2013

On the spectral decomposition for kernel machines

Debashis Ghosh, *Penn State University*

On the spectral decomposition for kernel machines

Abstract

Recently, a class of machine learning-inspired procedures, termed kernel machine methods, has been extensively developed in the statistical literature. It has been shown to have large power for a wide class of problems and applications in genomics and brain imaging. Many authors have exploited an equivalence between kernel machines and mixed effects models and used attendant estimation and inferential procedures. In this note, we explore the theoretical foundations of the kernel machine using a spectral decomposition. This leads to simple characterizations of the kernel machine procedure and its bias and variance properties. In addition, we construct a so-called 'adaptively minimax' kernel machine. Such a construction highlights the role of thresholding in the observation space and limits on the interpretability of such kernel machines.

Keywords: Data mining; Decision theory; hard-thresholding; high-dimensional data; nonparametric regression; support vector machines.

1 Introduction

With the availability of massive datasets from scienitifc and medical disciplines, increasing attention is being paid to the use of data mining techniques. This has in turn sparked interest as to the statistical properties of the methodologies. One example is support vector machines (Cristianini and Shawe-Taylor, 2000). This is a supervised learning procedure that attempts to find a margin-maximizing hyperplane that separates two groups. Liu et al. (2007) developed a statistical framework and equivalence in which the support vector machine regression with a continuous outcome is identical to a certain mixed effects model. This equivalence is reviewed in §2. The kernel machine has been utilized heavily with applications to genomics (Liu et al., 2007; 2008; Kwee et al., 2008; Cai et al., 2009; Wu et al., 2010, 2011; Pan, 2009; Kim et al., 2012) and imaging (Ge et al., 2012). These articles have typically show power gains for the kernel machine-based tests relative to their fixed-effects counterparts due to shrinkage brought on by the use of random effects.

Given the recent popularity of the kernel machine methodology, it is important to understand its theoretical foundations. Many of the previous authors have used estimation and attendant inference results from the mixed model framework. In this article, we seek to offer another justification and viewpoint on the kernel machine methodology. This is done using a spectral decomposition and leads to simple characterization of the kernel machine in terms of bias and variance. The decomposition has a very simple geometric characterization and motivates the construction of adaptively minimax kernel machines. The concept of minimaxity is well-studied in statistics and has also been addressed in the context of nonparametric density estimation and regression problems by many authors (e.g., Fan, 1992). In this paper, the adaptively minimax estimator will be very different in structure relatively to previously studied adaptive kernel methods. The structure of the paper is as follows. In Section 2 we review the kernel machine methodology as previously developed in the literature. Section 3 presents the new results on the spectral decomposition of the kernel machines along with bias and variance results. In section 4, we construct adaptive kernel machines using simple thresholding ideas and then prove asymptotic minimaxity results about them. Section 5

concludes with some discussion.

2 Kernel Machine Methodology: Model and Estimation

We first review the kernel machine framework of Liu et al. (2007). For the sake of exposition, we will work in the case that there is no parametric component. For i $(i = 1, \cdots, n)$, we observe $(Y_i, \boldsymbol{X}_i)$, where Y_i is a normally distributed continuous outcome, and $\boldsymbol{X}_i$ is a $p \times 1$ vector of covariates. We assume the following model:

$$Y_i = \beta_0 + h(\boldsymbol{X}_i) + e_i, \tag{1}$$

where β_0 is an intercept term, $h(\boldsymbol{z}_i)$ is an unknown centered smooth function, and the errors e_i are assumed to be independently and identically distributed from a $N(0, \sigma^2)$ distribution. Assume that y_i $(i = 1, \ldots, n)$ are centered so that β_0 drops out of the model (1).

