Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation

G. de los Campos,* D. Gianola,*‡ and G. J. M. Rosa†

*Departments of Animal Sciences, †Dairy Science, and ‡Biostatistics and Medical Informatics, University of Wisconsin, Madison 53706

ABSTRACT: Reproducing kernel Hilbert spaces (RKHS) methods are widely used for statistical learning in many areas of endeavor. Recently, these methods have been suggested as a way of incorporating dense markers into genetic models. This note argues that RKHS regression provides a general framework for genetic evaluation that can be used either for pedigree- or marker-based regressions and under any genetic model, infinitesimal or not, and additive or not. Most of the standard models for genetic evaluation, such as infinitesimal animal or sire models, and marker-assisted selection models appear as special cases of RKHS methods.

Key words: animal model, dense marker, marker-assisted selection, reproducing kernel Hilbert spaces, sire model

INTRODUCTION

Statistical models based on reproducing kernel Hilbert spaces (RKHS) have been useful for regression (e.g., Wahba, 1990), classification (e.g., Vapnik, 1998), and smoothing in highly dimensional problems such as in Vert and Kanehisa (2003). Recently, Gianola et al. (2006) and Gianola and van Kaam (2008) suggested using this approach for incorporating information on dense whole-genome markers into models for prediction of genetic values of animals for quantitative traits. Further, González Recio et al. (2008) used this approach to infer genetic values for early mortality rate in broilers. On the other hand, BLUP (Henderson, 1975) of breeding values is ubiquitous in animal breeding, and has gained attention in plant breeding recently (Piepho et al., 2007).

This note shows that RKHS regression provides a framework for genetic evaluation of quantitative traits that can be used to incorporate information on pedigrees, markers, or any other ways of characterizing the genetic background of individuals that may become available in the future. Some of the most important models of quantitative genetics [e.g., pedigree-based infinitesimal models (see Fisher 1918; Wright, 1921)], as well as some marker-assisted selection (MAS) models (e.g., Fernando and Grossman, 1989), are special cases of RKHS regression. To illustrate this fact, reference is made to an animal model for additive effects (e.g., Quaas and Pollak, 1980). However, all results extend to other models, such as sire models (e.g., Henderson, 1975), models for maternal effects (e.g., Henderson and Quaas, 1976), and multilocus models for genetic effects attributable to dominance or epistasis (e.g., Cockerham, 1954; Kempthorne, 1954), and to some models for MAS (e.g., Fernando and Grossman, 1989).

METHODS

RKHS Regression Models

In the literature on splines (e.g., Wahba, 1990) and in computer science (e.g., Vapnik, 1998), the method of RKHS is motivated by using the theory of Hilbert spaces of functions (e.g., Akhiezer and Glazman, 1963). Alternatively, the Bayesian view of RKHS regressions can be used to motivate this methodology by using Gaussian processes (e.g., Rasmussen and Williams, 2006). A brief and elementary description of the RKHS regression framework is provided below.

A standard regression problem involves a response variable, $y_i$, and some information set, $t_i$, ($i = 1, 2, \ldots$,
n), used to construct a predictor of the response. The information set consists of, for example, vectors of covariates (i.e., $t_i = x_i$); a pedigree, in which case $t_i = (s_i, d_i)$, where $(s_i, d_i)$ indicates the sire and dam of the $i$th subject; or information about several genetic markers (e.g., SNP). In a regression model, $y_i$ is viewed as the sum of a conditional expectation function, $f_i = E(y_i \mid t_i)$, and a model residual $\varepsilon_i$, that is,

$$y_i = f_i + \varepsilon_i.$$

Above, $f$ maps from an information set $T = \{t_1, \ldots, t_n\}$ to evaluations of a conditional expectation function, $E(y_i \mid T)$. For finite $n$, and without making parametric assumptions, $f$ can be viewed as an $n$-dimensional vector, $f = (f_1, \ldots, f_n)'$, where each entry gives the evaluation of the function $f$ at a point in $T$. Standard estimators, such as maximum likelihood or least squares, are based on minimization of some “loss” function, that is,

$$\hat{f}(T, i) = \arg\min_{f} l(y, f; T),$$

here, $l(y, f, T)$ is typically some measure of goodness of fit. The optimization problem is solved over the space of functions. If no further restrictions are imposed, the above problem is ill-posed and the solution is a function that passes through the data.

