DOI: 10.1534/genetics.107.084285

# Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits

**Daniel Gianola**[*,†,‡,1] **and Johannes B. C. H. M. van Kaam**[§]

*[*]Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706, [†]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway, [‡]Scienze Entomologiche, Fitopatologiche, Microbiologiche Agrarie e Zootecniche, Universitá degli Studi di Palermo, 90128 Palermo, Italy and [§]Istituto Zooprofilattico Sperimentale della Sicilia "A. Mirri," 90129 Palermo, Italy*

Manuscript received November 7, 2007
Accepted for publication February 8, 2008

## ABSTRACT

Reproducing kernel Hilbert spaces regression procedures for prediction of total genetic value for quantitative traits, which make use of phenotypic and genomic data simultaneously, are discussed from a theoretical perspective. It is argued that a nonparametric treatment may be needed for capturing the multiple and complex interactions potentially arising in whole-genome models, *i.e.*, those based on thousands of single-nucleotide polymorphism (SNP) markers. After a review of reproducing kernel Hilbert spaces regression, it is shown that the statistical specification admits a standard mixed-effects linear model representation, with smoothing parameters treated as variance components. Models for capturing different forms of interaction, *e.g.*, chromosome-specific, are presented. Implementations can be carried out using software for likelihood-based or Bayesian inference.

A massive quantity of genomic information is increasingly available for several species. For example, Wong *et al.* (2004) reported 2.8 million single-nucleotide polymorphisms (SNPs) in the chicken genome, and Hayes *et al.* (2004) found 2507 putative SNPs in salmons. Hundreds of thousands of SNPs have been identified in humans (*e.g.*, Hartl and Jones 2005). It is natural to consider use of this information as an aid in genetic improvement of livestock or plants or in molecular classification (or prediction) of diseases. In medicine and agriculture, for example, genomic information could also be used for designing diet or plant fertilization regimes that are genotype specific.

Early discussions on the use of molecular markers in genetic selection programs are given by Soller and Beckmann (1982) and Fernando and Grossman (1989). Subsequently, much work has addressed determining location and use of a single or a few quantitative trait loci (QTL). However, Dekkers and Hospital (2002), in a review of many studies, observed that there are an abundant number of loci associated with variation in quantitative traits. These authors noted that most statistical methods for marker-assisted selection proposed so far do not deal adequately with the complexity (in the sense of number of loci) posed by many traits. A relevant issue to be addressed is how a massive number of SNPs, viewed as covariates with potential explanatory power, can be incorporated reasonably into a statistical model specification. Some hurdles in the process of model building include multiple testing, strong dependence of inferences on assumptions, ambiguous interpretation of effects in a multiple-marker analysis due to collinearity, the famous "curse of dimensionality," as the number of markers, *e.g.*, SNPs, exceeds by far the number of cases in a sample, and the handling of nonadditive gene action. Balding (2006) discusses many of these problems.

A main challenge is how the many interactions between genotypes at different loci ought to be dealt with. A stylized treatment of epistatic variability from an evolutionary perspective is presented by Cheverud and Routman (1995). Translating this into whole-genome data analysis is another matter: if thousands of marker genotypes are fitted in a model for genomic-assisted prediction, the number of potential interactions and their interpretation can be mind boggling.

First, consider an analysis with random-effects models, so that the variance component parameterization or, more generally, the dispersion structure becomes the focus of the problem. Due to smoothing or "regularization" induced by, *e.g.*, a multivariate normal assumption, all random effects can be predicted uniquely. This is illustrated in Meuwissen *et al.* (2001), Gianola *et al.* (2003), and Xu (2003). For instance, animal breeders typically infer a number of breeding values that amply exceed the number of observations available (Quaas and Pollak 1980). However, coping with nonadditive genetic variability introduces additional difficulty. Theoretically, epistatic variance can be partitioned into

[1] *Corresponding author:* Department of Animal Sciences, University of Wisconsin, 1675 Observatory Dr., Madison, WI 53706.
E-mail: gianola@ansci.wisc.edu

orthogonal additive × additive, additive × dominance, dominance × dominance, etc., variance components, only under idealized conditions. These include linkage equilibrium, absence of mutation and of selection, and no inbreeding and assortative mating (Cockerham 1954; Kempthorne 1954). These assumptions are violated in nature and in breeding programs. Also, estimation of nonadditive components of variance is very difficult, even under standard assumptions (Chang 1988), leading to imprecise inference. Therefore, whether or not standard random-effects models for quantitative genetic analysis account for nonadditive relationships between genotypes and phenotypes accurately remains an open question.

