Representer Theorem

By *Grace Wahba and Yuedong Wang*

Abstract

The representer theorem plays an outsized role in a large class of learning problems. It provides a means to reduce infinite dimensional optimization problems to tractable finite dimensional ones. This article reviews the representer theorem for various learning problems under the reproducing kernel Hilbert spaces framework. We present solutions to the penalized least squares and penalized likelihood for nonparametric regression, and support vector machines for classification as a solution to the penalized hinge loss. We discuss extensions of the representer theorem for regression with functional data.

KEY WORDS: classification, functional data, nonparametric regression, penalized least squares, penalized likelihood, regularization, reproducing kernel Hilbert space, smoothing spline ANOVA, support vector machines

1 What Is A Representer Theorem

Briefly, a representer theorem tells us that the solutions to some regularization functionals in high or infinite dimensional spaces lie in finite dimensional subspaces spanned by the representers of the data. It effectively reduces the computationally cumbersome or infeasible problems in high or infinite dimensional spaces to optimization problems on the scalar coefficients. This neat and striking result first appeared in the work of Kimeldorf and Wahba

[1, 2]. The widespread applications of the representer theorem started much later in the late 1980s with the explosion in large complex data and computational power.

Suppose our task is to learn a function f in a model space based on data. Learning in high or infinite dimensional spaces is usually ill-posed and regularization is commonly used to overcome this problem. A regularization functional is a map from the model space to real line with two components:

$$C(f|\text{data}) + \lambda J(f) \tag{1}$$

where $C(f|\text{data})$ is a cost function measuring goodness-of-fit to data and $J(f)$ is a penalty to prevent overfitting as well as to incorporate prior knowledge such as smoothness of the function. The tuning parameter λ balances the trade-off between two conflicting components in (1). In their original work, Kimeldorf and Wahba considered the Sobolev space as the model space, least squares as the cost, and $J(f) = \int_a^b (Lf)^2$ as the penalty where L is a linear differential operator [1, 2]. Much research has been devoted to extending one or both of these two components for different purposes [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Necessary and sufficient conditions for the representation theorem to hold have also been studied [8, 13, 14]. We limit the scope of our review to regularization problems in **Reproducing Kernel Hilbert Spaces (RKHS)** for function estimation and classification. The purpose is to illustrate the usefulness of the representer theorem. We do not intend to cover all extensions of the representer theorem which is next to impossible.

We first provide a brief review of the RKHS in Section 2. Sections 3 and 4 present the representer theorem for nonparametric regression and classification. Section 5 presents various extensions of the representer theorem for regression with functional data.

2 What is an RKHS?

In this section we provide a brief review of the RKHS. Details about the RKHS can be found in Aronszajn [15] and Wahba [4]. Readers familiar with RKHS may skip this section. Formally, an RKHS is a Hilbert space of functions on some domain $\mathcal{T}$ in which all the evaluation functionals are bounded linear functionals. This means that, by the Riesz representation theorem, for every $t \in \mathcal{T}$ there exists a representer δ_t in the space such that for any g in the space, $\langle g, \delta_t \rangle = g(t)$, where $\langle \cdot, \cdot \rangle$ is the inner product in the space. Let $K_t \equiv \delta_t$. The bivariate function $K(s,t) = \delta_t(s)$ is called the reproducing kernel (RK) of the RKHS since it has the reproducing property: $\langle K_s, f \rangle = f(s)$ and $\langle K_s, K_t \rangle = K(s,t)$.

The RK $K(s,t)$ is symmetric and positive definite: $K(s,t) = K(t,s)$, and for any $t_1, \cdots, t_r \in \mathcal{T}$ and $a_1, \cdots, a_r \in \mathbb{R}$,

$$\sum_{i,j=1,\cdots,r} a_i a_j K(t_i, t_j) \geq 0. \tag{2}$$

Every RKHS has a unique RK that is positive definite. Conversely, the Moore-Aronszajn theorem [16] states that for every positive definite function, $K(\cdot,\cdot)$, there exists a unique RKHS with K as its RK. In practice we usually assume that the RK is known. Therefore, the representer of any bounded linear functional can be obtained explicitly in terms of the RK.

We note that there is no assumption on the nature of $\mathcal{T}$ which allows us to deal with functions defined on different domains in a unified manner. In practice the domain may be an interval of the real line, d-dimensional Euclidean space, a circle, or a sphere [4, 17, 18].

3 Nonparametric Regression

3.1 Penalized least squares

Denote $\{(x_i, y_i), \ i = 1, \cdots, n\}$ as n observations of a covariate $x \in \mathcal{X}$ and a response $y \in \mathbb{R}$. A nonparametric regression model assumes that

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3}$$

where ϵ_i are iid random errors with mean zero. The goal of regression analysis is to model and estimate the function f. We assume that $f \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS of functions from $\mathcal{X}$ to $\mathbb{R}$. Since the space $\mathcal{H}$ is usually infinite dimensional, certain regularization is necessary for estimation. We estimate f as the solution to the penalized least squares

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda J(f) \right\} \tag{4}$$

where the first part measures the goodness-of-fit, and $J(f)$ is a square seminorm penalty [17]. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0 = \{f : \ J(f) = 0\}$ is a finite dimensional space with basis functions $\phi_1, \cdots, \phi_m$ and $\mathcal{H}_1$ is an RKHS with RK denoted as K_1. Let $\xi_i(\cdot) = K_1(x_i, \cdot)$ for $i = 1, \cdots, n$ be representers.