One issue that arises in (1) is how to specify basis functions for h, especially in the case of high-dimensional $\mathbf{X}$. The advantage of kernel methods as defined in machine learning contexts is that one specifies instead a kernel function $K(\mathbf{x}, \mathbf{x}')$ instead of the basis functions. Specifically, a kernel function $K(\mathbf{x}, \mathbf{x}')$ is a bounded, symmetric, positive function satisfying

$$\int K(\mathbf{x}, \mathbf{x}')h(\mathbf{x})h(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \geq 0, \tag{2}$$

for any arbitrary square integrable function $h(\mathbf{x})$ and all $\mathbf{x}, \mathbf{x}' \in R^p$. The kernel function can be viewed as a measure of similarity between two values of the covariate vector $\mathbf{x}$ and $\mathbf{x}'$. Following from the Mercer Theorem (Cristianini and Shawe-Taylor, 2000, p 33), any kernel function satisfying some regularity conditions implicitly specifies an unique function space spanned by a particular set of basis functions (features), and vice versa.

Assume that the nonparametric function $h(\cdot) \in \mathcal{H}_K$, a reproducing kernel Hilbert space (Wahba, 1990). Then there is a 1-1 correspondence between K and the corresponding RKHS. Estimation of $\boldsymbol{\beta}$ and $h(\cdot)$ proceeds by maximizing the scaled penalized likelihood function

$$J(h, \beta) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - h(\mathbf{x}_i)\}^2 - \frac{1}{2}\lambda\|h\|^2_{\mathcal{H}_K}, \tag{3}$$

where λ is a tuning parameter and controls the tradeoff between goodness of fit and complexity of the model. Exploiting a primal/dual equivalence from Karush-Kuhn-Tucker theory, we can

show that the estimator of the nonparametric function $h(\cdot)$ evaluated at the design points $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$ is estimated as

$$\widehat{\boldsymbol{h}} = \lambda^{-1}\boldsymbol{K}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$. In Liu et al. (2007), it was shown that the estimates of h in (4) can be derived as arising from a random effects model of the following form:

$$\boldsymbol{y} = \boldsymbol{h} + \boldsymbol{e}, \tag{5}$$

where $\boldsymbol{h}$ is an $n \times 1$ vector random effects following $\boldsymbol{h} \sim N\{\boldsymbol{0}, \tau\boldsymbol{K}\}$, τ is a scale parameter, and $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Because of this equivalence, all regression parameters in the model can be estimated by maximum likelihood, while the variance component parameters can be estimated by restricted maximum likelihood.

Remark 1. Liu et al. (2007, 2008) used the standard mixed effects model framework for estimation and attendant inference result. While they did not prove asymptotic normality results in that work, one could use Theorems 1 and 2 from Mardia and Marshall (1984) or the work of Cressie and Lahiri (1993) to derive consistency and asymptotic normality results for the kernel machine estimators of the fixed and random effects. Here, we will investigate different properties of the kernel machine relative to those studied by the previous authors.

3 Bias and Variance calculations: some insights

In this section, we characterize the finite-sample bias and variance of the kernel machine estimator for (4). We will begin by assuming that $\lambda > 0$ is fixed. In addition, we assume the model (1) in which we assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are treated as fixed. We will use lower-case notation, i.e. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, to denote this. We can write $\widehat{\boldsymbol{h}}$ from (4) as $\mathbf{H}(\mathbf{I} + \mathbf{H})^{-1}\mathbf{Y}$, where $\mathbf{H} = \lambda^{-1}\boldsymbol{K}$ and $\mathbf{I}$ is the $n \times n$ identity matrix. Thus, we have defined the kernel machine in terms of an operator $\mathbf{P}$ mapping from $\mathbf{R}^n$ to $\mathbf{R}^n$. Because $\mathbf{H}$ is positive definite, the following equivalences hold:

$$\mathbf{H} = \mathbf{U}\mathbf{D}_n\mathbf{U}' \tag{6}$$

$$\mathbf{I} + \mathbf{H} = \mathbf{U}(\mathbf{I} + \mathbf{D}_n)\mathbf{U}' \tag{7}$$

where $\mathbf{U}$ is an $n \times n$ orthonormal matrix such that $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$, and $\mathbf{D}_n$ is a diagonal matrix with entries equalling the eigenvalues of $\mathbf{H}$. Because $\mathbf{H}$ is symmetric and positive definite, all the eigenvalues will be positive. Note that these results are a straightforward application of the singular value decomposition (Golub and Van Loan 1996, Section 2.5). Combining (6) and (7), we have that (4) can be expressed as

$$\widehat{\boldsymbol{h}} \equiv \mathbf{P}\mathbf{Y} = \mathbf{U}\mathbf{D}_n(\mathbf{I} + \mathbf{D}_n)^{-1}\mathbf{U}'\mathbf{Y} \tag{8}$$

