Advanced Statistical Methods

KERNEL RIDGE REGRESSION

December 6, 2016

Leonard Strnad
University of Colorado at Denver
Department of Mathematics and Statistics

Introduction

In this paper we investigate the approximation of continuous nonlinear functions using kernel ridge regression using a gaussian radial basis kernel function. This is a different approach than typical statistical methods as it assumes the data are finite realizations of an infinite dimensional random element. Kernel methods are extremely robust in that they are capable of defining an implicit map on the data which can map the data into a higher dimension (potentially infinite dimensional) where the data is then in a linear feature space. There are strong ties to other kernel methods like Gaussian Processes [1], Support Vector Machines [2] and Neural Networks among many others. There has also been a lot of development in speeding up the process–specifically the Nystrom method [3].

The results that are the foundation for these statistical methods stem from RKHS theory. RKHS is short for the Reproducing Kernel Hilbert Space which is a complete inner product space of functionals which can be represented as a linear combination of positive definite kernels evaluated at each x in $\mathcal{X}$. The theory provides a closed form for the solution that minimizes the typical loss function and regularization term. In fact, the results show that we can reduce the infinite dimensional minimization problem to a computationally tractable finite dimensional minimization problem. We will also discuss the "kernel trick" which provides a significant computational advantage when mapping data into a feature space. We will also discuss how the implicit feature space that results from a radial basis function or gaussian kernel maps the realization x into a infinite dimensional feature space. We also discuss the spectral decomposition of the kernel operator/matrix and discuss its similarity to principal component analysis.

The application portion of this paper explores two data sets: Donoho's doppler data set and the noisy sinc data set [4] found in the CVST package [5]. We will use the CVST package to perform kernel ridge regression on this data set with varying levels of noise. The packages also includes a improved cross validation process that is exceptionally fast which we will use for model selection for each level of noise in the data. The R tutorial for this project can be found at `https://github.com/ljstrnadiii/KernelRidgeRegressionProject`

We begin by giving a brief introduction to Reproducing Kernel Hilbert Spaces (RKHS), the regularization framework, discuss the spectral decomposition of Kernel matrices and their relation to the Kernel operator, introduce kernel ridge regression (KRR), and discuss the gaussian radial basis function. Lastly, we discuss the data sets considered, demonstrate KRR on these data sets and perform model selection using the fast cross validation method in the CVST package.

Methods

Reproducing Kernel Hilbert Spaces

A Hilbert space is a complete inner product space. A RKHS, $\mathcal{H}$, is a Hilbert space of functions defined over some bounded domain, $\mathcal{X}$ equipped with an evaluation functional defined as

$$\mathcal{F}_x[f] = f(x) \quad \forall\ f \in \mathcal{H} \tag{1}$$

that is linear and bounded i.e. there exists an M s.t.

$$|\mathcal{F}_x[f]| = |f(x)| \leq M\|f\|. \tag{2}$$

The Riesz representation theorem implies that $\forall\ x \in \mathcal{X}$ there exists $k_x = k(\cdot, x) \in \mathcal{H}$ with the *reproducing property*

$$f(x) = \mathcal{F}_x[f] = \langle f, k_x \rangle_{\mathcal{H}} \quad \forall\ f \in \mathcal{H} \tag{3}$$

and since $k_y \in \mathcal{H}$ we have $\forall\ y \in \mathcal{X}$ that

$$k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} \tag{4}$$

which allows us to define the reproducing kernel of $\mathcal{H}$ as a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}} \tag{5}$$

which implies that k is both symmetric and positive definite. If we characterize the symmetric positive definite kernel, k, through the integral operator (Hilbert-Schmidt integral operator) as

$$[T_k\phi](x) = \int_{\mathcal{X}} k(x, s)\phi(s)ds \tag{6}$$

then Mercer's theorem shows that the operator is compact, self adjoint and provides a series representation of k composed of the eigenvalues and eigenfunctions of T_k. We also have from [6] the spectral theorem for compact, self-adjoint operators which says that the nonzero eigenvalues of T_k form a finite or countably infinite set. Thus, we assume that k takes on the form

$$k(x, y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x)\phi_i(y) \tag{7}$$

where $\gamma_i \geq 0$, $\sum \gamma_i^2 < \infty$. Now, consider arbitrary linear combinations of the form $f(x) = \sum_m \alpha_m k_x(y_m)$. Then we have that elements of H are of the form