Theorem 1. *(Representer Theorem) The solution to (4), $\hat{f}$, is a linear combination of the basis functions $\phi_1, \cdots, \phi_m$ and representers $\xi_1, \cdots, \xi_n$:*

$$\hat{f} = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{n} c_j \xi_j. \tag{5}$$

[Proof] Any $f \in \mathcal{H}$ can be expressed as $f = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{n} c_j \xi_j + \rho$ where $\rho \in \mathcal{H}$ is orthogonal to the space spanned by $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$. Then the penalized least squares (4) reduces to

$$\sum_{i=1}^{n} \left(y_i - \langle K_{x_i}, \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{n} c_j \xi_j + \rho \rangle \right)^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K_1(x_i, x_j) + \lambda \|\rho\|^2 \tag{6}$$

where $K_x(\cdot) = K(x, \cdot)$ and $K(x,z) = \sum_{\nu=1}^{m} \phi_\nu(x)\phi_\nu(z) + K_1(x,z)$ is the RK of $\mathcal{H}$. Note that K_{x_i} belongs to the subspace spanned by $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$. Therefore, ρ drops out the first term in (6) since it is orthogonal to K_{x_i}. Consequently $\rho = 0$ and the conclusion follows.

Remarks

1. The representer theorem was first derived by Kimeldorf and Wahba [1, 2] in the setting of Chebyshev splines. The results for general RKHS first appeared in [4].

2. The significance of the representer theorem is that the solution in an infinite dimensional space falls in a finite dimensional space. This property makes it possible to compute estimates of general regularization problems in infinite dimensional spaces.

3. The proof of Theorem 1 is quite simple, considering how important it turned out to be. Two key facts used in the proof are the reproducing property and orthogonality. The proofs for various extensions usually follow similar arguments.

4. The solution to (4) is unique when the least squares has a unique minimizer in $\mathcal{H}_0$ [17].

5. The least squares in (4) may be replaced by weighted least squares when observations are heteroscedastic and/or correlated. The representer theorem still holds [18]. In fact the representer theorem holds when the least squares is replaced by a general cost function $c((x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n)))$ and the penalty is replaced by $g(J(f))$ where g is a strictly monotone increasing function on $[0, \infty]$ [7].

In some applications the function f is observed through bounded linear functionals $\mathcal{L}_i : \mathcal{H} \rightarrow \mathbb{R}$ plus random errors:

$$y_i = \mathcal{L}_i f + \epsilon_i, \quad i = 1, \cdots, n. \tag{7}$$

Model (3) is a special case of model (7) with $\mathcal{L}_i$ being the evaluational functional: $\mathcal{L}_i f = f(x_i)$. The estimate of f, $\hat{f}$, is a solution to the penalized least squares with $f(x_i)$ being replaced by $\mathcal{L}_i f$ in (4). Define new representers as $\xi_i(\cdot) = \mathcal{L}_{i(z)} K_1(z, \cdot)$ for $i = 1, \cdots, n$ where $\mathcal{L}_{i(z)}$ indicates that $\mathcal{L}_i$ is applied to what follows as a function of z. Then the representer theorem holds with simple adjustment of the proof: since $\mathcal{L}_i$ are bounded linear functionals, by the Riesz representation theorem, there exist representers $\eta_i \in \mathcal{H}$ such that $\mathcal{L}_i f = \langle \eta_i, f \rangle$. Then $\eta_i(x) = \langle \eta_i, K_x \rangle = \mathcal{L}_i K_x = \sum_{\nu=1}^m (\mathcal{L}_i \phi_\nu) \phi_\nu(x) + \xi_i(x)$. Write any $f \in \mathcal{H}$ as $f = \sum_{\nu=1}^m d_\nu \phi_\nu + \sum_{j=1}^n c_j \xi_j + \rho$ where $\rho \in \mathcal{H}$ is orthogonal to the space spanned by $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$. Then $\mathcal{L}_i \rho = \langle \eta_i, \rho \rangle = 0$ and $J(f) = \sum_{i=1}^n \sum_{j=1}^n c_i c_j \langle \xi_i, \xi_j \rangle + \|\rho\|^2$. The conclusion follows.