Parametric methods impose restrictions on the form of $f$, by expressing it in a parametric form, say $f_{\theta}$, where $\theta$ is a vector of unknowns. However, in many instances, there is no theory for specifying such a function, and it may be difficult to choose a parametric specification balancing fitness and complexity optimally. The “regularization” class of statistical methods (to which RKHS models belong) addresses the problem in a different way: $f$ is kept general and a penalty is added to the objective function of the optimization problem. This penalty, denoted as $J(f)$, measures model complexity, in some sense. The optimization problem becomes

$$\hat{f}(T, i) = \arg\min_{f} \{l(y, f; T) + \lambda J(f)\},$$

where $\lambda$ is a parameter used to control the trade-off between model goodness of fit and complexity, as measured by $J(f)$.

The method of RKHS regression is a special case of [1], and it is obtained by assuming that $f$ belongs to a Hilbert space of real-valued functions, denoted as $f \in H$, and by using the square of the norm of $f$ in that space as penalty. Then, $J(f) = \|f\|_H^2$, where $\|f\|_H$ denotes the norm in Hilbert space $H$. For a detailed account of RKHS, the reader is referred to Akhiezer and Glazman (1963); a more amenable treatment is in Shawe-Taylor and Cristianini (2004). The RKHS regression problem can then be expressed as

$$\hat{f}(T) = \arg\min_{f \in H} \left\{l(y, f; T) + \lambda \|f\|_H^2 \right\}.$$  \[2\]

Model specification in RKHS regression refers to the choice of the loss function, $l(\cdot)$, the Hilbert space, $H$, and the smoothing parameter, $\lambda$. Standard choices of loss function are the negative of the log-likelihood, $l(y, f) = -\log p(y \mid f)$, and the residual sum of squares, $(y - f)'(y - f)$. If the response is a binary outcome, coded as $y_i \in \{-1, 1\}$, and the loss function is taken to be a hinge function, the above problem is the standard support vector machine (e.g., Vapnik, 1998). The smoothing parameter $\lambda$ can be chosen by cross-validation or generalized cross-validation (e.g., Craven and Wahba, 1979), or by Bayesian methods (e.g., Gianola and van Kaam, 2008).

The choice of $H$ is central in RKHS regression models; it defines the space of functions over which the search of $f$ is performed, and, because Hilbert spaces are normed spaces (e.g., Akhiezer and Glazman, 1963), by choosing $H$, one automatically defines its norm, and, therefore, $J(f) = \|f\|_H^2$. Each RKHS is uniquely associated with a positive-definite function, or reproducing kernel (RK), and vice versa. Because of this, choosing an RK is tantamount to defining $H$. The duality between Hilbert spaces of functions and positive-definite functions is convenient, because it is easier to define a positive definite function on $T$ than to define $H$ explicitly.

Let $K = \{K(t_i, t_j)\}$ be an $n \times n$ positive definite matrix with elements $K(t_i, t_j)$, which are evaluations of a RK, and $l(y, f) = (y - f)'(y - f)$ be a residual sum of squares. Under this setting, Kimeldorf and Wahba (1970) showed that the optimization problem [2] can be expressed equivalently as

$$\min_{c} \left\{(y - K\hat{c})'(y - K\hat{c}) + \lambda c'Kc\right\}. \hspace{1cm} [3]$$

Above, $c$ is an $n \times 1$ vector of unknown constants. The first-order-conditions of problem [3] are

$$\hat{c} = K'K + \lambda K'^{-1}c = K'y.$$

Further, since $K = K'$ and $K^{-1}$ exists, premultiplication by $K^{-1}$ yields

$$\hat{c} = K'y = P(K, \lambda)y,$$

where $P$ is an identity matrix of appropriate dimension. The estimated conditional expectation function is

$$\hat{f} = K\hat{c} = K[K + \lambda I]^{-1}y = P(K, \lambda)y,$$
where \( \mathbf{P} = \mathbf{K}[\mathbf{K} + \lambda \mathbf{I}]^{-1} \) is a projection matrix, or averaging operator. Therefore, \( \mathbf{f} \) is a weighted sum of the observations,

\[
\hat{f}_i = \sum_{j=1}^{n} w_{ij} y_j,
\]

where the weights, \( w_{ij} \), are the entries of \( \mathbf{P} = \{w_{ij}\} \) that are defined by the kernel, the information set \( T \), and the value of the smoothing parameter.