We will refer to representation (8) as the spectral decompostion of $\widehat{\boldsymbol{h}}$. The geometric insight given by (8) is quite revealing. The matrix $\mathbf{U}'$ is an orthogonal transformation and hence rotates the data $\mathbf{Y}$ in $n-$dimensional space. In the transformed space, we apply a shrinkage operation given by $\mathbf{D}_n(\mathbf{I} + \mathbf{D}_n)^{-1}$. Since this is a diagonal matrix, univariate shrinkage is applied to each component of $\mathbf{Y}$. Finally, these shrunken variables are back-transformed to the original space. In Figure 1, we illustrate the sequence of transformations with a specific $\mathbf{D}_n$ and $\mathbf{U}$ in two dimensions.

Having such a formula will allow for easy characterization of the bias and variance of the kernel machine estimator. To study these quantities, we will consider the mean squared error (MSE) of the estimator, defined as

$$\text{MSE} = n^{-1}\sum_{i=1}^{n} E\{\widehat{\boldsymbol{h}}(\boldsymbol{x}_i) - h(\boldsymbol{x}_i)\}^2, \tag{9}$$

where the covariates are treated as deterministic, and the expectation inside the summation is taken with respect to the distribution of Y. If the covariates were treated as stochastic, then (9) converges in probability to $E\{\widehat{\boldsymbol{h}}(\boldsymbol{X}) - h(\boldsymbol{X})\}^2$, where $\boldsymbol{X}$ is a new but statistically observation from same distribution as the original design distribution and where the expectation is taken over the training and test sets. We first have the following result.

Theorem 1. Assume the model (1) with attendant estimator given by (4), where $\lambda > 0$ is treated as fixed. Then quantity 9 can be decomposed into a squared bias and variance term, where

$$\text{Bias}^2 = n^{-1}h'\mathbf{U}(\mathbf{I} + \mathbf{D}_n)^{-2}\mathbf{U}'h$$

and

$$\text{Var} = \sigma^2 n^{-1}\mathbf{U}\mathbf{D}_n(\mathbf{I} + \mathbf{D}_n)^{-2}\mathbf{D}_n\mathbf{U}'.$$

Proof: The squared bias is given by

$$\begin{aligned} \text{Bias}^2 &= \{E(\mathbf{PY}) - h\}'\{E(\mathbf{PY}) - h\} \\ &= \{(\mathbf{P} - \mathbf{I})h\}'\{(\mathbf{P} - \mathbf{I})h\}. \end{aligned}$$

It is easily seen using (8) that

$$\begin{aligned} (\mathbf{P} - \mathbf{I}) &= \mathbf{U}\{\mathbf{D}_n(\mathbf{I} + \mathbf{D}_n)^{-1} - \mathbf{I}\}\mathbf{U}' \\ &= \mathbf{UVU}', \end{aligned}$$

where $\mathbf{V} = -(\mathbf{I} + \mathbf{D}_n)^{-1}$. This leads to the result for the squared bias term. For the variance term, we have that

$$\begin{aligned} \text{Var}(\mathbf{PY}) &= \mathbf{PP}' \\ &= \{\mathbf{UD}_n(\mathbf{I} + \mathbf{D}_n)^{-1}\mathbf{U}'\}\{\mathbf{UD}_n(\mathbf{I} + \mathbf{D}_n)^{-1}\mathbf{U}'\}' \\ &= \mathbf{UD}_n(\mathbf{I} + \mathbf{D}_n)^{-1}\mathbf{U}'\mathbf{U}(\mathbf{I} + \mathbf{D}_n)^{-1}\mathbf{D}_n\mathbf{U}', \end{aligned}$$

which upon simplification leads to the final result in the theorem.