The representer theorem reduces the difficult computational problems in high or infinite dimensional spaces to optimization problems on the scalar coefficients $\boldsymbol{c} = (c_1, \cdots, c_n)^T$ and $\boldsymbol{d} = (d_1, \cdots d_m)^T$. Letting T be the $n \times m$ matrix with the $i\nu$th entry $\phi_\nu(x_i)$ and Σ be the $n \times n$ matrix with the ijth entry $K_1(x_i, x_j)$, we need to compute $\boldsymbol{c}$ and $\boldsymbol{d}$ as minimizers of $\|\boldsymbol{y} - T\boldsymbol{d} - \Sigma\boldsymbol{c}\|^2 + \lambda \boldsymbol{c}^T \Sigma \boldsymbol{c}$. The computational details for $\boldsymbol{c}$ and $\boldsymbol{d}$ can be found in Wahba [4], Gu [17], and Wang [18].

The penalty $J(f)$ penalizes departure from the null space $\mathcal{H}_0$. The choice of the penalty depends on several factors such as the domain of the function $\mathcal{X}$, prior knowledge about the function f, and the purpose of the study. For example, one may choose $J(f)$ to incorporate indefinite information that f is close to, but not necessarily in, the space $\mathcal{H}_0$ (often called L-spline). One may also test the hypothesis that f is a parametric model in the space $\mathcal{H}_0$ against the general alternative that $f \in \mathcal{H}$ and $f \notin \mathcal{H}_0$ [19, 18]. To penalize all function in $\mathcal{H}$, one may set $J(f) = \|f\|^2$ and $\mathcal{H}_0$ as an empty set.

The well-known polynomial spline assumes that $\mathcal{X} = [a, b]$, $\mathcal{H}$ is the Sobolev space

$$W_2^m[a, b] = \Big\{ f : \ f, f', \cdots, f^{(m-1)} \text{ are absolutely continuous}, \int_a^b (f^{(m)})^2 dx < \infty \Big\},$$

and $J(f) = \int_a^b (f^{(m)})^2 dx$ penalizing the roughness of the function measured by squared mth derivative. The null space $\mathcal{H}_0$ consists of polynomials of degree $m-1$ or less. The thin-plate spline assumes that $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{f : J_m^d(f) < \infty\}$, $J(f) = J_m^d(f)$, and $2m > d$ where

$$J_m^d(f) = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \right)^2 \prod_{j=1}^{d} dx_j.$$

The null space $\mathcal{H}_0$ consists of polynomials in d variables of total degree up to $m-1$. Other smoothing spline models including periodic spline, spherical spline, and L-spline with different penalties $J(f)$ can be found in Wahba [4], Gu [17], and Wang [18].

The tuning parameter λ in (4) balances the trade-off between goodness-of-fit and penalty. How to choose λ is a separate topic in itself. The commonly used methods in spline smoothing include the Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML), and unbiased risk [20, 21, 4, 17, 18].

3.2 Smoothing spline ANOVA

Consider model (3) where f is a function of multiple covariates denoted as a vector $\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and $\mathcal{X}_1, \cdots, \mathcal{X}_d$ are arbitrary sets. A smoothing spline ANOVA (SS ANOVA) decomposition expresses a function in the tensor product RKHS $\mathcal{H} = \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d$ as

$$f(x_1, \cdots, x_d) = \mu + \sum_{k=1}^{d} f_k(x_k) + \sum_{k<l} f_{kl}(x_k, x_l) + \cdots + f_{1\cdots d}(x_1, \cdots, x_d) \tag{8}$$

where μ represents the grand mean, $f_k(x_k)$ represents the main effect of x_k, $f_{kl}(x_k, x_l)$ represents the two-way interaction between x_k and x_l, and the remaining terms represent higher-order interactions. The SS ANOVA decomposition leads to a hierarchical structure that facilitates model selection and interpretation as the classical ANOVA models. To overcome the curse of dimensionality problem, as in classical ANOVA, high-order interactions are often dropped from the model space. A model consisting of any subset of components in the SS ANOVA decomposition (8) is referred to as an SS ANOVA model. Given an SS ANOVA model, we can regroup and write the model space as

$$\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \cdots \oplus \mathcal{H}^q, \tag{9}$$

where $\mathcal{H}^0$ is a finite dimensional space containing functions that are not penalized, and $\mathcal{H}^1, \cdots, \mathcal{H}^q$ are orthogonal RKHS's with RKs K^j for $j = 1, \cdots, q$. See Wahba [4], Gu [17], and Wang [18] for details about the SS ANOVA models.

