\ell\|^2+\sum_{i=1}^{n}\xi_i\right)$$

subject to

$$\xi_i\ge 0,\ y_i(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i\ge 1-\xi_i\quad\text{for all}\quad 1\le i\le n.$$

The Lagrangian primal function is

$$\begin{aligned}L_P=&\sum_{\ell=1}^{L}\lambda_\ell\|\mathbf{u}_\ell\|^2+\sum_{i=1}^{n}\xi_i\\ &-\sum_{i=1}^{n}\alpha_i\{y_i(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i-(1-\xi_i)\}-\sum_{i=1}^{n}\tau_i\xi_i,\end{aligned} \tag{14}$$

where $\alpha_i,\tau_i\ge 0$ for all $1\le i\le n$. Setting the derivatives of L_P with respect to $\boldsymbol{\beta}$, $\mathbf{u}_\ell$ and ξ_i to zero results in the equalities

$$\begin{aligned}&\mathbf{X}^T(\boldsymbol{\alpha}\odot\mathbf{y})=\mathbf{0};\ \mathbf{u}_\ell=(2\lambda_\ell)^{-1}\mathbf{Z}_\ell^T(\boldsymbol{\alpha}\odot\mathbf{y}),\\ &\qquad 1\le\ell\le L;\ \text{and}\ \tau_i=1-\alpha_i\quad 1\le i\le n,\end{aligned}$$

where here, and subsequently, $\mathbf{A}\odot\mathbf{B}$ denotes the element-wise product of equal-sized matrices (i.e., same number of rows and columns) $\mathbf{A}$ and $\mathbf{B}$. Substitution into (14) leads to the Lagrangian dual function

$$L_D = \mathbf{1}^T\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{D}\boldsymbol{\alpha},$$

where

$$\mathbf{D} = \frac{1}{2}(\mathbf{y}\mathbf{y}^T)\odot(\mathbf{Z}\boldsymbol{\Lambda}^{-1}\mathbf{Z}^T), \tag{15}$$

and $\boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1\mathbf{1},\ldots,\lambda_L\mathbf{1})$. The fitted $\widehat{\alpha}_i$ values are then found by solving the quadratic programming problem

$$\min_{\boldsymbol{\alpha}}(-\mathbf{1}^T\boldsymbol{\alpha}+\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{D}\boldsymbol{\alpha}) \tag{16}$$

subject to $0\le\alpha_i\le 1$, for all $1\le i\le n$, and $\mathbf{X}^T(\boldsymbol{\alpha}\odot\mathbf{y})=\mathbf{0}$. The Karush-Kuhn-Tucker constraints include

$$\alpha_i[y_i(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i-(1-\xi_i)]=0,\ \tau_i\xi_i=0,$$

and

$$y_i(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i - (1 - \xi_i) \geq 0 \tag{17}$$

for all $1 \leq i \leq n$.

Let $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\xi}}$ be the solution to (16) and (17). The fitted $\mathbf{u}$ is then

$$\widehat{\mathbf{u}} = \frac{1}{2}\boldsymbol{\Lambda}^{-1}\mathbf{Z}^T(\widehat{\boldsymbol{\alpha}} \odot \mathbf{y}).$$

A fitted value for $\boldsymbol{\beta}$ needs to be determined from the points in the set $\{\mathbf{x}_i : 0 < \widehat{\alpha}_i < 1\}$; the *support vectors* for which $\widehat{\xi}_i = 0$. Let $\mathcal{S}$ be set of $1 \leq i \leq n$ corresponding to these points. From the first Karush-Kuhn-Tucker constraint we obtain the set of equations:

$$(\mathbf{X}\beta)_i = (1/y_i) - (\mathbf{Z}\mathbf{u})_i = (\mathbf{y} - \mathbf{Z}\mathbf{u})_i, \quad i \in \mathcal{S}. \tag{18}$$

(the last equality follows from $y_i \in \{-1, 1\}$). If q is the length of $\boldsymbol{\beta}$ and m is the cardinality of $\mathcal{S}$ then (18) represents a system of q unknowns with m linear equations. Most of the support vector machine literature only treats the case $q = 1$, corresponding to an unpenalized intercept. In this case, Cristianini and Shawe-Taylor (2000) solved for $\boldsymbol{\beta} = \beta_0$ using an arbitrary margin point while Hastie et al. (2001) recommended averaging all m solutions. For general q our current recommendation for obtaining $\widehat{\boldsymbol{\beta}}$ is to take the median values from the $\binom{m}{q}$ possible solutions, assuming this number is not too large. Otherwise, take the median from a random sample of the $\binom{m}{q}$ possible solutions. Some safeguards are necessary to avoid degenerate linear systems.