$$f(x) = \sum_m \alpha_m k_x(y_m) = \sum_{i=1}^{\infty} c_i \phi_i(x) \tag{8}$$

where $\|f\| = \sum c_i^2/\gamma_i < \infty$. That is, we can express an element in the RKHS that may be an arbitrary sum as an element of a countable sum. We now have a feature space defined by $\{\phi_i(x)_{i=1}^{\infty}, x \in \mathcal{X}\}$. Generally, it can be shown [7] that whenever k is of form (6), then we can construct a RKHS and vice versa. The $\{\phi_i\}$ construct a basis for the RKHS and the kernel k is the "correlation" associated with these basis functions [8]. There is a close relationship between regularization, RKHS, and gaussian processes [7, 9, 10, 11] . In the next section we will discuss the RKHS in the context of regularization networks.

The Regularization Framework

The method we use to estimate f is a specific class of regularization problems of the form

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \tag{9}$$

where L is any loss function and J is any regularization/penalty function defined on a Hilbert space, $\mathcal{H}$. Next, we restrict to a subclass of $\mathcal{H}$ generated by some positive definite kernel, $k(x, y)$, which gives us a RKHS we will call $\mathcal{H}_k$ that has all the properties discussed above. We will see that $J(f)$ defined on this space can be expressed in terms of the kernel, k. So, we are now minimizing the expression

$$\min_{f \in \mathcal{H}_k} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \tag{10}$$

but we have that $f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$ which implies that we are similarly minimizing the expression

$$\min_{\{c_i\}_i^{\infty}} \left[\sum_{i=1}^{N} L(y_i, \sum_{i=1}^{\infty} c_i \phi_i(x)) + \lambda J(f) \right]. \tag{11}$$

It has been shown by Wahba [7] that (11) lends itself to a finite dimensional solution of the form

$$\hat{f}(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i) \tag{12}$$

This solution form can be interpreted as a linear combination of inner products of the data in some implicit feature space. It is saying that our estimated function is in the span of the image of k on $\mathcal{X}$. Consider $k(x_i, x_j) = \langle k(\cdot, x_i), k(\cdot, x_j) \rangle$. This allows us to directly express the inner product of the data in some feature space implicitly defined by k which can be

expressed in terms of a matrix, $\mathbf{K}$. If we consider our penalty function to be the norm of f, then it can be shown from (8) that

$$J(f) = \sum^{N}\sum^{N} k(x_i, x_j)\alpha_i\alpha_j = \alpha^{\mathbf{t}}\mathbf{K}\alpha. \tag{13}$$

In light of the things discussed above we see that the minimization problem has reduced to the finite dimensional problem

$$\min_{\alpha} L(\mathbf{y}, \mathbf{K}\alpha) + \lambda\alpha^{\mathbf{t}}\mathbf{K}\alpha \tag{14}$$

from a potentially uncountably infinite dimensional space. This property of reducing the infinite minimization problem to a finite dimensional problem is called the *kernel property.* Lastly, in light of (8) can be shown that $\|f\| = \sum_{i=1}^{\infty} c_i/\gamma_i < \infty$ which expresses a similar penalization as regular kernel regression where functions with large eigenvalues in the expansion of (8) get penalized less i.e. the process "flattens" the data before fitting. Flattening refers to the eigenbasis of the feature space which has a low associated eigenvalue. So, if we increase λ we will see that we are reducing the dimension of the solution space. In the next section, we look more closely at the integral operator mention in the RKHS section above and discuss its spectral decomposition.

Kernel Operators and Spectral Decomposition

The kernel function, k, is a defined positive definite mapping $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and as discussed above it also has a unique function space–the reproducing kernel hilbert space, $\mathcal{H}_k$. This RKHS is generated by the span of k on $\mathcal{X}$. We consider the integral operator of the kernel function because we are actually assuming the. underlying data comes from a random function defined on $\mathcal{X}$ that is an element of the RKHS. Since we are working with finite realizations of $\mathcal{X}$, the kernel matrix, $\mathbf{K}$, is actually a finite approximation of the kernel defined as an operator on $\mathcal{H}_k$. The kernel operator is defined as

$$[K_p f](y) = \int_{\mathcal{X}} k(x, y)f(x)p(x)d\mathcal{X} \tag{15}$$

where p is a probability distribution over $\mathcal{X}$ which exists for a gaussian radial basis kernel by virtue of Bochner's theorem. If we consider this operator we can discuss the notion of eigenvalues and eigenfunctions through the expression

$$K_p\phi = \lambda\phi \quad or \quad \int_{\mathcal{X}} k(x, y)\phi(x)p(x)d\mathcal{X} = \lambda\phi(y) \tag{16}$$

where ϕ is an eigenfunction of K_p in $\mathcal{H}_k$. The eigenvalue and eigenfunction both depend on p and k. Suppose λ_i and $\mathbf{v_i}$= $(v_1, ..., v_2)^t$ are an eigenvalue and eigenvector of the kernel matrix, $\mathbf{K}$ s.t.