The solution to optimization problem [3] can be interpreted as a conditional (given \( \lambda \)) posterior mean and mode of a Bayesian model with a Gaussian likelihood and a normal prior for “regression coefficients” \( \mathbf{c} \). To see this, let the equation for the data be

\[
\mathbf{y} = \mathbf{Kc} + \mathbf{e},
\]

where \( \mathbf{y} = (y_1, \ldots, y_n)' \) and \( \mathbf{e} = (e_1, \ldots, e_n)' \) are vectors of data and model residuals, respectively, and the kernel matrix, \( \mathbf{K} (n \times n) \), is viewed as an incidence matrix for the vector of regression coefficients \( \mathbf{c} (n \times 1) \). If the residuals satisfy \( \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e) \), the likelihood function is \( N(\mathbf{y}|\mathbf{Kc}, \mathbf{I}\sigma^2_e) \). Now, assume that the prior distribution of regression coefficients is \( p(\mathbf{c}) = N(\mathbf{c}|\mathbf{0}, \mathbf{K}^{-1}\sigma^2_c) \). If \( \sigma^2_e \) and \( \sigma^2_c \) are known, the density of the conditional posterior distribution of \( \mathbf{c} \) is

\[
p(\mathbf{c}|\mathbf{K},\sigma^2_e,\sigma^2_c,\mathbf{y}) \propto \exp \left( -\frac{1}{2\sigma^2_e} (\mathbf{y} - \mathbf{Kc})' (\mathbf{y} - \mathbf{Kc}) \right) \exp \left( -\frac{1}{2\sigma^2_c} \mathbf{c}'\mathbf{Kc} \right).
\]

This density is known to be multivariate-normal, with a mean (mode) equal to the solution to [4]. From the Bayesian perspective, the smoothing parameter is interpreted as a ratio of variance components, that is,

\[
\lambda = \frac{\sigma^2_e}{\sigma^2_c}.
\]

Large values of \( \lambda \) shrink the solution to [4] toward zero, whereas small values lead to relatively low shrinkage.

**The Animal Model**

This model is reviewed for the case in which each animal has a single record, and assuming that data have been centered. With this setting, the model is defined by the following assumptions,

**Animal model:**

\[
\begin{align*}
\mathbf{y} &= \mathbf{u} + \mathbf{e} \\
\mathbf{e} &\sim N(\mathbf{0}, \mathbf{I}\sigma^2_e) \\
\mathbf{u} &\sim N(\mathbf{0}, \mathbf{A}\sigma^2_u) \\
\end{align*}
\]

where \( \mathbf{u} = (u_1, \ldots, u_j)' \) is a vector of additive genetic effects, and \( \mathbf{y} \) and \( \mathbf{e} \) are as before. Above, \( \mathbf{A} = \{a(i, j)\} \) is known as the matrix of numerator-relationships, and its entries are evaluations of a positive-definite function, additive-relationships, on a pedigree. This (co)variance structure is derived from genetic theory, and it is the expected (given the pedigree) (co)variance structure under an infinitesimal additive mode of gene action (e.g., Fisher, 1918; Wright, 1921). The posterior mean and mode in this model (assuming that variance components are known) is the solution to the equations:

\[
\left[ \mathbf{I} + \frac{\sigma^2_e}{\sigma^2_u} \mathbf{A}^{-1} \right] \hat{\mathbf{u}} = \mathbf{y}.
\]

The above system is a special case of the mixed model equations of Henderson, and the solution is known to be BLUP.