The bulk of the computation is concerned with the solution of (16). For penalized splines kernels (15) shows that the Gram matrix admits the factorization

$$\frac{1}{2}\mathbf{Z}\boldsymbol{\Lambda}^{-1}\mathbf{Z}^T = \{\mathbf{Z}(2\boldsymbol{\Lambda})^{-1/2}\}\{\mathbf{Z}(2\boldsymbol{\Lambda})^{-1/2}\}^T$$

and thus has rank corresponding to the number of columns in $\mathbf{Z}$. Fine and Scheinberg (2001) described interior point algorithms that take advantage of such low-rank kernels. See Ormerod and Wand (2006) for R implementation. The algorithms involve $O(nK^2)$ operations, where K is the rank of the Gram matrix and corresponds to the number of columns in $\mathbf{Z}$ for penalized splines. Interior point algorithms with full-rank kernels involve $O(n^3)$ operations. Decomposition algorithms—such as Platt (1998) and Joachims (1998)—allow reductions to $O(n^2)$. In the $n \gg K$ situation low-rank kernels offer considerable savings. For example, Figure 3 of Fine and Scheinberg (2001) illustrates a more than 20-fold improvement in computational time over the algorithm of Platt (1998).

Penalized splines are one of several ways to obtain a low-rank kernel. Another general approach is to apply a thinning strategy to the basis functions of a full-rank kernel. In the context of support vector classification, such strategies have been studied by, for example, Smola and Schölkopf (2000) and Williams and Seeger (2001). The relative advantages of these approaches is the topic of ongoing research.

Table 1. Mean (standard error of the mean) Misclassification Rates Over 50 Simulations for the "Skin of the Orange" Example. Classifiers 1–6 are described in Section 12.3.4 of Hastie et al. (2001). Classifier 7 is a support vector classifier with additive penalized spline kernel as described in Section 6.

	Classifier	No noise features (4 dimensions)	Six noise features (10 dimensions)
1	SVC/orig.	0.450 (0.003)	0.472 (0.003)
2	SVC/poly. 2	0.078 (0.003)	0.152 (0.004)
3	SVC/poly. 5	0.180 (0.004)	0.370 (0.004)
4	SVC/poly. 10	0.230 (0.003)	0.434 (0.002)
5	BRUTO	0.084 (0.003)	0.090 (0.003)
6	MARS	0.156 (0.004)	0.173 (0.005)
7	**SVC/add. pen. spline**	**0.095 (0.004)**	**0.123 (0.003)**
	Bayes error	0.029	0.029

6.1 "Skin of the Orange" Example

We tested additive penalized spline support vector classifiers on the "skin of the orange" simulation settings described by Hastie et al. (2001, sec. 12.3.4). Table 1 is mostly a reproduction of their Table 12.2 but with addition of classifier 7—and lists the mean misclassification rates from the simulation study (along with standard errors). Descriptions of classifiers 1–6 are given there and classifier 7, based on ideas in the current article, is described in the next paragraph. At the time of writing, data from the Hastie et al. (2001) simulation study were available on the Internet and classifier 7 was applied to those data, making the results directly comparable. Note that the Bayes error for each setting is 0.029 and represents a lower bound on the expected misclassification rate.

Classifier 7 involved the 4- and 10-dimensional extension of the truncated line additive model (10) with 20 knots in each direction. A relatively simplistic rule was used for choice of the smoothing parameters in the additive penalized spline classifier. We roughly mimicked the "4 degrees of freedom per smooth function" default used in the R and S-PLUS function `gam()` (Chambers and Hastie 1992; Hastie 2005). For hinge loss the usual degrees of freedom definitions for penalized spline additive models (e.g., Ruppert et al. 2003, sec. 11.4) are not immediate due to its nondifferentiability. We got around this by using the Bernoulli log-likelihood loss as a rough approximation.

Table 1 shows that this "rough-and-ready" additive penalized spline support vector classifier performs quite well compared with the classifiers from the original study. Classifier 5 (BRUTO) performs better than classifier 7 in both settings, but uses much more sophisticated smoothing parameter and variable selection strategies. Classifier 2 performs better than classifier 7 when there are no additional noise features, but the two-degree polynomial kernel is ideal for the spherical Bayes classification boundary of this setting. It should also be mentioned that classifiers 1–4 had their smoothing parameters chosen for optimal performance using the test data; while classifiers 5–7 used data-driven rules for smoothing parameter selection, and possibly variable selection, using only the training data.

7. DISCUSSION

The connection between penalized splines and reproducing kernel methods has the potential to be extremely fruitful. As is made clear in Section 6, support vector machines, which are not seriously hindered by large sample sizes, are a major payoff from this connection. It is also anticipated that many features of semiparametric regression including variable selection, smoothing parameter selection, interpretability, robustness, and

low-dimensional structure will prove to be beneficial in data mining and machine learning applications. The simple structure of penalized splines will aid research in this direction.