$$\mathbf{K}\mathbf{v_i} = \lambda_\mathbf{1}\mathbf{v_i} \tag{17}$$

then

$$\frac{1}{n}\sum_{j=1}^{n} k(x_i, x_j)v_j = \frac{\lambda_1}{n}v_i \tag{18}$$

where the LHS is an approximation to its integral counterpart

$$\frac{1}{n}\sum_{j=1}^{n} k(x_i, x_j)\phi(x_j) \approx \int_{\mathcal{X}} k(x, y)\phi(x)p(x).d\mathcal{X}. \tag{19}$$

This implies that λ_i/n is an approximation to λ of the kernel operator with eigenfuction ϕ. This result discussed by [12] demonstrates that the eigenpair of the kernel matrix $\mathbf{K}$ are discrete approximations to the eigenbasis elements of K_p. These approximations are given by the expression

$$\hat{\phi}(x) = \frac{1}{\lambda_i n}\sum_{i=1}^{n} v_i k(x_i, x). \tag{20}$$

We will focus on the gaussian radial basis kernel and discuss how the eigendecomposition of the kernel matrix has eigendecomposition with eigenvalues that effectively converge to zero. This has a close similarity with principal component analysis in that we can express our solution/feature space as a sum of $m < n$ eigenvectors of the kernel matrix which are approximating the eigendecomposition of our feature space. Next, we will discuss the specific setting of our regularization network that is kernel ridge regression.

Kernel Ridge Regression

If we specify our regularization network (14) by choosing the squared-error loss function then we are in the Ridge Regression framework. Specifically, we have the minimization problem

$$\min_{\alpha}(\mathbf{y} - \mathbf{K}\alpha)^\mathbf{t}(\mathbf{y} - \mathbf{K}\alpha) + \lambda\alpha^\mathbf{t}\mathbf{K}\alpha \tag{21}$$

which has the solution of the form $\hat{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{y}$ shown in [13]. This gives us our estimated function

$$\hat{f} = \mathbf{K}\hat{\alpha} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K}^{-\mathbf{1}})^{-\mathbf{1}}\mathbf{y} \tag{22}$$

where $\mathbf{K}$ is the kernel matrix and $\mathbf{y}$ is the vector of responses that correspond to the data being implicitly mapped by $\mathbf{k}$. The QR decomposition is similar to that of regular linear regression, however, $\mathbf{K}$ is itself $O(N^3)$. Faster methods like the Nystrom method have been studied to increase the performance of the algorithm [3].

Gaussian Radial Basis Function Kernel

The goal is to choose the proper kernel which implicitly maps $\mathcal{X}$ into a higher dimensional feature space that constructs a linear representation of the data. The gaussian radial basis function kernel is a common kernel used for learning algorithms and has great success at mapping data into a linear feature space. The gaussian kernel

$$k(x, y) = e^{-v\|x-y\|} \tag{23}$$

corresponds to an RKHS that is an expansion in the gaussian radial basis functions in the form of (4) as

$$k_y(x) = e^{-v\|x-y\|^2} \quad y \in \mathcal{X}. \tag{24}$$

In order for the theory to work we just need to know that $k(x, y)$ is a valid inner product on some implicit feature space. It can be shown that (24) can be expressed as

$$k(x, y) = Ce^{-\langle x,y\rangle} \tag{25}$$

which has been shown to have the taylor series expansion of

$$k(x, y) = C\sum_{k=1}^{\infty} \frac{\langle x, y\rangle^k}{k!}. \tag{26}$$

This suggests that $k(x, y)$ is an inner product of feature elements. In fact, we are taking the inner product of infinite dimensional elements which means that $k(x, y)$ finds the inner product of x and y in some infinite dimensional feature space. From a computational perspective, this is a cheap and tractable method to find an inner product of an infinite dimensional feature space and is often called the *kernel trick*. This is tremendously powerful because it is a way of projecting the data into a feature space that is linear which allows for ridge regression in that feature space.