**RESULTS AND DISCUSSION**

**RKHS View of the Animal Model**

The connection between BLUP and RKHS regression estimates was recognized long ago. Kimeldorf and Wahba (1970) discussed the correspondence between Bayesian models and smoothing by splines (a special case of RKHS regressions). Harville (1983) pointed out that the problems of prediction, estimation, interpolation, smoothing, and filtering could be unified into a single framework: that of BLUP. A more recent recognition is in a discussion of Robinson’s (1991) review by Speed (1991). To see this connection, recall that a Bayesian RKHS model is defined by the following assumptions:

**RKHS:**

\[
\begin{align*}
\mathbf{y} &= \mathbf{Kc} + \mathbf{e} \\
\mathbf{e} &\sim N(\mathbf{0}, \mathbf{I}\sigma^2_e) \\
\mathbf{u} &\sim N(\mathbf{0}, \mathbf{A}\sigma^2_u) \\
\end{align*}
\]

Now, consider the change of variable \( \mathbf{u} = \mathbf{Kc} \). Because \( \mathbf{c} \) is assumed to be normal, so is \( \mathbf{u} \). Further, \( E(\mathbf{u}) = E(\mathbf{Kc}) = \mathbf{K} E(\mathbf{c}) = \mathbf{0} \), and \( \text{Cov}(\mathbf{u}) = \text{Cov}(\mathbf{Kc}) = \mathbf{K} \text{Cov}(\mathbf{c}) \mathbf{K}' = \mathbf{K} \sigma^2_c \). After the change of variable the model becomes

\[
\begin{align*}
\mathbf{y} &= \mathbf{u} + \mathbf{e} \\
\mathbf{e} &\sim N(\mathbf{0}, \mathbf{I}\sigma^2_e) \\
\mathbf{u} &\sim N(\mathbf{0}, \mathbf{A}\sigma^2_u) \\
\end{align*}
\]

If the information set \( T \) is a pedigree and the kernel is taken to be \( K(i, j) = a(i, j) \), the model above is the standard additive genetic animal model, and \( \sigma^2_u \) is interpreted as the additive genetic variance. The 2 parameterizations lead to different sets of mixed-model equations, as shown in Table 1.
The fact that the animal model appears as a special case of RKHS regressions provides a semiparametric view of this model. This naturally extends to other pedigree-based infinitesimal models (e.g., infinitesimal models for effects due to dominance or epistasis) and to classical MAS models, where both marker and pedigree information are used to arrive at some prior (co)variance structure (e.g., Fernando and Grossman, 1989).

The Role of the Reproducing Kernel

From the point of view of the theory of RKHS, one can use kernels to construct a Hilbert space of functions, and the learning problem consists of choosing a function within the space defined by the kernel. This is, however, a rather abstract view of the role of the kernel. Parameterization II in Table 1 provides a more concrete view of the role of the kernel: $K(i, j)$ represents the expected (co)variance structure of the evaluations of the conditional expectation function at the points in the input space, that is, $\text{Cov}(f_i, f_j) \propto K(i, j)$. At the same time, this (co)variance function defines a notion of smoothness of the conditional expectation function with respect to the input space. The choice between kernels can simply be stated as a model comparison problem, or, as discussed later, can be formulated as an estimation problem.

Generality of RKHS

The RKHS provides a general and flexible framework for prediction of genetic values for at least 2 reasons:

1. The methodology can be used regardless of the nature of the information set ($T$): pedigree, sparse markers, SNP, or any other method of characterizing the genome that may become available in the future. The only requirement is for one to be able to define one or several positive definite functions evaluated in $T$. This generality is important in a context in which techniques available for characterizing the genome progress rapidly.

2. A second convenient fact is that the same framework can be used in the context of theory-based models (e.g., animal model or MAS models), or, alternatively, as a semiparametric tool in which case positive definite functions are chosen based on some property (e.g., good predictive ability).

Interpretability of Predictions from RKHS Models

What is the interpretation of predictions from RKHS regressions? The answer is simple: if a kernel expresses the (co)variance structure implied by a given theory, then estimates are interpretable in terms of that theory. Otherwise, no clear interpretation can be attached to predictions. To illustrate, consider a model that uses additive relationships, $a(i, j)$, as kernel and a competitive model that uses some other kernel $K(i, j)$. From theory of RKHS, each of these kernels is associated with a space of functions, $f_a \in H_a$ and $f_K \in H_K$. Depending on the form of $K(i, j)$, several situations may occur. For instance, if $H_a \subseteq H_K$ (i.e., $H_K$ includes $H_a$), then predictions using $K(i, j)$ will capture additive effects and some nonadditive effects as well. However, if $H_a \not\subseteq H_K$, predictions from a model that uses $K(i, j)$ as a kernel will not fully capture additive effects.