Second, interactions between markers could be studied using fixed-effects models; this is what Cheverud and Routman (1995) refer to as "physiological epistasis," to disassociate inference from the gene and genotype frequencies that generate a probability distribution. Such an analysis "runs out" of degrees of freedom quickly in a whole-genome treatment, because there are 2 d.f. per biallelic SNP locus. Even if the number of parameters is reduced in some manner, estimates of effects are expected to be unstable and imprecise, due to severe lack of orthogonality induced, partly, by extant linkage disequilibrium. Also, interactions involving more than three loci are very difficult to interpret. A standard parametric treatment may require a formidable model selection exercise, with any model in particular probably having little plausibility or predictive power. Bayesian model averaging (*e.g.*, Hoeting *et al.* 1999) is an option, but how can this be made free from some strong and possibly untestable parametric assumptions?

A third and distinct avenue is to explore model-free approaches, which may be useful for phenotypic prediction under subtle or cryptic forms of epistasis. There is little evidence that such methods have been considered in quantitative genetics. Gianola *et al.* (2006) discussed semiparametric procedures for analysis of complex phenotypic data involving massive genomic information. These authors argued that application of the parametric additive genetic model in selective breeding of livestock produced tangible dividends, as shown in Dekkers and Hospital (2002), and proposed combining a nonparametric treatment of effects of molecular SNPs with features of the additive polygenic mode of inheritance.

The objective of this article is to develop further a reproducing kernel Hilbert spaces (RKHS) mixed model proposed by Gianola *et al.* (2006), with a focus on its theoretical aspects. The accompanying article by González-Recio *et al.* (2008, this issue) presents an application of the methodology to data on chicken mortality.

This article is organized as follows. The semiparametric mixed model section sets the stage and introduces notation. The nonparametric treatment (RKHS) adopted here is sketched in the reproducing kernel hilbert spaces regression section, where the main theoretical results are presented; additional details are in the appendix. dual formulation shows how the problem can be embedded into a mixed-effects model structure and discusses how statistical learning proceeds in a penalized-likelihood framework. The rkhs chromosome mixed model section presents a linear model aimed at capturing interactions between many loci at different chromosomes and presents a Bayesian implementation. The article concludes with a discussion of some standing issues.

## SEMIPARAMETRIC MIXED MODEL

**Setting:** The notation follows that of Gianola *et al.* (2006). Each of $n$ individuals possesses a measurement for some quantitative trait denoted as $y$ and information on a possibly massive number of SNP genotypes represented by a vector $\mathbf{x}$. An SNP locus is considered biallelic, so at most three genotypes are observed. Genotype instances can be coded uniquely via two linearly independent variables per locus as in an analysis-of-variance setting, *i.e.*, with 2 d.f. per locus. In standard quantitative genetics settings, the two dummy variates are coded such that the corresponding effects are interpretable as "additive" and "dominance." This is irrelevant from the predictive point of view taken here, in the sense that parameters (most of which lack a mechanistic interpretation) serve as transition tools, to pass from observed to predicted data.

Suppose, temporarily, that there are no nuisance variables and that the focus is on discovering a function relating $\mathbf{x}_i$ to $y_i$. Three alternative modeling possibilities are considered, for illustrative purposes.

1. Let the relationship between $y$ and $\mathbf{x}$ be represented as

$$y_i = g(\mathbf{x}_i) + e_i; \quad i = 1, 2, \ldots, n, \tag{1}$$

where $y_i$ is a measurement on the quantitative trait for individual $i$, $\mathbf{x}_i$ is a $p \times 1$ vector of dummy SNP instance variates observed on $i$, and $g(.)$ is some unknown function relating genotypes to phenotypes. Define $g(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i)$ as the conditional expectation function, that is, the mean phenotypic value of an infinite number of individuals, all possessing the $p$-dimensional genotypic attribute vector $\mathbf{x}_i$. $e_i \sim (0, \sigma_e^2)$ is a random residual, distributed independently of $\mathbf{x}_i$ and with variance $\sigma_e^2$. Typically, the residual distribution is assumed normal.

The vector $\mathbf{x}$ may have a probability distribution reflecting frequencies of the SNP attributes in the population. However, the prediction problem normally centers on what can be expected about the phenotypic distribution, given some specific configuration $\mathbf{x} = \mathbf{x}^*$, say. In nonparametric regression, $g(\mathbf{x}_i)$ is left unspecified and estimated as a smooth $\hat{g}(\mathbf{x}_i)$; this function represents pertinent signals on the phenotype from elements of $\mathbf{x}_i$, acting either additively or as members of some genetic network. Several techniques for inferring $g(\mathbf{x}_i)$ are described in Takezawa (2005).

2. A second specification is the additive regression model

$$g(\mathbf{x}_i) = \sum_{j=1}^{p} E(y_i \mid x_{ij}) = \sum_{j=1}^{p} g_j(x_{ij}) \tag{2}$$

(Hastie and Tibshirani 1990; Fox 2005), where $x_{ij}$ is the value of attribute $j$ in individual $i$. Each of the "partial-regression" functions $g_j(x_{ij})$ allows exploration of effects of individual attributes on phenotypes. This model is expected to pick up additive and dominance effects at each of the marker loci, but not epistatic interactions. It does not possess any clear advantage over a standard regression model with additive and dominance effects, the main difference residing in the nonparametric treatment that (2) would receive.