Theorem 1 reveals a very simple tradeoff between the bias and the variance that is similar to the standard one seen in the nonparametric regression literature. Note that the result does not depend on the choice of kernel. It is also exact and is valid for any sample size. This theorem also shows clearly the role and importance of the eigenvalues of $\lambda^{-1}\mathbf{K}$. They appear in two locations in the variance, compared to one location for the bias. The following corollary is immediate.

Corollary 1. Suppose the conditions of Theorem 1 can be assumed. Then the following hold:

(a). If the eigenvalues of $\lambda^{-1}\mathbf{K}$ are all greater than one, then the variance will be greater than the squared bias.

(b). If the eigenvalues of $\lambda^{-1}\mathbf{K}$ are all between zero and one, then the squared bias will be greater than the variance.

Remark 2. In Theorem 1, we assumed that λ, the smoothing parameter was treated as a fixed but known quantity. In practice, this parameter has to be estimated. While Liu et al.

(2007) proposed using restricted maximum likelihood to estimate it, other options include generalized cross-validation (Golub et al., 1979). If we now allow $\mathbf{P}$ to depend upon this estimator, say $\hat{\lambda}$, then Theorem 1 no longer directly applies. However, it holds conditional on $\hat{\lambda}$. If one can assume in a functional sense that $\mathbf{P}$ is stochastically equicontinuous (Pollard, 1984) in large samples with respect to λ, then the result of Theorem 1 will hold in an approximate sense.

Remark 3. In practice, the kernel matrix will also include parameters that have to be estimated. A canonical example of such a kernel is the Gaussian kernel. The formula for the Gaussian kernel is given by $K(\boldsymbol{z}, \boldsymbol{z}'; \rho) = \exp\{-||\boldsymbol{z} - \boldsymbol{z}'||^2/\rho\}$, where $||\boldsymbol{z} - \boldsymbol{z}'|| = \sum_{k=1}^{p}(z_k - z'_k)^2$. The Gaussian kernel generates the function space spanned by radial basis functions, whose nice properties can be found in Buhmann (2003). The function space determined by the Gaussian kernel is infinite-dimensional. The result of Theorem 1is again conditional on an estimate of ρ. One could argue as in Remark 2 that the theorem holds asymptotically under conditions regarding the dependence of $\mathbf{P}$ on ρ.

4 Shrinkage, Decision-theoretic framework and minimaxity

In this section, we describe two shrinkage interpretations of kernel machines. The first relies on the fact that we can write the scaled kernel matrix as

$$\lambda^{-1}\mathbf{K} = \mathbf{H}\mathbf{H}',$$

where $\mathbf{H}$ is a matrix of the same rank as $\mathbf{K}$. Then we can write $\hat{h}$ as

$$\begin{aligned}
\hat{h} &= \lambda^{-1}\mathbf{K}(\mathbf{I} + \lambda^{-1}\mathbf{K})^{-1}\mathbf{Y} \\
&= \mathbf{H}\mathbf{H}'(\mathbf{I} + \mathbf{H}\mathbf{H}')^{-1}\mathbf{Y} \\
&= \mathbf{H}\mathbf{H}'(\mathbf{I} - \mathbf{H}(\mathbf{I} + \mathbf{H}'\mathbf{H})^{-1}\mathbf{H}')\mathbf{Y} \\
&= \lambda^{-1}\mathbf{K}\mathbf{Y} - \mathbf{H}\mathbf{A}(\mathbf{I} + \mathbf{A})^{-1}\mathbf{H}'\mathbf{Y}. \qquad (10)
\end{aligned}$$

Notice that going from second to the third equality above required using the Woodbury matrix identity, a nice review for which can be found in Hager (1989). The representation (10) shows that the kernel machine consists of two pieces. The first term is effectively the correlation

between $\lambda^{-1}\mathbf{K}$ with the response vector. The second term consists of "transforming" the data $\mathbf{Y}$ by the 'matrix square root' of $\lambda^{-1}\mathbf{K}$, followed by shrinkage by $\mathbf{A}(\mathbf{I}+\mathbf{A})^{-1}$, followed by untransforming using the matrix square root. Using (10), one interpretation of the kernel machine is as a correlation between the kernel matrix and the response, shrunk by the second factor