Building Kernels

Selecting a kernel is the most critical stage in applying kernel-based algorithms (Shawe-Taylor and Cristianini, 2004). Prior knowledge about a problem may not be enough for choosing a specific kernel. Kernels can be selected using model comparison techniques, or, alternatively, the uncertainty about the kernel can be formulated as an estimation problem. In this case, the algorithm has 2 tasks: choosing a kernel and a pattern. For example, one can index a kernel with one or several parameters, and these parameters may be chosen in some optimal way (e.g., cross-validation, generalized cross-validation, or Bayesian methods). A standard method of generating a family of kernels is to introduce bandwidth parameters that control how fast the (co)variance drops as points get further apart in input space. For example, in a Gaussian kernel, $\exp\left(-h \| h_i - h_j \|^2 \right)$, $h > 0$ may be used to control how local the regression is.

Another simple way of generating flexible kernels is to exploit the fact that polynomial kernels with positive constants, such as

$$K(i, j) = \sigma_h^2 K_1(i, j) + \sigma_h^2 K_2(i, j) + \sigma_h^2 K_1(i, j) K_2(i, j),$$

Table 1. Mixed model equations under 2 parameterizations

<table>
<thead>
<tr>
<th>Parameterization I (Bayesian RKHS)</th>
<th>Parameterization II (animal model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_y^2 K + \sigma_e^2 K \hat{c} = \sigma_y^2 y$</td>
<td>$\sigma_y^2 \text{I} + \sigma_e^2 K^{-1} \hat{u} = \sigma_y^2 y$</td>
</tr>
</tbody>
</table>

1RKHS = reproducing kernel Hilbert spaces.
are also valid kernels. From a Bayesian perspective, polynomial kernels can be interpreted as follows: suppose that the model is \( y = f_1 + f_2 + f_{12} + \varepsilon \). Further, let the prior distribution be

\[
p(f_1, f_2, f_{12}) = N(f_1 | 0, \Sigma_{h_1}^2) \cdot N(f_2 | 0, \Sigma_{h_2}^2) \cdot N(f_{12} | 0, \Sigma_{h_{12}}^2).
\]

This model is equivalently expressed using \( y = f + \varepsilon \) with prior

\[
p(f) = N(f | 0, \Sigma_{h_1}^2 + \Sigma_{h_2}^2 + \Sigma_{h_{12}}^2).
\]

With this, it becomes clear that variance parameters define weights over (co)variance structures. In a polynomial kernel, \( K_i(i, j) \) may be a kernel computed from a pedigree, whereas \( K_2(i, j) \) may be computed from marker genotypes. Alternatively, the 2 kernels may just be different positive definite functions applied to the same information set, for example, \( K_1(i, j) = a(i, j) \), and \( K_2(i, j) = d(i, j) \) may reflect infinitesimal effects attributable to dominance, in which case the above model becomes a standard infinitesimal model that accounts for additive, dominance, and dominance \( \times \) additive effects.

**Computational Considerations**

The fact that any RKHS regression can be parameterized as a standard animal model with \( K = A \) implies that some packages for genetic evaluation can be used to perform RKHS regressions with any information set and kernel. In addition, when developing software, one can choose between parameterizations I and II of Table 1 (or many others that can be defined using factorizations of the kernel; e.g., Cholesky or Eigen-value decompositions) depending on computational convenience. If there is an efficient algorithm for computing \( K^{-1} \) directly from \( T \), parameterization II will be preferred. However, if such an algorithm is not available, parameterization I may be preferred. This situation arises, for example, in models for nonadditive genetic effects (e.g., Cockerham, 1954; Kempthorne, 1954), where there are no efficient algorithms for computing the inverses of relationship matrices that represent infinitesimal effects attributable to dominance or epistasis (see Gianola and de los Campos, 2008, for a RKHS treatment of this issue).

**LITERATURE CITED**