3. One could also think in terms of an additive "chromosome" model, as follows. Let $C$ be the number of pairs of chromosomes, and partition vector $\mathbf{x}_i$ as $\mathbf{x}_i = [\mathbf{x}'_{i1}\,\mathbf{x}'_{i2}\,\ldots\,\mathbf{x}'_{iC}]'$, so that $\mathbf{x}'_{ij}$ contains the values of the SNP instance variates at chromosome pair $j$ ($j = 1, 2, \ldots, C$), and so on. If the number of SNPs in chromosome pair $j = p_j/2$, then the order of $\mathbf{x}_{ij}$ is $p_j$, and the dimension of $\mathbf{x}$ is $p = \sum_{j=1}^{C} p_j$. The additive chromosome model is

$$g(\mathbf{x}_i) = \sum_{j=1}^{C} g_j(\mathbf{x}_{ij}), \tag{3}$$

with $\mathbf{x}_{ij}$ being the attributes observed on chromosome pair $j$ for individual $i$. This model would account for chromosome-specific signals (reflecting additive, dominance, and any relevant epistatic effects involving genes in chromosome pair $j$) and combine all these additively over pairs of chromosomes. Examples of tightly linked genes having epistatic effects are the major histocompatibility complex and the lac operon in *Escherichia coli*. Evidence of epistatic interactions among linked loci in plants is in Fenster and Galloway (2000), who studied fitness traits in the annual legume *Chamaecrista fasciculata*. The interplay between epistasis, linkage, and linkage disequilibrium is an old topic in population genetics (Kimura 1965; Franklin and Lewontin 1970).

Another modeling option consists of dividing all chromosomes somehow into $R$ genomic regions of equal or different sizes and then combining the $R$-region-specific signals additively.

Models (1)–(3) are nonparametric descriptors of situations in which epistasis plays different roles, *i.e.*, a major one in (1), none in (2), or involving only linked genes in (3). In what follows, model (1) is retained for presentation of theoretical developments, which are extended to model (3) later on.

**Additional structure:** Animal breeders have exploited to advantage the additive model of quantitative genetics, embedding it into a mixed-effects linear model specification. Basing selection of parents on predictions of additive genetic values, notable genetic progress has been attained in many species, such as dairy cattle, pigs, and poultry. While it is possible to accommodate some types of nonadditive gene action in a parametric manner, the assumptions are very strong. Further, construction and inversion of "epistatic relationship matrices" are daunting and a realistic parametric treatment is simply not available. Hence, as argued by Gianola *et al.* (2006), it seems reasonable to expand (1) as

$$y_i = \mathbf{w}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u} + g(\mathbf{x}_i) + e_i; \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\beta}$ is an $f \times 1$ vector of nuisance location parameters and $\mathbf{u}$ is a $q \times 1$ vector containing additive genetic effects of $q$ individuals (these effects are assumed here to be independent of those of the markers), some of which may lack a phenotypic record, so typically $n << q$; $\mathbf{w}'_i$ and $\mathbf{z}'_i$ are known nonstochastic incidence vectors. As before, $g(\mathbf{x}_i)$ is an unknown function of the SNP data, to be inferred. It is assumed that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{u}})$, where $\sigma^2_{\mathrm{u}}$ is the additive genetic variance due to unmarked polygenes and $\mathbf{A}$ is the additive relationship matrix, whose entries are twice the coefficients of coancestry between individuals. Let $\mathbf{e} = \{e_i\}$ be the $n \times 1$ vector of residuals, and take $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{e}})$. In matrix notation

$$\mathbf{y} = \{y_i\} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{g}(\mathbf{X}) + \mathbf{e},$$

where $\mathbf{W} = \{\mathbf{w}'_i\}$ and $\mathbf{Z} = \{\mathbf{z}'_i\}$ are incidence matrices of appropriate order. Further, $\mathbf{g}(\mathbf{X}) = \{g(\mathbf{x}_i)\}$ is a vector of order $n \times 1$, an unknown function of marker matrix $\mathbf{X}$, with $n$ rows and $p$ columns; a row of $\mathbf{X}$ contains the $p$ SNP instance variates (two per marker locus) observed in individual $i$.