Given observations $\{(\boldsymbol{x}_i, y_i),\ i = 1, \cdots, n\}$, the estimate of the multivariate function f, $\hat{f}$, is the solution to the following penalized least squares:

$$\min_{f \in \mathcal{M}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \sum_{j=1}^{q} \lambda_j \|P_j f\|^2 \right\} \tag{10}$$

where P_j is the projection operator onto $\mathcal{H}^j$ and λ_j's are the tuning parameters which allows different penalties for components in different spaces. Let $\mathcal{H}_1^* = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_q$, $\lambda_j = \lambda/\theta_j$, and define a new inner product in $\mathcal{H}_1^*$ as

$$\langle f, g \rangle_* = \sum_{j=1}^{q} \theta_j^{-1} \langle f, g \rangle.$$

Then it is easy to verify that the RK of $\mathcal{H}_1^*$ under the new inner product is $K_1^* = \sum_{j=1}^{q} \theta_j K^j$ and the penalized least squares reduces to

$$\min_{f \in \mathcal{M}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda J(f) \right\} \tag{11}$$

where $J(f) = \|P_1^* f\|^2$ and P_1^* is the projection in $\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}_1^*$ onto $\mathcal{H}_1^*$. From Theorem 1, the estimate $\hat{f}(\boldsymbol{x})$ has the same representation as equation (5) where $\phi_1, \cdots, \phi_m$ are basis functions of $\mathcal{H}^0$ and $\xi_i(\cdot) = K_1^*(\boldsymbol{x}_i, \cdot) = \sum_{j=1}^{q} \theta_j K^j(\boldsymbol{x}_i, \cdot)$.

3.3 Penalized likelihood

The likelihood function may be used as the cost function when the distribution is known. One important example is the nonparametric regression in exponential family. Assume that y_i are generated from a distribution in the exponential family with the conditional density function

$$g(y|x) = \exp\left\{\frac{yf(x) - b(f(x))}{a(\phi)} + c(y, \phi)\right\}, \tag{12}$$

where $f(x) = h(\mathrm{E}(y|x))$, h is the canonical link, $a > 0$, b, and c are known functions, and ϕ is either known or a nuisance parameter. The goal is to model and estimate the function f. Given observations $\{(x_i, y_i),\ i = 1, \cdots, n\}$ and assuming that $f \in \mathcal{H}$, we estimate f as the solution to the following penalized likelihood

$$\min_{f\in\mathcal{H}}\left\{-\sum_{i=1}^{n}(y_i f(x_i) - b(f(x_i))) + \lambda J(f)\right\} \tag{13}$$

where the first term is part of the negative log likelihood with components $c(y_i, \phi)$ being removed since they are independent of f and the component $a(\phi)$ being absorbed into λ. Again, the representer theorem holds [3]: the solution to (13) has the representation (5) where $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$ are defined in Section 3.1.

Another important example is the condition density estimation. Denote $g(y|x)$ as the conditional density of y given x. To deal with the constraints of the density function, assume that $g > 0$ and consider the logistic transformation [17]

$$g(y|x) = \frac{\exp\{f(x, y)\}}{\int_{\mathcal{Y}} \exp\{f(x, y)\}dy}.$$

The bivariate function f is free of constraints. A model space for f may be derived through SS ANOVA decomposition with certain terms being removed for identifiability [17]. Denote the model space for f as $\mathcal{M}$ given in (9). Given observations $\{(x_i, y_i),\ i = 1, \cdots, n\}$, we estimate f as the solution to the following penalized likelihood

$$\min_{f\in\mathcal{M}}\left\{-\sum_{i=1}^{n}\left[f(x_i,y_i)-\log\int_{\mathcal{Y}}\exp\{f(x_i,y)\}dy\right]+\sum_{j=1}^{q}\lambda_j\|P_jf\|^2\right\} \tag{14}$$

where the first term is the negative log likelihood, and P_j is the projection operator onto $\mathcal{H}^j$. The representer theorem no longer holds for this situation since the cost function depends on f through the integral $\int_{\mathcal{Y}}\exp\{f(x_i,y)\}dy$. Nevertheless, the representer theorem provides an approximate estimate. Let $\phi_1,\cdots,\phi_m$ be basis functions of $\mathcal{H}^0$ and $\xi_i(\cdot)=\sum_{j=1}^{q}\theta_jK^j(\boldsymbol{z}_i,\cdot)$ where $\boldsymbol{z}_i=(x_i,y_i)$ and $\lambda_j=\lambda/\theta_j$. An approximate solution based on the form of the representer theorem is $\hat{f}(x,y)=\sum_{\nu=1}^{m}d_\nu\phi_\nu(x,y)+\sum_{j=1}^{r}c_j\tilde{\xi}_j(x,y)$ where $\{\tilde{\xi}_1,\cdots,\tilde{\xi}_r\}$ is a subset of $\{\xi_1,\cdots,\xi_n\}$. The approximate estimate $\hat{f}$ has nice theoretical properties [17]. This example illustrates that the influence of the representer theorem is not limited to situations when solutions to regularization problems fall in finite dimensional spaces. For many complicated applications where there is no finite dimensional solution, the representer theorem can be used to derive approximate estimates with theoretical guarantees [17].

4 Classification

4.1 Soft and hard classification

The goal of classification is to assign an observation to one of the two or more categories. We consider binary classification problem in this section. Based on the training sample $\{(x_i,y_i),\ i=1,\cdots,n\}$ where y_i are class labels taking values 1 or -1, the task is to create a classification rule labeling a new observation as one of the two categories.