We now start with equation (8) and consider it in more detail. It can be rewritten as $\mathbf{USU}'\mathbf{Y}$, where $\mathbf{S} = \mathbf{D}_n(\mathbf{I}+\mathbf{D}_n)^{-1}$. Let us now consider the $n-$dimensional vector of observations $\mathbf{V} = \mathbf{U}'\mathbf{Y}$. Recall that we are conditioning on $\mathbf{X}$ throughout the paper. We can then consider the following model for the components of $\mathbf{V}$:

$$V_i = \mu_i + \epsilon_i \tag{11}$$

for $i = 1, \ldots, n$, where ϵ_i are a random sample from $N(0, \sigma^2)$, V_i is the ith component of $\mathbf{V}$ and $\mu_i = \mathbf{u}_i'\mathbf{h}(\mathbf{X})$, with $\mathbf{u}_i$ denoting the ith column of $\mathbf{U}$. While at first glance this model appears to be restrictive, if we let n approach infinity, then this model is identical to the classical nonparametric regression model first considered in Pinsker (1980). Note that conditional on $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and λ, we can view μ and $\mathbf{h}$ interchangeably. The kernel machine provides one class of estimators for h, which induces an estimator $\hat{\mu}$.

An elegant decision-theoretic formulation for models of the form (11) has been developed by Donoho and Johnstone (1994). We follow that approach here. Assuming that (11) holds, we consider the L^q minimax risk functional for an estimator μ^E, defined as

$$R(\mu^E) = \inf_{\mu^E} \sup_{\Theta_{p,n}} E_\mu \sum_{i=1}^{n} |\mu_i^E - \mu_i|^q, \tag{12}$$

where $\Theta_{p,n} \equiv$ denotes a ball of radius r for the parameter vector μ. We now seek to understand the behavior of R in (12) as n approaches infinity. In our framework, we treat σ and r as known functions of n. In addition, we define the range of (p, q) to be $(0, \infty] \times [1, \infty)$ so that the results presented here generalize to many types of loss functions.

We reiterate again that the spectral decomposition (8), along with conditioning on covariates and λ, allows for an equivalence between the kernel machine estimator with the Donoho-Johnstone framework. Application of their work yields the following results in our setting.

Theorem 2: Define the $n \times n$ diagonal matrices $\tilde{\mathbf{S}} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$ and $\mathbf{S}^* = \text{diag}(d_1^*, \ldots, d_n^*)$, where

$$\tilde{d}_i = \text{sign}(V_i)(|V_i| - \lambda\sigma)_+,$$

$$d_i^* = V_i I(|V_i| \geq \lambda),$$

with $(a)_+ = \max(a, 0)$. Setting $\tilde{\mu} \equiv \tilde{\mathbf{S}}\mathbf{U}'\mathbf{Y}$ and $\mu^* = \mathbf{S}^*\mathbf{U}'\mathbf{Y}$, we have the following:

(a). If $0 < p < 2$, $n\sigma^p \to \infty$ and $\sigma^2 \log n\sigma^p \to 0$, then as $n \to \infty$,

$$\frac{R(\hat{\mu})}{R(\tilde{\mu})} = (1 + n\sigma^2)^{-1} n\sigma^p (2 \log n\sigma^p)^{-1+p/2}(1 + o(1)). \tag{13}$$

(b). If $\lambda^2 = 2 \log n\sigma^p + \alpha \log(2 \log n\sigma^p)$ for $\alpha > p - 1$, then as $n \to \infty$,

$$\frac{R(\hat{\mu})}{R(\mu^*)} = (1 + n\sigma^2)^{-1} n\sigma^p (2 \log n\sigma^p)^{-1+p/2}(1 + o(1)). \tag{14}$$

Proof: The result follows from representation 11 and application of Theorem 3 and Corollary 4 of Donoho and Johnstone (1994).