Gianola *et al.* (2006) suggested backfitting-type algorithms in which, first, $g(\mathbf{x}_i)$ is estimated for $i = 1, 2, \ldots, n$, via some nonparametric estimate $\hat{g}(\mathbf{x}_i)$, and then a standard (frequentist or Bayesian) mixed-model analysis is carried out using the "corrected" data vector and pseudomodel

$$\mathbf{y}^* = \{y_i - \hat{g}(\mathbf{x}_i)\} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is a residual vector. The pseudomodel ignores uncertainty about $g(\mathbf{x})$, because $\hat{g}(\mathbf{x}_i)$ is treated as if it were the true regression (on SNPs) surface and $\boldsymbol{\varepsilon}$ is regarded as having the same distribution as $\mathbf{e}$, which is of course not true in finite samples. Subsequently, some estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ are obtained, and the offset $\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ is evaluated at these estimates, to produce a new fit of $g(\mathbf{x}_i)$. The backfitting algorithm iterates back and forth between the nonparametric and parametric phases. At convergence, the "total" genetic value of individual $i$ is assessed as $\tilde{T}_i = \tilde{u}_i + \tilde{g}(\mathbf{x}_i)$, where $\tilde{u}_i$ is the converged value of the empirical best linear unbiased predictor (or of a posterior mean in a Bayesian analysis) of $u_i$ and $\tilde{g}(\mathbf{x}_i)$ is the converged nonparametric smooth of $g(\mathbf{x}_i)$. Instead, a self-contained approach for inferring $\mathbf{u}$ and $g(\mathbf{x}_i)$ is discussed in what follows.

## REPRODUCING KERNEL HILBERT SPACES REGRESSION

**Theory:** A precise account of the theory is beyond the scope of this article, so only essentials are given here. Foundations and some applications are in Aronszajn (1950), Kimeldorf and Wahba (1971), and Wahba (1990, 1999, 2002). Some essential theoretical details and term definitions are presented in the appendix.

Consider inferring a function $g$ from data $\mathbf{y}$, without any assumptions. The problem is ill-posed, because any function passing through the data would be acceptable (Rasmussen and Williams 2006). Bayesians introduce assumptions via a prior over functions, but this problem has also been tackled using "regularization," *i.e.*, by imposing some smoothness assumptions on $g$. This second approach starts by considering the functional (a function containing functions as part of an argument)

$$J(g) = Q[\mathbf{y}, \mathbf{g}(\mathbf{X})] + a\|g(\mathbf{x})\|^2_{\mathcal{H}}, \tag{4}$$

where $\mathbf{g}(\mathbf{X}) = [\,g(\mathbf{x}_1)\,g(\mathbf{x}_2)\,\ldots\,g(\mathbf{x}_n)\,]'$; $Q[\mathbf{y}, \mathbf{g}(\mathbf{X})]$ is some function of the data and of $\mathbf{g}(\mathbf{X})$; $a$ is a positive smoothing parameter (typically unknown); and $\|g(\mathbf{x})\|_{\mathcal{H}}$ is some norm or "stabilizer" under a Hilbert space $\mathcal{H}$, a space of functions on a set having an inner product $\langle g_1, g_2\rangle$ and a norm $\|g_1\| = \sqrt{\langle g_1, g_1\rangle}$ for $g_1, g_2 \varepsilon \mathcal{H}$ (Wahba 2002; Mallick *et al.* 2005).

**Optimizing function:** Consider functional (4), and let

$$Q[\mathbf{y}, \mathbf{g}(\mathbf{X})] = \frac{1}{2\sigma^2_{\mathrm{e}}}\sum_{i=1}^{n}[y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u} - g(\mathbf{x}_i)]^2,$$

which is a deviance measure, assuming temporarily that $\mathbf{u}$ is a fixed parameter in the frequentist sense; subsequently, a random-effects treatment of $\mathbf{u}$ is made. Making explicit the dependency of the functional on the positive smoothing parameter $a$, write

$$J(g \mid a) = \frac{1}{2\sigma^2_{\mathrm{e}}}\sum_{i=1}^{n}[y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u} - g(\mathbf{x}_i)]^2 + \frac{a}{2}\|g(\mathbf{x})\|^2_{\mathcal{H}}, \tag{5}$$

where the factor $\frac{1}{2}$ is introduced for convenience. The second term in (5) acts as a penalty because it adds up to the deviance. It is also known as a regularizer, representing smoothness assumptions encoded in the RKHS. The issue here is finding the function $g(\mathbf{x})$ that minimizes (5), which is a calculus-of-variations problem over a space of smooth curves. The solution is given by the representer theorem of Kimeldorf and Wahba (1971); see Wahba (1999) for a more recent account and O'Sullivan *et al.* (1986) for extensions to generalized linear model deviances. The representer theorem states that the minimizer has the form

$$g(\mathbf{x} \mid h) = \alpha_0 + \sum_{j=1}^{n}\alpha_j k_h(\mathbf{x}, \mathbf{x}_j), \tag{6}$$

where the α's are unknown coefficients and the basis function $k_h(\mathbf{x}, \mathbf{x}_j)$ is a reproducing kernel, possibly dependent on some parameter $h$. While $\mathbf{x}$ is $p \times 1$, there are $n + 1$ coefficients in the function. The intercept $\alpha_0$ can be included as part of $\boldsymbol{\beta}$, so that the focus is on $\alpha_1, \alpha_2, \ldots, \alpha_n$. A possible kernel to be used as a basis function (Mallick *et al.* 2005) is the single-smoothing-parameter squared exponential (Gaussian) function

$$k_h(\mathbf{x}, \mathbf{x}_j) = \exp\left[-\frac{(\mathbf{x} - \mathbf{x}_j)'(\mathbf{x} - \mathbf{x}_j)}{h}\right].$$