Denote $p(x)=P(y=1|x)$ as the conditional probability and $f(x)=\log\{p(x)/(1-p(x))\}$ as the log odds ratio. The function $f(x)$ describes how the relative risk of two categories varies with x which are of interest in a wide range of applications. Treating y_i as binary data with binomial distribution in the exponential family, the negative log likelihood is $\log(1 + \exp\{y_i f(x_i)\})$. Results in Section 3.3 assures that the penalized likelihood estimate of f in an RKHS has the representation (5).

Under the assumption of equal costs of misclassification for both kinds of misclassification, the optimal classification rule is to label an observation as 1 when $\hat{f}(x) > 0$ and -1 when $\hat{f}(x) < 0$ [6, 22]. This approach, often referred to as soft classification, estimates the logit function f first and then perform classification. A misclassification happens when $-y\hat{f}(x) > 0$, that is, the signs of y and $\hat{f}(x)$ do not match. This suggests that one may consider a cost function in a form of $V(yf(x))$ for the purpose of minimizing classification error. The quantity $yf(x)$ is commonly referred to as the functional margin.

The SVM was proposed in Boser, Guyon and Vapnik [23], and Vapnik [24] with a geometrical interpretation of finding a separating hyperplane in a multidimensional input space. In a meeting in Mt. Holyoke it came about that the SVM could be obtained as the solution to an optimization problem in an RKHS with the hinge loss function $V(u) = (1-u)_+$ where $(u)_+ = u$ when $u > 0$ and $(u)_+ = 0$ otherwise. Specifically, we find $f \in \text{span}\{1\} \oplus \mathcal{H}$ as the solution to

$$\min_{f \in \text{span}\{1\} \oplus \mathcal{H}} \left\{ \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|Pf\|^2 \right\} \tag{15}$$

where P is the projection operator onto $\mathcal{H}$. The representer theorem again holds [6]:

$$\hat{f}(x) = d + \sum_{i=1}^{n} c_i K(x_i, x)$$

where K is the RK of the RKHS $\mathcal{H}$. The SVM provides a classification rule that directly targets on the classification decision boundary. This approach is often referred to as hard classification since it does not produce the probability estimation.

4.2 Multicategory classification

Many classification problems involve more than two categories. Suppose that there are $k > 2$ categories. Denote $p_j(x)$ as the conditional probability that y belong to class j given x. If the misclassification costs are all equal, then the Bayes decision rule assigns a new x to the class with the largest $p_j(x)$.

As in the binary classification, the soft classification approach models and estimates the probabilities $p_j(x)$ first and then use them for classification. Let $f_j(x) = \log\{p_j(x)/p_1(x)\}$ be the odds ratio between classes j and 1, $j = 2, \cdots, k$. Let $y_{ij} = 1$ if the ith observation is in class j and 0 otherwise for $i = 1, \cdots, n$ and $j = 2, \cdots, k$. Assume that $f_j \in \mathcal{H}_j$. The penalized likelihood estimates $\hat{f}_j$'s of f_j's are solutions to

$$\min_{f_2 \in \mathcal{H}_2, \cdots, f_k \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \left[-\sum_{j=2}^{k} y_{ij} f_j(x_i) + \log\left(1 + \sum_{j=2}^{k} \exp\{f_j(x_i)\}\right) \right] + \sum_{j=2}^{k} \lambda_j J_j(f_j) \right\}, \tag{16}$$

where x_i is the feature of the ith observation, $J_j(f_j)$'s are square seminorm penalties and λ_j's are smoothing parameters. The representer theorem in this case states that [25]

$$\hat{f}_j(x) = \sum_{\nu=1}^{m_j} d_{\nu j} \phi_{\nu j}(x) + \sum_{i=1}^{n} c_{ij} K_{1j}(x_i, x), \tag{17}$$

where $\phi_{1j}, \cdots, \phi_{m_j j}$ are basis functions of the null space $\mathcal{H}_{j0} = \{h: \ J_j(h) = 0\}$ and K_{1j} is the RK of $\mathcal{H}_j \ominus \mathcal{H}_{j0}$.

To obtain a symmetric generalization of the two class SVM to the multicategory case, define $y_{ij} = 1$ if the ith observation is in class j and $y_{ij} = -1/(k-1)$ otherwise, $i = 1, \cdots, n$ and $j = 1, \cdots, k$. Let L be the $k \times k$ matrix with 0 on the diagonal and 1 elsewhere, a cost matrix when all of the misclassification costs are equal. Denote the jrth elements of L as L_{jr}. Consider k separating functions $f_j \in \text{span}\{1\} \oplus \mathcal{H}$ with sum-to-zero constraint $\sum_{j=1}^{k} f_j(x) = 0$. The multicategory SVM is the solution to

$$\min_{f_1, \cdots, f_k \in \text{span}\{1\} \oplus \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} L_{\text{cat}(i)j} (f_j(x_i) - y_{ij})_+ + \lambda \sum_{j=1}^{k} \|P f_j\|^2 \right\} \tag{18}$$

under constraint $\sum_{j=1}^{k} f_j(x) = 0$, where $\text{cat}(i) = v$ if the ith observation is from category v, and P is the projection operator onto $\mathcal{H}$. Denote K as the RK of $\mathcal{H}$. The representer theorem in this case states that the optimization problem (18) is equivalent to finding f_j's of the form

$$f_j(x) = d_j + \sum_{i=1}^{n} c_{ij} K(x_i, x)$$

with the sum-to-zero constraint at x_i for $i = 1, \cdots, n$ [26].