Experiments

Overview

The application portion of this paper uses the CVST R packages to implement KRR and perform model selection using a fast cross validation scheme called sequential cross validation. In order to provide a sense of performance, the mean squared error (mse) and time to process is reported for the cross validation. There are two data sets considered: Donoho's doppler

data set and the noisy sinc data set found in the CVST R package. The noisy sinc data set is generated from

$$f(x) = \frac{\sin(4x)}{4x} + \frac{\sin(15dx)}{5} + \epsilon \quad \epsilon \sim N(0, \sigma^2),\ x \in [-\pi, \pi],\ d \in \{1, 2\} \tag{27}$$

and Donoho's doppler data set is generated from

$$f(x) = (x(1-x))^{1/2} \sin(2\pi(1+\epsilon)/(x+\epsilon)) \quad \epsilon \sim N(0, \sigma^2). \tag{28}$$

The noisy sinc data set (figure 1) has an underlying function which exemplifies periodicity within a sinusoidal function with varying amplitude. Demonstrating that KRR can fit well to this underlying function provides strong evidence of its flexibility. It is an excellent data set for testing nonlinear estimation because with increasing noise the second term in the sum of (27) will begin to hide. We will consider two levels of variance in the noise of this data set to see how this component is recognized by KRR under the influence of noise.

Donoho's doppler data set (figure 2) has an underlying function that has increasing periodicity as x approaches 0. As the noise component increases the overall underlying function becomes entirely hidden as x approaches zero. We will consider two levels of variance in the noise component to demonstrate how sensitive KRR is to a function with such increasing periodicity.

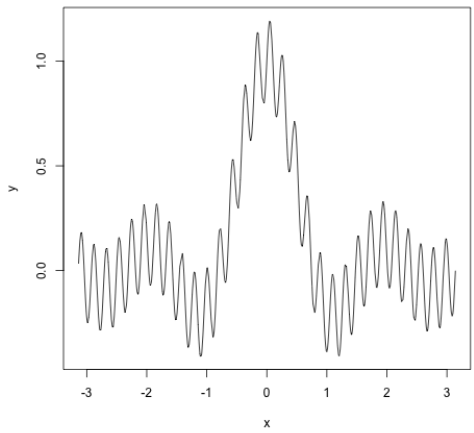

Figure 1: The noisy sinc function

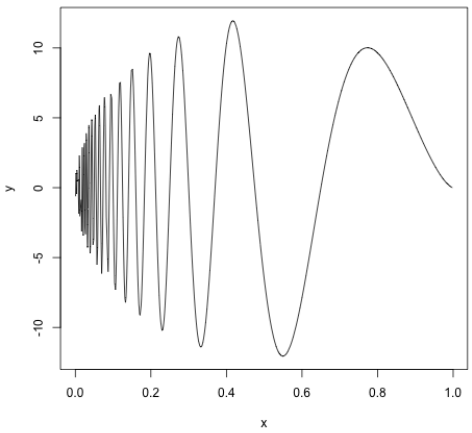

Figure 2: Donoho's doppler function

Feature Space

As discussed above, the solution to KRR lies in the span of k over all combinations of data points. We can think of the eigendecomposition of the kernel matrix $\mathbf{K}$ as approximations to the eigen expansion of the kernel operator K_p. We can also get the implicit feature space by considering

$$h_i = \sqrt{\lambda_i}\hat{\phi}(x) \tag{29}$$

where $\hat{\phi}(x)$ is shown in (20). The σ parameter in the gaussian radial basis kernel function plays a large role in the feature space and fitting the model. We will use cross validation to choose the best σ but let us consider two cases: $\sigma = 5$ and $\sigma = 40$. Figure 3 below shows us that a smaller $\sigma = 5$ has smoother functions and figure 4 shows how the complexity increases for larger $\sigma = 40$. These are essentially the non-linear principal components of the infinite dimensional functional space constructed by the span of the kernel and its parameters. The figures below shows the approximate scaled eigenfunctions that are the basis for the corresponding RKHS associated with the particular kernel and its parameters.