The values of $k_h(\mathbf{x}, \mathbf{x}_j)$ range between 0 and 1, so the kernel is positive definite and acts as a correlation, in the sense that the closer $\mathbf{x}_j$ is to $\mathbf{x}$, the stronger the correlation is. Parameter $h$ controls the rate of decay of the correlation: smaller $h$ values produce a sharper correlogram. Define now the $1 \times n$ row vector

$$\begin{aligned}\mathbf{k}_i'(h) &= \{k_h(\mathbf{x}_i, \mathbf{x}_j)\}\\ &= \left\{\exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h}\right]\right\},\\ &\quad j = 1, 2, \ldots, n;\end{aligned}$$

the $n \times n$ symmetric matrix $\mathbf{K}_h = \{k_h(\mathbf{x}_i, \mathbf{x}_j)\}$ of kernels, which can be interpreted as a correlation matrix; and the $n \times 1$ column vector $\boldsymbol{\alpha} = \{\alpha_j\}$, $j = 1, 2, \ldots, n$. Then, the minimizing function (6) can be expressed in vectorial manner as the linear function of $\boldsymbol{\alpha}$:

$$\mathbf{g}(\mathbf{X} \mid h) = \begin{bmatrix}\mathbf{k}_1'(h)\boldsymbol{\alpha}\\ \mathbf{k}_2'(h)\boldsymbol{\alpha}\\ \cdot\\ \cdot\\ \cdot\\ \mathbf{k}_n'(h)\boldsymbol{\alpha}\end{bmatrix} = \mathbf{K}_h\boldsymbol{\alpha}. \tag{7}$$

These results can now be employed in (5), leading to a function having $\boldsymbol{\beta}$, $\mathbf{u}$, and $\boldsymbol{\alpha}$ as arguments, given $a$ and $h$. One obtains

$$
\begin{aligned}
&J(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\alpha}\mid a,h)\\
&=\frac{1}{2\sigma_{\mathrm{e}}^2}\sum_{i=1}^{n}[y_i-\mathbf{w}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{u}-\mathbf{k}_i'(h)\boldsymbol{\alpha}]^2+\frac{a}{2}\|\mathbf{K}_h\boldsymbol{\alpha}\|_{\mathcal{H}}^2\\
&=\frac{1}{2\sigma_{\mathrm{e}}^2}(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})\\
&\quad+\frac{a}{2}\|\mathbf{K}_h\boldsymbol{\alpha}\|_{\mathcal{H}}^2.
\end{aligned}
\tag{8}
$$

Using (A1) in the APPENDIX,

$$
\begin{aligned}
\|\mathbf{K}_h\boldsymbol{\alpha}\|_{\mathcal{H}}^2=\langle\mathbf{K}_h\boldsymbol{\alpha},\mathbf{K}_h\boldsymbol{\alpha}\rangle&=\int\boldsymbol{\alpha}'\mathbf{K}_h'\mathbf{K}_h\boldsymbol{\alpha}p(\mathbf{t})\,d\mathbf{t}\\
&=\boldsymbol{\alpha}'\int\mathbf{K}_h'\mathbf{K}_h\boldsymbol{\alpha}p(\mathbf{t})\,d\mathbf{t}.
\end{aligned}
\tag{9}
$$

Now, a vectorial generalization of the result in (A4) of the APPENDIX is

$$
\begin{bmatrix}\langle g(.),\,k(\mathbf{x}_1,.)\rangle_{\mathcal{H}}\\ \langle g(.),\,k(\mathbf{x}_2,.)\rangle_{\mathcal{H}}\\ \cdot\\ \cdot\\ \cdot\\ \langle g(.),\,k(\mathbf{x}_n,.)\rangle_{\mathcal{H}}\end{bmatrix}=\begin{bmatrix}g(\mathbf{x}_1)\\ g(\mathbf{x}_2)\\ \cdot\\ \cdot\\ \cdot\\ g(\mathbf{x}_n)\end{bmatrix}=\mathbf{K}_h\boldsymbol{\alpha}.
$$

This can be used in (9), because the integral is a vector valued inner product of the kernel and of each minimizer (7); that is,

$$
\|\mathbf{K}_h\boldsymbol{\alpha}\|_{\mathcal{H}}^2=\boldsymbol{\alpha}'\int\mathbf{K}_h'\mathbf{K}_h\boldsymbol{\alpha}p(\mathbf{t})\,d\mathbf{t}=\boldsymbol{\alpha}'\mathbf{K}_h\boldsymbol{\alpha}.
$$

Finally, this can be employed in (8), producing

$$
\begin{aligned}
&J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)\\
&=\frac{1}{2\sigma_{\mathrm{e}}^2}(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})\\
&\quad+\frac{a}{2}\boldsymbol{\alpha}'\mathbf{K}_h\boldsymbol{\alpha}.
\end{aligned}
\tag{10}
$$

Note that (10) does not include a penalty for the random vector **u**. This is added later on.