5 Functional Regression

Functional data analysis (FDA) deals with data that are functions [27, 28, 29]. We consider functional regression that investigates the relationship between a covariate x and a response y where at least one of the x and y is a function.

First consider the case when y is a scalar and x is a real-valued function on an arbitrary domain $\mathcal{T}$. Denote the observations as $\{(x_i, y_i),\ i = 1, \cdots, n\}$. Consider the model

$$y_i = F(x_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{19}$$

where $x \in \mathcal{X}$ and F is a functional that maps $\mathcal{X}$ to $\mathbb{R}$. The goal is to model and estimate the functional F.

One simple model assumes that $\mathcal{X}$ is a Hilbert space (e.g. $L_2(\mathcal{T})$) with inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ and

$$F(x) = \alpha + \langle x, \beta \rangle_{\mathcal{X}} \tag{20}$$

where $\beta \in \mathcal{H}$, $\mathcal{H} \subset \mathcal{X}$ is an RKHS of real-valued functions on the domain $\mathcal{T}$ [27, 30, 18]. The estimates of the scalar α and function β are solution to the penalized least squares

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathcal{H}} \left\{ \sum_{i=1}^{n} \left(y_i - \alpha - \langle x_i, \beta \rangle_{\mathcal{X}} \right)^2 + \lambda J(\beta) \right\} \tag{21}$$

where $J(\beta)$ is a square seminorm penalty. Assuming that $\mathcal{L}_i\beta \equiv \langle x_i, \beta\rangle_{\mathcal{X}}$ for $i = 1, \cdots, n$ are bounded linear functionals on $\mathcal{H}$, then by the results in Section 3, the estimate of β is a linear combination of $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$ where $\phi_1, \cdots, \phi_m$ are basis functions of $\mathcal{H}_0 = \{\beta : \ J(\beta) = 0\}$, $\xi_i(t) = \langle x_i(\cdot), K_1(t, \cdot)\rangle_{\mathcal{X}}$, and K_1 is the RK of $\mathcal{H}_1 = \mathcal{H} \ominus \mathcal{H}_0$.

Model (20) may be too restrictive for some applications. One may adopt the RKHS approach by considering F as a function on the domain $\mathcal{X}$ of functions [31]. Assume that $F \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with functions (functionals) on $\mathcal{X}$. Since the representer theorem holds for arbitrary domain, so the solution to the penalized least squares

$$\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} (y_i - F(x_i))^2 + \lambda J(F) \right\} \tag{22}$$

is a linear combination of $\phi_1, \cdots, \phi_m$ and $\xi_1, \cdots, \xi_n$ where $\phi_1, \cdots, \phi_m$ are basis functions of $\mathcal{H}_0 = \{F : \ J(F) = 0\}$, $\xi_i(x) = K_1(x_i, x)$, and K_1 is the RK of $\mathcal{H}_1 = \mathcal{H} \ominus \mathcal{H}_0$. Preda [31] considered the Gaussian kernel $K_1(x, z) = \exp\{-\|x - z\|_{\mathcal{X}}^2/(2\sigma^2)\}$ and inhomogeneous polynomial kernel $K_1(x, z) = (c + \langle x, z\rangle_{\mathcal{X}})^d$ where $\|\cdot\|_{\mathcal{X}}$ and $\langle \cdot, \cdot\rangle_{\mathcal{X}}$ are the norm and inner product of the space $\mathcal{X}$, σ and c are real numbers and d is a natural number.

Next consider the case when y is a function on an arbitrary domain $\mathcal{W}$ and $x \in \mathcal{X}$ is either a scalar or a function. Wahba [5] proposed the following discrete approach. Denote observations as $\{(x_i, y_i(w_{ij})), \ i = 1, \cdots, n; \ j = 1, \cdots, J_i\}$ where functions y_i are observed at discrete points w_{ij} while x_i is observed in whole if it is a function. Assume the model

$$y_i(w_{ij}) = f(w_{ij}, x_i) + \epsilon_{ij}, \quad i = 1, \cdots, n; \ j = 1, \cdots, J_i, \tag{23}$$

where f is a bivariate function (functional) on $\mathcal{W} \times \mathcal{X}$ and ϵ_{ij} are iid random errors with mean zero. Let $K((w, x), (u, z))$ be an RK where $(w, x), (u, z) \in \mathcal{W} \times \mathcal{X}$. For a fixed x, let $k_x(w, u) = K((w, x), (u, x))$. Since k_x is a positive definite function on $\mathcal{W} \times \mathcal{W}$, there exists an RKHS denoted as $\mathcal{H}_{k_x}$ such that its RK is k_x. It was shown that for each fixed $(w_*, x_*) \in \mathcal{W} \times \mathcal{X}$, the RK K defines an $\mathcal{H}_{k_x}$-valued function of x [5]. That is, letting