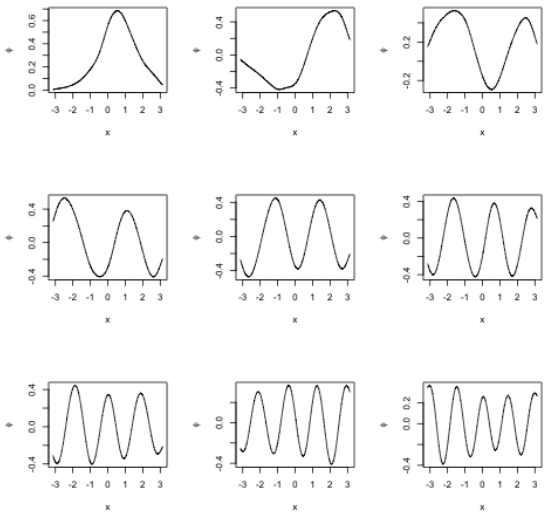

Figure 3: Feature space induced by gaussian radial basis kernel of Noisy Sinc data: $\sigma = .1$ $d = 2$. Gaussian kernel: $\sigma = 5$

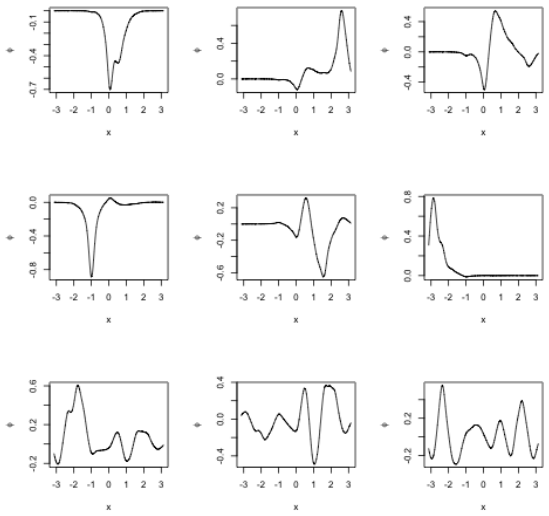

Figure 4: Feature space induced by gaussian radial basis kernel of Noisy Sinc data: $\sigma = .1$ $d = 2$. Gaussian kernel: $\sigma = 40$

Furthermore, we inspect the eigenvalues of the two different sets of eigenfunctions approximated above. We see in figure 5 and 6 that the eigenvalues approach zero, but at a different index of eigenvector. Similar to principal component analysis, a positive definite matrix will have eigenvalues that behave this way. So, for the noisy sinc data with $\sigma = .1$ we have an

effective dimension of 20 for the gaussian kernel with $\sigma = 5$ (figure 5) and dimension of 50 for the gaussian kernel with $\sigma = 40$ (figure 6).

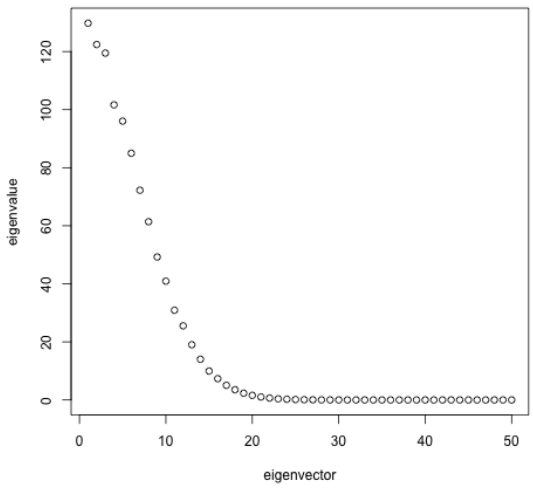

Figure 5: Eigenvalues of the eigen decomposition of Gaussian kernel matrix: $\sigma = 5$

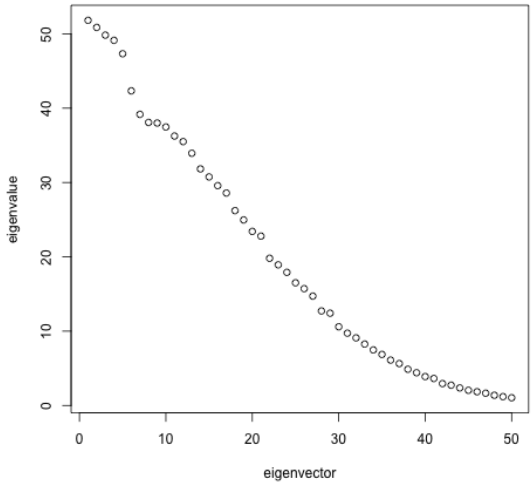

Figure 6: Eigenvalues of the eigen decomposition of Gaussian kernel matrix: $\sigma = 40$

The feature space when using a gaussian radial kernel function is largely determined by the σ parameter. The σ parameter in the gaussian radial kernel function is familiar to us and we know it as the variance of a statistical distribution. In this context the thickness of the tails corresponds to larger weights when computing the inner product of points that are further from a reference point. That is, the larger that σ is for the gaussian kernel, the more globally aware the kernel is when computing the inner product of all points and a reference point. The take away is that varying the kernel-σ parameter adjusts the effective dimension of the feature space in addition to increasing the flexibility by introducing more complex feature elements (demonstrated in figure 4). It is not obvious at this point which σ level to use for the kernel. The next section we will see fitted models for various levels of σ and see how the flexibility increases as the σ parameter for the kernel increases.