**Minimizer of the penalized sum of squares:** The gradients of $J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)$ with respect to the parametric $(\boldsymbol{\beta},\,\mathbf{u})$ and nonparametric $(\boldsymbol{\alpha})$ coefficients are

$$
\frac{\partial}{\partial\boldsymbol{\beta}}J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)=-\frac{1}{\sigma_{\mathrm{e}}^2}\mathbf{W}'(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha}),
$$

$$
\frac{\partial}{\partial\mathbf{u}}J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)=-\frac{1}{\sigma_{\mathrm{e}}^2}\mathbf{Z}'(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha}),
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\alpha}}J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)=-\frac{1}{\sigma_{\mathrm{e}}^2}&\mathbf{K}_h'(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})\\
&+a\mathbf{K}_h\boldsymbol{\alpha}.
\end{aligned}
$$

Noting that $\mathbf{K}_h$ is symmetric (so that $\mathbf{K}_h'\mathbf{K}_h=\mathbf{K}_h^2$; the notation $\mathbf{K}_h'$ is retained to facilitate analogies with mixed-model methodology), the first-order condition is satisfied by the system

$$
\begin{bmatrix}\mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{K}_h\\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{K}_h\\ \mathbf{K}_h'\mathbf{W} & \mathbf{K}_h'\mathbf{Z} & \mathbf{K}_h'\mathbf{K}_h+\frac{\sigma_{\mathrm{e}}^2}{a^{-1}}\mathbf{K}_h\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_{a,h}\\ \hat{\mathbf{u}}_{a,h}\\ \hat{\boldsymbol{\alpha}}_{a,h}\end{bmatrix}=\begin{bmatrix}\mathbf{W}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\\ \mathbf{K}_h'\mathbf{y}\end{bmatrix},
\tag{11}
$$

with the notation emphasizing the dependence of the solutions on $a$ and $h$. There is no unique solution to this system, because the number of equations $(p+q+n)$ exceeds the rank of the coefficient matrix. This problem is solved by a random-effects treatment of **u** via the assumption $\mathbf{u}\sim N(\mathbf{0},\,\mathbf{A}\sigma_{\mathrm{u}}^2)$, stated earlier. Under this assumption and the penalized-likelihood framework, the objective function to minimize becomes

$$
\begin{aligned}
J(\boldsymbol{\beta},\,\mathbf{u},\,\boldsymbol{\alpha}\mid a,\,h)=&\frac{1}{2\sigma_{\mathrm{e}}^2}(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})'\\
&\times(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}-\mathbf{K}_h\boldsymbol{\alpha})\\
&+\frac{1}{2\sigma_{\mathrm{u}}^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}+\frac{a}{2}\boldsymbol{\alpha}'\mathbf{K}_h\boldsymbol{\alpha}.
\end{aligned}
\tag{12}
$$

Taking derivatives as before, setting to zero and rearranging produces now

$$
\begin{bmatrix}\mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{K}_h\\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z}+\left(\frac{\sigma_{\mathrm{e}}^2}{\sigma_{\mathrm{u}}^2}\right)\mathbf{A}^{-1} & \mathbf{Z}'\mathbf{K}_h\\ \mathbf{K}_h'\mathbf{W} & \mathbf{K}_h'\mathbf{Z} & \mathbf{K}_h'\mathbf{K}_h+\left(\frac{\sigma_{\mathrm{e}}^2}{a^{-1}}\right)\mathbf{K}_h\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_{a,h}\\ \hat{\mathbf{u}}_{a,h}\\ \hat{\boldsymbol{\alpha}}_{a,h}\end{bmatrix}=\begin{bmatrix}\mathbf{W}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\\ \mathbf{K}_h'\mathbf{y}\end{bmatrix},
\tag{13}
$$

which has full rank if the elements of $\boldsymbol{\beta}$ are defined uniquely, *i.e.*, as a set of linearly independent estimable functions. There is a clear parallel between the forms of the *u*-equations and of the α-equations. In particular, in the nonparametric part, $(\sigma_{\mathrm{e}}^2/a^{-1})\mathbf{K}_h$ plays the role of $(\boldsymbol{\sigma}_{\mathrm{e}}^2/\sigma_{\mathrm{u}}^2)\mathbf{A}^{-1}$. Hence, one can arrive at representation (12) by making the assumption $\boldsymbol{\alpha}\mid a,\,h\sim N(0,\,\mathbf{K}_h^{-1}\sigma_{\alpha}^2)$, where $\sigma_{\alpha}^2=(1/a)$ is the "variance" of the $\boldsymbol{\alpha}$-effects, and $\mathbf{K}_h^{-1}$ is their correlation matrix. Fortunately, this inverse is not needed for solving (13) as $\mathbf{K}_h$ is a dense $n\times n$ matrix. The α-equations can be rearranged such that the solution for the nonparametric coefficients is

$$
\hat{\boldsymbol{\alpha}}_{a,h}=\left(\mathbf{K}_h+\frac{\sigma_{\mathrm{e}}^2}{a^{-1}}\mathbf{I}\right)^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\beta}}_{a,h}-\mathbf{Z}\hat{\mathbf{u}}_{a,h}).
$$