$K_{(w_*,x_*)}(w,x) = K((w,x),(w_*,x_*))$, $K_{(w_*,x_*)}(\cdot,x)$ is an element of $\mathcal{H}_{k_x}$ for each $x \in \mathcal{X}$. Let $\mathcal{H}_K$ be the linear span of all such $\mathcal{H}_{k_x}$-valued functions $K_{(w_*,x_*)}(\cdot,x)$ for all $(w_*,x_*) \in \mathcal{W}\times\mathcal{X}$ which is closed with respect to the inner product $\langle K_{(w,x)}, K_{(u,z)}\rangle_K = K((w,x),(u,z))$. Assuming that $f \in \mathcal{H}_K$, the estimate of f, $\hat{f}$, is the solution to

$$\min_{f\in\mathcal{H}_K} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{J_i}(y_i(w_{ij}) - f(w_{ij},x_i))^2 + \lambda\|f\|_K^2 \right\}. \tag{24}$$

Then the estimate has a representation [5]

$$\hat{f}(\cdot,x) = \sum_{i=1}^{n}\sum_{j=1}^{J_i} c_{ij}K_{(w_{ij},x_i)}(\cdot,x). \tag{25}$$

This approach starts with the RK K on $\mathcal{W}\times\mathcal{X}$. One example is to assume that K is the product of RK's on $\mathcal{W}$ and $\mathcal{X}$: $K((w,x),(u,z)) = K_{\mathcal{W}}(w,u)K_{\mathcal{X}}(x,z)$. This is equivalent to assuming that f belong to the tensor product of two RKHS's with RK's $K_{\mathcal{W}}$ and $K_{\mathcal{X}}$ respectively. SS ANOVA decomposition may be applied to the tensor product space to construct different models for f. If $K_{\mathcal{X}}$ is isotropic (i.e. $K_{\mathcal{X}}(x,z) = E(\|x-z\|)_{\mathcal{X}}$), then $K((w,x),(u,x))$ does not depend on x and consequently $\mathcal{H}_{k_x}$ does not depend on x.

Finally consider the case when both x and y are functions. To model and estimate the map F from $\mathcal{X}$ to $\mathcal{Y}$ nonparametrically, Lian [32] and Kadri et al [33] extended the RKHS from function space to function-valued space. Assume that $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are separable Hilbert spaces of real-valued functions. Denote $\mathcal{L}(\mathcal{Y})$ as the set of bounded linear operators from $\mathcal{Y}$ to $\mathcal{Y}$. An $\mathcal{L}(\mathcal{Y})$-valued kernel K is a map from $\mathcal{X}\times\mathcal{X}$ to $\mathcal{L}(\mathcal{Y})$. Such a kernel is positive definite on $\mathcal{X}$ if (i) $K(x,z) = K(z,x)^*$ for any $x,z \in \mathcal{X}$ where the superscript * denotes the adjoint operator; and (ii) for any $z_1,\cdots,z_r \in \mathcal{X}$ and $u_1,\cdots,u_r \in \mathcal{Y}$, the matrix with the ijth entry $\langle K(z_i,z_j)u_i,u_j\rangle_{\mathcal{Y}}$ is positive semi-definite. A function-valued Hilbert space $\mathcal{H}$ of functions (operators) from $\mathcal{X}$ to $\mathcal{Y}$ is an RKHS if there exists a positive definite $\mathcal{L}(\mathcal{Y})$-valued kernel K on $\mathcal{X}\times\mathcal{X}$ such that (a) $K(\cdot,x)y \in \mathcal{H}$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$; and (b) $\langle F, K(x,\cdot)y\rangle_{\mathcal{H}} = \langle F(x),y\rangle_{\mathcal{Y}}$ for any $x\in\mathcal{X}$, $y\in\mathcal{Y}$, and $F\in\mathcal{H}$.

Similar to the classical RKHS, every function-valued RKHS has a unique $\mathcal{L}(\mathcal{Y})$ valued RK that is positive definite, and every positive definite $\mathcal{L}(\mathcal{Y})$-valued kernel K is associated with a functional-valued RKHS with K as its RK.