Implementation

This section performs the model fitting for various values of σ in the gaussian radial basis kernel. The last section discussed how varying the level of kernel-σ in the kernel affects the dimension of the feature space and affects the complexity of the feature elements. It is not

exactly apparent how kernel-σ will effect the flexibility of the model. The purpose of this section is to examine how flexibility is effected and discuss overfitting and the need for cross validation.

We fit the model for various levels of kernel-σ. Figure 7 displays 15 fitted models on top of the noisy sinc data set with very low noise. The kernel-σ values for Figure 7 range from 4 to 10 by incremements of .4. The kernel-σ values for Figure 8 range from 10 to 80 by incremements of 5. The lowest kernel-σ values are the least flexible fitted lines and the highest level of kernel=σ values are the most flexible. Similarly for Figure 8, we see that higher values of σ fit the local periodicity which demonstrated increased flexibility.

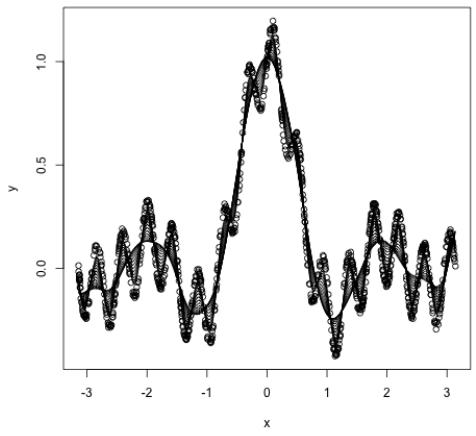

Figure 7: Noisy Sinc: $\sigma = .01$ $d = 1$

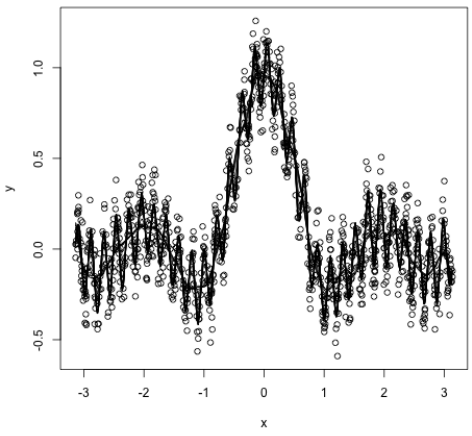

Figure 8: Noisy Sinc: $\sigma = .1$ $d = 2$

Figure 9 and figure 10 also show the effect of increasing the σ parameter for the gaussian kernel. Figure 9 and 10 are the Donoho doppler data set with two levels of noise. The σ values for figure 9 and 10 are generated by $2^{(1:13)}$. The fitted lines increase in flexibility as λ increases. We see that in order to learn an underlying function with such increasing periodicity the value of sigma must get very large. This data set is not a good candidate for cross validation as a larger kernel-σ will always increase the flexibility of the fitted model and decrease the MSE. An interested question to research would be to see at which level of noise would cross validation become advantageous. The next section will consider model selection in the context of varying levels of noise in the noisy sinc data set.