[Received September 2005. Revised June 2006.]

REFERENCES

Aronszajn, N. (1950), "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, 68, 337–404.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160–169.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ed. D. Haussler, New York: ACM Press, pp. 144–152.

Brumback, B. A., Ruppert, D., and Wand, M. P. (1999), Comment on Shively, Kohn, and Wood, *Journal of the American Statistical Association*, 94, 794–797.

Burges, C. (1998), "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, 121–167.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, New York: Chapman & Hall.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), "Bayesian Analysis for Penalized Spline Regression using WinBUGS," *Journal of Statistical Software*, 14, 14.

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge: Cambridge University Press.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, 97, 77–87.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties" (with discussion), *Statistical Science*, 11, 89–121.

Evgeniou, T., Pontil, M., and Poggio, T. (2000), "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, 13, 1–50.

Fine, S., and Scheinberg, K. (2001), "Efficient SVM Training Using Low-Rank Kernel Representations," *Journal of Machine Learning Research*, 2, 243–264.

French, J. L., Kammann, E. E., and Wand, M. R (2001), Comment on Ke and Wang, *Journal of the American Statistical Association*, 96, 1285–1288.

Girosi, R., Jones, M., and Poggio, T. (1995), "Regularization Theory and Neural Networks Architectures," *Neural Computation*, 7, 219–269.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Hansen, M. H., and Kooperberg, C. (2002), "Spline Adaptation in Extended Linear Models" (with discussion), *Statistical Science*, 17, 2–51.

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society*, Series B, 58, 379–396.

——— (2005), `gam 0.94`, R package, available online is *cran.r-project.org*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Joachims, T. (1998), "Making Large-Scale SVM Learning Practical," in *Advances in Kernel Methods—Support Vector Learning*, eds. B. Scholkopf, C. Burges, and A. Smola, Cambridge, MA: MIT Press.

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Applied Statistics*, 52, 1–18.

Lee, Y., Kim, Y., Lee, S., and Koo, J.-Y. (2006), "Structured Multicategory Support Vector Machines with ANOVA Decomposition," *Biometrika*, to appear.

Lin, Y., Lee, Y., and Wahba, G. (2002), "Support Vector Machines for Classification in Nonstandard Situations," *Machine Learning*, 46, 191–202.

Lin, Y., Wahba, G., Zhang, H., and Lee, Y. (2002), "Statistical Properties and Adaptive Tuning of Support Vector Machines," *Machine Learning*, 48, 115–136.

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *The Annals of Statistics*, 34, 5.

Marx, B. D., and Eilers, P. H. C. (1998), "Direct Generalized Additive Modeling With Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Ngo, L., and Wand, M. P. (2004), "Smoothing With Mixed Model Software," *Journal of Statistical Software*, 9, 1.

Ormerod, J. T., and Wand, M. P. (2006), `LowRankQP 1.0`, R package, available online at *cran.r-project.org*.

Parker, R. L., and Rice, J. A. (1985), Discussion of "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting" by B. W. Silverman, *Journal of the Royal Statistical Society*, Series B, 47, 40–42.

Platt, J. (1998), "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," in *Advances in Kernel Methods—Support Vector Learning*, eds. B. Scholkopf, C. Burges, and A. Smola, Cambridge, MA: MIT Press.

Rudin, W. (1991), *Functional Analysis* (2nd ed.), New York: McGraw Hill.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Schoenberg, I. (1969), "Monosplines and Quadrature Formulae," in *Theory and Application of Spline Functions*, ed. T. Greville, Madison, WI: University of Wisconsin Press.

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

Simmons, G.F. (1983), *Introduction to Topology and Modern Analysis*, Melbourne: Krieger.

Smola, A. J., and Schölkopf, B. (2000), "Sparse Greedy Matrix Approximation for Machine Learning," in *Proceedings of the 17th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann.

Stone, C. J. (1994), "The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation" (with discussion), *The Annals of Statistics*, 22, 118–184.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999), "The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines" (with discussion), *Applied Statistics*, 48, 269–311.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

——— (1999), "Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV," in *Advances in Kernel Methods: Support Vector Learning*, eds. B. Scholkopf, C. Burges, and A. Smola, Cambridge, MA: MIT Press, pp. 69–88.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Williams, C.K.I, and Seeger, M. (2001), "Using the Nyström Method to Speed up Kernel Machines," in *Advances in Neural Information Processing Systems* (vol. 13), eds. T. K. Leen and T. G. Diettrich, Cambridge, MA: MIT Press, pp. 682–688.

Wood, S. N. (2006), `mgcv 1.3`, R package, available online at *cran.r-project.org*.