We can construct kernel machines given as $\tilde{h} = \mathbf{U}\tilde{\mu}$ and $h^* = \mathbf{U}\mu^*$. Note that the risk quantities for h analogous to (12) will be the same as risk quantities for estimators of μ due to invariance under orthogonal matrices. We note that the estimators being constructed in Theorem 1 involve thresholding estimators as given by the definition of $\tilde{d}_i$ and d_i^*, $i = 1, \ldots, n$. We refer to $\tilde{h}$ nad h^* as the soft-thresholding and hard-thresholding kernel machines, respectively. One implication of Theorem 2 is that the minimax risk of $\hat{h}$, which involves a linear shrinkage estimator after transformation by $\mathbf{U}$, is asymptotically dominated by that of the soft- and hard-thresholding kernel machines as the sample size approaches infinity. However, another insight from Theorem 2 is that for kernel machines, the beneficial effects of sparsity operate in the space of the response variable (i.e., Y). This is very different from penalized regression estimation procedures (Tibshirani, 1996; Fan and Li, 2001), where sparsity operates on the regression coefficients in the predictor space. Thus, asymptotic minimaxity for kernel machines requires sparsity/thresholding in a space that is less interpretable. While it is easy to interpret a zero of a regression coefficient as the corresponding variable not being important in a regression, it is unclear what a zeroing of an individual observation means. The

latter is required to construct kernel machines with desirable risk properties. However, we note in passing that a similar phenomenon aappears to happen with support vector machines (Cristianini and Shawe-Taylor, 2000). Such a link deserves further exploration.

Next, we wish to expand the notion of the minimax risk considered in (11) to include nonlinear estimators. Now we define

$$R_N^* \equiv R_N^*(\sigma, \Theta_{p,n}(r)) = \inf_{\mu^E} \sup_{\Theta_{p,n}(r)} E_\mu \sum_{i=1}^n |\mu_i^E - \mu_i|^q, \tag{15}$$

where the subscript N denotes that nonlinear procedures are allowed. We conclude with an oracle risk inequality for the thresholded kernel machines in terms of the minimax risk defined in (15).

Theorem 3: There exist constants $\tilde{C}, C^* \geq 1$ such that if (i) $p \geq q$ or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then

(a). The following bound holds for $\tilde{\mu}$:

$$\inf_{\lambda} \sup_{\mu \in \Theta_{p,n}(r)} E_\mu |\tilde{\mu} - \mu|^q \leq \tilde{C} R_N^*(\sigma, \Theta_{p,n}(r))(1 + o(1)).$$

(b). The following bound holds for μ^*:

$$\inf_{\lambda} \sup_{\mu \in \Theta_{p,n}(r)} E_\mu |\mu^* - \mu|^q \leq C^* R_N^*(\sigma, \Theta_{p,n}(r))(1 + o(1)).$$

Proof: We will prove part (a) of the theorem; part (b) will follow by a similar argument. We begin by defining the following quantity:

$$R_B^*(\sigma, \Theta(r)) = \inf_{\mu^E} \sup_{\pi} \{E_\pi E_{\mu^E} |\mu^E - \mu|^q : \pi \ \text{ s.t.} E_\pi \sum_{i=1}^n |\mu_i|^p \leq r\}. \tag{16}$$

This quantity represents the Bayes risk for estimation of μ with respect to prior distributions π satisfying the moment conditions define in (16), which are effectively constraints on the pth order moments. If we restrict attention to coordinatewise estimators of μ, then we have the following result, along the lines of Proposition 9 in Donoho and Johnstone (1994):

$$R_B^*(\sigma, \Theta_{n,p}(r)) = n\rho(rn^{-1/p}, \sigma),$$

where

$$\rho(\tau, \sigma) = \inf_{\sigma} \sup_{F} \{E_F E_\mu |\delta(v) - \mu|^q : F \in \mathcal{F}_P(\tau)\},$$

with $\delta(v)$ representing a univariate esimator, and $\mathcal{F}_P(\tau)$ denoting the set of distribution functions satisfying $\int |\mu|^p F(d\mu) \leq \tau^p$. We can define quantities analogous to (16) for soft- and hard-thresholding estimators, say R_s^* and R_h^*. Then using arguments similar to those in Section 3 of Donoho and Johnstone (1994), we have that $R_s^*(\tau, \sigma) = \sigma^q \rho_s(\tau/\sigma, 1)$ and $R_h^*(\tau, \sigma) = \sigma^q \rho_h(\tau/\sigma, 1)$ where