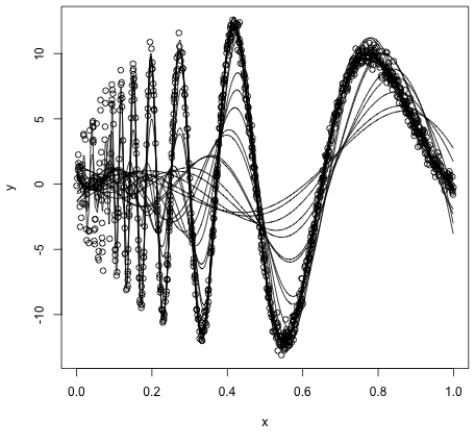

Figure 9: Doppler: $\sigma = .5$

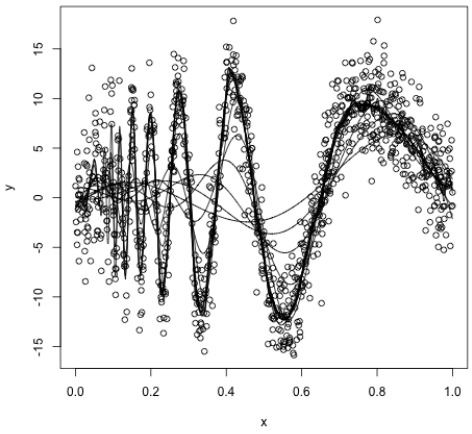

Figure 10: Doppler: $\sigma = 3$

Model Selection

Cross validation is used to fit kernel ridge regression to three different data sets with varying levels of noise. Figure 11 has a noise variance component of $\sigma = .1$, figure 12 has a noise variance component of $\sigma = .2$ and figure 13 has a noise variance component of $\sigma = .3$. AS discussed above, increasing the level of noise begins to hide the second term in the sum of (27). We perform k=10 cross validation to see how the optimal σ parameter for the gaussian kernel varies. Intuitively, we would expect that as noise increases the structure underlying the data set will begin to hide. We created a grid of parameters for the cross validation method considering kernel-σ values in $[1, 300]$ and λ values in $\{.01, .1, .3\}$. The optimal λ value for each data set is .01. The cross validation process took approximately 2 minutes to search over the grid of parameters on a simple notebook computer.

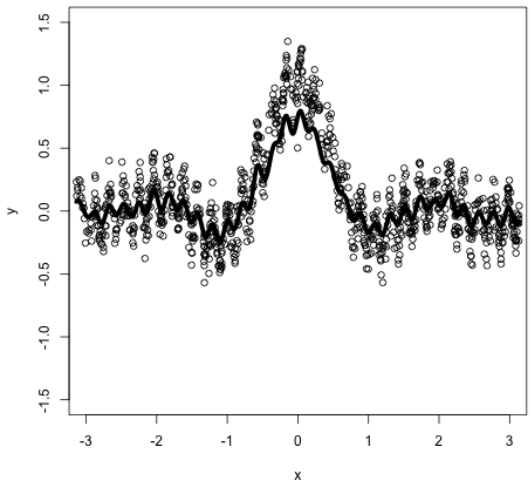

Figure 11: KRR fitted model on noisy sinc data: $\sigma = .1$ $d = 2$

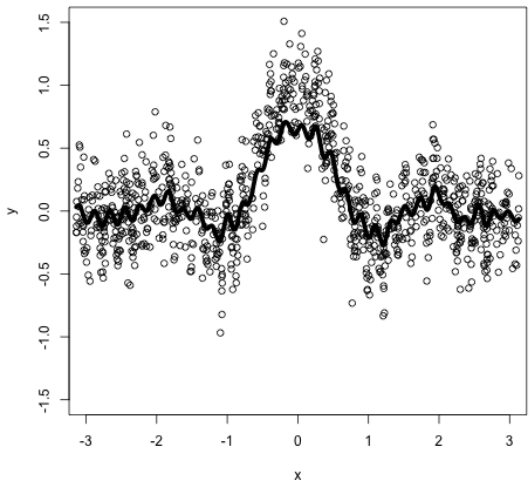

Figure 12: KRR fitted model on noisy sinc data: $\sigma = .2$ $d = 2$

Based on the k-fold cross validation offered in the CVST package the data set in figure 11 has an optimal kernel-σ value of 130 with a corresponding MSE of .030. The underlying structure is still apparent since there is low levels of noise. We see that the fitted model which is plotted on the data is quite flexible and is capable of capturing the local periodicity or sinusoidal behavior.

The optimal kernel-σ parameter for figure 12's data set is a bit lower than figure 11's data set. It has more noise and thus hiding the underlying local sinusoidal behavior. The optimal kernel-σ is 120 with a corresponding MSE of .063. Therefore, their is still moderate flexibility

in the fitted model.

Lastly, the model with the most noise is displayed in Figure 13. The optimal kernel-σ value is 3 with a corresponding MSE of .117. At this level of noise in the data we see that the model prefers a less flexible kernel-σ value. This demonstrates how KRR handles the complexity of underlying processes with significant noise. This is analogous to most of statistics: the underlying structure may be partially hidden by the noise and the best thing we can do is to perform models selection. if we were just given the data set in figure 13 we would not be able to tell if there is an underlying local sinusoidal behavior or excessive amount of noise.