The estimate of F, $\hat{F}$, is the solution to

$$\min_{F\in\mathcal{H}}\left\{\sum_{i=1}^{n}\|y_i - F(x_i)\|^2_{\mathcal{Y}} + \lambda\|F\|^2_{\mathcal{H}}\right\}. \tag{26}$$

Lian [32] and Kardi et al [33] extended the representer theorem under this setting:

$$\hat{F}(x) = \sum_{i=1}^{n} K(x_i, x)u_i, \quad u_i \in \mathcal{Y}. \tag{27}$$

To apply the representer theorem one needs a specific form of the RK which is usually difficult to construct under this setting. Note that $u_i \in \mathcal{Y}$ are functions rather than scalars. Computation of these function "coefficients" are not straightforward. Methods for the RK construction and computation of functions u_i are given in Kadri [33]. In particular, a separable kernel construction assumes that [33]

$$K(x, z) = K_{\mathcal{X}}(x, z)T \tag{28}$$

where $K_{\mathcal{X}}$ is a scalar-value kernel on $\mathcal{X}$ and T is an operator in $\mathcal{L}(\mathcal{Y})$. Lian [32] considered a spacial case of (28) with $K_{\mathcal{X}}(x, z) = E(\|x - z\|_{\mathcal{X}})$ and $T = I$ where E is a real valued positive definite function and I is the identity operator.

In this section we have seen that the RKHS can be used to model functional regression nonparametrically. The technical difficulty lies in the construction of flexible RKs for different applications where future research is needed.

Acknowledgments

Grace Wahba's research is supported in part under National Science Foundation Grant DMS 1308877. Yuedong Wang's research is supported in part under National Science Foundation Grant DMS 1507620.

6 Related Articles

See also **Classification; Functional Data Analysis; Maximum Penalized Likelihood Estimation; Nonparametric Regression; Regression; Regularization Methods; Reproducing Kernel Hilbert Space; Smoothing; Splines in Nonparametric Regression; Spline Smoothing; Support Vector Machines**.

References

[1] Kimeldorf, G. & Wahba, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495–502.

[2] Kimeldorf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, **33**, 82–95.

[3] O'Sullivan, F., Yandell, B. & Raynor, W. (1986), Automatic smoothing of regression functions in generalized linear models, *J. Amer. Statist. Assoc.*, **81**, 96–103.

[4] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

[5] Wahba, G. (1992). Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII*, Casdagli, M. & Eubank, S. eds, 95–112.

[6] Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods-Support Vector Learning*, Schölkopf, B., Burges, C. & Smola, A. eds, 69–88.

[7] Schölkopf, B., Herbrich, R. & Smola, A. (2001). A generalized representer theorem. In *Lecture Notes in Computer Science, Vol. 2111*, pages 416–426.

[8] Argyiou, A., Micchelli, C. & Pontil, M. (2009). When is there a representer theorem? Vector versus matrix regularizers. *J. Mach. Learn. Res.*, **10**, 2507–2529.

[9] Zhang, H. & Zhang, J. (2012), Regularized learning in Banach spaces as an optimization problem: representer theorems. *J. Glob. Optim.*, **54**, 235–250.

[10] Bohn, B., Griebel, M. & Rieger, C. (2017). A representer theorem for deep kernel learning. *arXiv preprint arXiv:1709.10441.*

[11] Unser, M. (2018), A representer theorem for deep neural networks. *arXiv preprint arXiv:1802.09210.*

[12] Diwale, S. & Jones, C. (2018). A generalized representer theorem for Hilbert space - valued functions. *arXiv preprint arXiv:1809.07347.*

[13] Dinuzzo, F. & Schölkopf, B. (2012). The representer theorem for Hilbert spaces: a necessary and sufficient condition. in *Adv. Neural Inf. Process. Syst. 25*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., 189–196.

[14] Yu, Y., Cheng, H., Schuurmans, D. & Szepesvari, C. (2013). Characterizing the representer theorem. In *Proceedings of the 30th International Conference on Machine Learning*, 570–578.

[15] Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.*, **68**, 337–404.

[16] Akhiezer, N.I. & Glazman, I.M. (1963). *Theory of Linear Operators in Hilbert Space.* Ungar, New York.

[17] Gu, C. (2013). *Smoothing Spline ANOVA Models*. Springer-Verlag, New York.

[18] Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Chapman & Hall, New York.

[19] Liu, A. and Wang, Y. (2004), Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, **74**: 581–597

[20] Craven, P. & Wahba, G. (1979), Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

[21] Golub, G., Heath, M. & Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–224.

[22] Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proc. Natl. Acad. Sci.*, **99**: 16524–16530.

[23] Boser, B., Guyon, I. & Vapnik, V. (1992). A training agorithn for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh*, Haussler, D. eds, 144–152.

[24] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[25] Lin, X. (1998). Smoothing spline analysis of variance for polychotomous response data. Technical Report 1003, Department of Statistics, University of Wisconsin, Madison WI.

[26] Lee, Y., Lin, Y. & Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Amer. Statist. Assoc.*, **99**, 67–81.

[27] Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

[28] Ferraty F. & Vieu P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York.

[29] Wang, J. L., Chiou, J. M. & Müller, H. G. (2015). Functional data analysis. *Annu. Rev. Statist.*, **3**: 257–295.

[30] Yuan, M. & Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear model, *The Annals of Statistics*, **38**: 3412–3444.

[31] Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, **137**: 829–840.

[32] Lian H. (2007). Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *The Canadian Journal of Statistics*, **35**, 597–606.

[33] Kardi, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A. & Audiffren, J. (2016). Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, **17**: 1–54.