Large linear systems such as (13) have been solved routinely in animal breeding since the 1980s (*e.g.*, QUAAS and POLLAK 1980). System (13) differs from Equation 26 in GIANOLA *et al.* (2006), which has $\mathbf{K}_h'\mathbf{K}_h+(\sigma_{\mathrm{e}}^2/a^{-1})\mathbf{I}$ instead of $\mathbf{K}_h'\mathbf{K}_h+(\boldsymbol{\sigma}_{\mathrm{e}}^2/a^{-1})\mathbf{K}_h$. The latter is the correct RKHS representation.

## DUAL FORMULATION

**The linear model:** The preceding results imply that the RKHS approach is equivalent (this is referred to as a "dual" formulation) to the linear model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{K}_h\boldsymbol{\alpha} + \mathbf{e}, \tag{14}$$

under the assumptions that $\boldsymbol{\beta}$ is a "fixed" vector and that the random effects $\mathbf{u}$, $\boldsymbol{\alpha}$, and $\mathbf{e}$ are distributed independently as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{u}})$, $\boldsymbol{\alpha} \mid h \sim N(0, \mathbf{K}_h^{-1}\sigma^2_\alpha)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{e}})$, respectively. Hence, given $h$, implementation of the RKHS regression is as in a standard mixed-effects linear model, especially if the kernel matrix does not involve any parameter(s) $h$. For instance, variance components can be estimated via restricted maximum likelihood; subsequently, point estimates of $\boldsymbol{\beta}$, $\mathbf{u}$, and $\boldsymbol{\alpha}$ are obtained by solving (13) evaluated at the variance-components estimates. If $\sigma^2_{\mathrm{e}}/\sigma^2_\alpha$ is large (implying that $a$ is large), the estimated $\boldsymbol{\alpha}$-coefficients are expected to be near 0. A remarkable aspect of the RKHS procedure is the mutual exchange of information between α-coefficients due to the nontrivial correlation structure induced by $\mathbf{K}_h$. This is similar to the exchange of information between relatives induced by $\mathbf{A}$ in the classical additive genetic model.

**Effective number of parameters:** In a standard regression model ($\boldsymbol{\beta}$-coefficients only) the degrees of freedom of the model are given by rank($\mathbf{W}$) $= f$, provided the incidence matrix has full-column rank (this can be assumed without loss of generality). The fitted value is $\hat{\mathbf{y}} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} = \mathbf{Sy}$, and the $n \times n$ matrix $\mathbf{S} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$ is called the smoother operator (Hastie and Tibshirani 1990). Note that the degrees of freedom can also be arrived at by taking tr[$\mathbf{S}$] $= f$.

Let now $\mathbf{Q}_h = [\mathbf{W}\,\mathbf{Z}\,\mathbf{K}_h]$ so that in the context of (14), the vector of fitted values is $\hat{\mathbf{y}}_h = \mathbf{Q}_h\mathbf{C}^{-1}\mathbf{Q}'_h\mathbf{y}$, where $\mathbf{C}$ is the coefficient matrix in (13). Hence, a measure of the effective number of parameters fitted in RKHS regression

$$\begin{aligned}\mathrm{tr}(\mathbf{Q}_h\mathbf{C}^{-1}\mathbf{Q}'_h) &= \mathrm{tr}(\mathbf{C}^{-1}\mathbf{Q}'_h\mathbf{Q}_h)\\ &= \mathrm{tr}[\mathbf{C}^{-1}(\mathbf{Q}'_h\mathbf{Q}_h + \mathbf{P} - \mathbf{P})]\\ &= p + q + n - \mathrm{tr}(\mathbf{C}^{-1}\mathbf{P}),\end{aligned}$$

where

$$\mathbf{P} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\mathrm{u}}}\mathbf{A}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \frac{\sigma^2_{\mathrm{e}}}{a^{-1}}\mathbf{K}_h \end{bmatrix}.$$