$$\rho_s(\tau, \sigma) = \inf_{\sigma} \sup_{F} \{E_F E_\mu |\delta_s(V_1) - \mu|^q : F \in \mathcal{F}_P(\tau)\},$$

and

$$\rho_h(\tau, \sigma) = \inf_{\sigma} \sup_{F} \{E_F E_\mu |\delta_h(V_1) - \mu|^q : F \in \mathcal{F}_P(\tau)\},$$

where the subscripts s and h denote soft- and hard-thresholding. We now define $\tilde{C}$ and C^* as $\tilde{C} = \sup_{\tau,\sigma} \rho_s(\tau, \sigma)/\rho(\tau, \sigma)$ and $C^* = \sup_{\tau,\sigma} \rho_h(\tau, \sigma)/\rho(\tau, \sigma)$. Then by the arguments in Sections 3 and 6 of Donoho and Johnstone (1994),

$$\begin{aligned}
\inf_{\lambda} \sup_{\mu \in \Theta_{p,n}(r)} E\|\tilde{\mu} - \mu\|^q &\leq R_s \\
&= n\rho_s(rn^{-1/p}, \sigma) \\
&\leq \tilde{C} n \rho_s(rn^{-1/p}, \sigma) \\
&= \tilde{C} R_B^*(\sigma, \Theta_{n,p}(r)) \\
&= \tilde{C} R_N^*(\sigma, \Theta_{n,p}(r))(1 + o(1)).
\end{aligned}$$

A similar argument holds for the hard-thresholding estimators.

The primary implication of Theorem 3 is that from an asymptotic minimaxity point of view, the thresholded kernel machines represent 'near-optimal' estimators for the function h.

5 Discussion

In this note, we have investigated the properties of kernel machine methods using a spectral decomposition. Using this decomposition makes explicit the operations that are needed for construction of a kernel machine. The first step is projection onto an orthogonal basis function space of the same dimension as the sample size, followed by a shrinkage operator applied to the transformed observations followed by a back-transformation. Such a decomposition opens up the range of methods for constructing kernel machines. However, another point to observe

is that the kernel machine operates in the $n-$dimensional space of the response variable. We have constructed a near-optimal kernel machine in an asymptotically minimax sense. However, such a kernel machine involves thresholding of observations and not of predictor variables in the way that most modern regression methods. This speaks to a limitation of kernel machine estimators, namely their interpretability. An analogous concept arises in principal components analysis, when one wishes to interpret the individual components. Because the components are linear combinations of the original variables, they have limited interpretability. While the scale of the kernel machine computations does not depend on p, this comes at the cost of limitations in interpretation.

6 References

Buhmann, M.D. (2003). *Radial Basis Functions.* Cambridge University Press, Cambridge, UK.

Cai, T., Tonini, G., and Lin, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics* **67**, 975-86.

Cressie, N. and Lahiri, S. N. (1993). The Asymptotic Distribution of REML Estimators. *Journal of Multivariate Analysis* **45**, 217 – 233.

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

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probability Theory and Related Fields* **99**, 277 – 303.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association* **87**, 998 – 1004.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M. and Nichols, T. E. (2012). Increasing power

for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *Neuroimage* **63**, 858 – 873.

Golub, G. and van Loan, C. F. (1996). *Matrix Computations*, 3rd ed. Baltimore: Johns Hopkins University Press.

Golub, G., H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215 – 223.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Review* **31**, 221 – 239.

Kim, I., Pang, H. and Zhao, H. (2012). Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. *Statistics in Medicine* **31**, 1633 – 1651.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D. and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics* **82**, 386 – 397.

Liu, D., Ghosh, D. and Lin, X. (2008). Assessing the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via mixed models. *BMC Bioinformatics* **9**, 292.

Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. *Biometrics* **63**, 1079 – 1088.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135 – 146.

Pan W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology* **35**, 211 – 216.

Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission* **16**, 52 – 68.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**, 267 – 288.

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

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86**, 929 – 942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* **89**, 82 – 93.