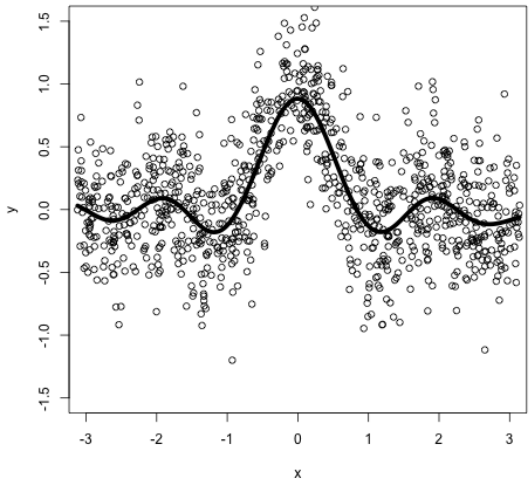

Figure 13: KRR fitted model on noisy sinc data: $\sigma = .3$ $d = 2$

Conclusion

Reproducing kernel Hilbert space theory is largely a result from development in functional analysis. It gives the regularization framework a way of reducing the hypothesis space to a manageable space where solutions exist through finite and computationally tractable methods. An additional benefit of working in a regularization framework is that it has the ability to help control overfitting and ensures that there is a closed form solution for the estimated function. The downside, however, is that it adds a dimension in the parameter space when performing model selection methods.

Kernel ridge regression is a specific instance of the regularization framework under the influence of restricting the hypothesis space using RKHS theory. There is a closed form solution for the estimated function, however, the kernel inversion process requires QR decomposition and it is often advantageous to use faster implementations like the Nystrom methods mentioned earlier. Unlike regular regression the inversion is of a matrix that is $N \times N$ instead of $p \times p$. The huge benefit of kernel ridge regression is the utilization of a kernel matrix to represent the data in a new feature space that intends to find a linear basis of non-linear patterns. The kernel trick in the context of the gaussian radial basis function kernel allows us to implicitly map the data into a infinite dimensional space. The advantage of this implicit map allows us to find any non-linear pattern in the class of continuous functions. This flexibility is demonstrated in the experimental portion of this paper

The data sets considered in the experimental portion of this paper are great at testing non-linear function approximation. The noisy sinc data set demonstrates how components of the underlying function may hide as the noise begins to increase. The data set also demonstrated how KRR is capable of being flexible enough to fit to local sinusoidal behavior is shown in figure 8. The Donoho doppler data set is great at demonstrating the capabilities of non-linear estimation in the context of underlying functions that become increasingly periodic over the domain. Adding noise to the Donono doppler data set essentially hides the underlying function all together. In the context of low levels of noise, KRR becomes increasingly flexible as kernel-σ increases which is shown in figure 10.

This was a brief introduction to KRR covering the basics of the RKHS theory that is necessary to have an intuitive understanding of the mathematical background behind the "kernel" part of the regression. It is shown that KRR is a very flexible non-linear estimation scheme and there are many methods to speed up the process. Further research in kernel logistic regression, kernel discriminant analysis, and gaussian processes is of interest. Specifically, focusing on how different kernels map nonlinear manifolds to linearly separable spaces is of interest.

References

[1] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[2] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[3] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Proceedings of the 14th annual conference on neural information processing systems*, no. EPFL-CONF-161322, pp. 682–688, 2001.

[4] T. Krueger, D. Panknin, M. Braun, *et al.*, "Fast cross-validation via sequential analysis," *Sierra Nevada, Spain*, 2011.

[5] T. Krueger and M. Braun, *CVST: Fast Cross-Validation via Sequential Testing*, 2013. R package version 0.2-1.

[6] J. K. Hunter and B. Nachtergaele, "The spectrum of bounded linear operators," in *Applied Analysis*, pp. 215–243, World Scientific, 2001.

[7] G. Wahba, *Spline models for observational data*, vol. 59. Siam, 1990.

[8] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in computational mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[9] F. Girosi, T. Poggio, and B. Caprile, "Extensions of a theory of networks for approximation and learning: outliers and negative examples," 1990.

[10] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Journal of the american statistical association*, vol. 82, no. 397, pp. 76–89, 1987.

[11] T. Poggio and F. Girosi, "A sparse representation for function approximation," *Neural computation*, vol. 10, no. 6, pp. 1445–1454, 1998.

[12] Z. Liang, *Eigen-analysis of kernel operators for nonlinear dimension reduction and discrimination*. PhD thesis, The Ohio State University, 2014.

[13] M. Kevin, "Machine learning: a probabilistic perspective," 2012.