Further,

$$\mathrm{tr}(\mathbf{Q}_h\mathbf{C}^{-1}\mathbf{Q}'_h) = p + \left[q - \mathrm{tr}\left(\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\mathrm{u}}}\mathbf{C}^{uu}\mathbf{A}^{-1}\right)\right] + \left[n - \mathrm{tr}\left(\frac{\sigma^2_{\mathrm{e}}}{a^{-1}}\mathbf{C}^{\alpha\alpha}\mathbf{K}_h\right)\right], \tag{15}$$

where $\mathbf{C}^{uu}$ and $\mathbf{C}^{\alpha\alpha}$ are the $u$- and α-blocks of $\mathbf{C}^{-1}$. Then, in some sense $q - \mathrm{tr}((\sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{u}})\mathbf{C}^{uu}\mathbf{A}^{-1})$ is the effective number of additive genetic effects fitted and $n - \mathrm{tr}((\sigma^2_{\mathrm{e}}/a^{-1})\mathbf{C}^{\alpha\alpha}\mathbf{K}_h)$ is the effective number of α-coefficients. If in (12) $\sigma^2_{\mathrm{u}} \to \infty$ and $a = 0$ (equivalently, $\sigma^2_\alpha = \infty$), (15) is equal to $p + q + n$ and, in the limit, the degrees of freedom of the model are given by the rank of $[\mathbf{W}\,\mathbf{Z}\,\mathbf{K}_h]$, so that the model interpolates the data. On the other hand, as $\sigma^2_{\mathrm{u}}$ and $\sigma^2_\alpha$ tend to 0 ($a \to \infty$), the effective number of parameters fitted decreases, and the model becomes less capable of reflecting potentially existing patterns in the data.

**Uncertainty about predictions:** Given $h$ and ignoring the error of estimation of variance components, an estimator of the variance–covariance matrix of the estimates of $\boldsymbol{\beta}$, and of the prediction errors of $\mathbf{u}$ and $\boldsymbol{\alpha}$, is given by

$$\mathbf{V}_{\boldsymbol{\beta},\mathbf{u},\boldsymbol{\alpha}}(h) = \mathbf{C}^{-1}\sigma^2_{\mathrm{e}}.$$

Let $\mathbf{S}_h = \mathbf{Q}_h\mathbf{C}^{-1}\mathbf{Q}'_h$ be the smoothing matrix. The variance–covariance matrix of the vector of fitted values is

$$\mathbf{V}(\hat{\mathbf{y}}_h) = \mathbf{S}_h\mathbf{V}(\mathbf{y})\mathbf{S}_h = \mathbf{S}_h(\sigma^2_{\mathrm{u}}\mathbf{ZAZ}' + \sigma^2_\alpha\mathbf{K}_h + \sigma^2_{\mathrm{e}}\mathbf{I})\mathbf{S}_h,$$

and the variance–covariance matrix of the fitted residuals is

$$\begin{aligned}\mathbf{V}(\mathbf{y} - \hat{\mathbf{y}}_h) &= \mathbf{V}(\mathbf{y}) - \mathbf{V}(\hat{\mathbf{y}}_h)\\ &= \sigma^2_{\mathrm{u}}(\mathbf{ZAZ}' - \mathbf{S}_h\mathbf{ZAZ}'\mathbf{S}_h)\\ &\quad + \sigma^2_\alpha(\mathbf{K}_h - \mathbf{S}_h\mathbf{K}_h\mathbf{S}_h) + \sigma^2_{\mathrm{e}}(\mathbf{I} - \mathbf{S}^2_h).\end{aligned}$$

In a genetic context, a relevant prediction problem is that of inferring a vector of future observations $\mathbf{y}^*$ in individuals possessing marker genotypes $\mathbf{X}^*$, so that the unknown molecularly marked genetic value, or contribution to phenotype, is $\mathbf{g}(\mathbf{X}^*)$. The model for the future observations is

$$\mathbf{y}^* = \mathbf{W}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} + \mathbf{K}^*_h\boldsymbol{\alpha} + \mathbf{e}^*,$$

with its coefficients inferred from currently available data $\mathbf{y}$. The point predictor is $\hat{\mathbf{y}}^* = \mathbf{Q}^*_h\mathbf{C}^{-1}\mathbf{Q}^{*\prime}_h\mathbf{y}$, and the prediction error variance–covariance matrix is

$$\mathbf{V}(\mathbf{y}^* - \hat{\mathbf{y}}^*) = (\mathbf{Q}^*_h\mathbf{C}^{-1}\mathbf{Q}^*_h + \mathbf{I}^*)\boldsymbol{\sigma}^2_{\mathrm{e}},$$

where $\mathbf{I}^*$ is an identity matrix with as many rows and columns as there are elements in $\mathbf{y}^*$. A confidence band for the elements of $\mathbf{y}^*$ based on the pointwise standard errors of prediction (Hastie and Tibshirani 1990) is given by $\hat{\mathbf{y}}^* \pm \mathrm{Diag}[2\sqrt{\mathbf{V}(\mathbf{y}^* - \hat{\mathbf{y}}^*)}]$, where Diag{.} denotes a vector whose elements are equal to twice the square root of the diagonal elements of $\mathbf{V}(\mathbf{y}^* - \hat{\mathbf{y}}^*)$. The confidence band does not consider the uncertainty in the estimates of variance components as well as that associated with $h$.

**Tuning parameter:** If the kernel matrix involves one or more $h$'s, some value(s) needs to be arrived at. Typically, cross-validation (CV) is used (*e.g.*, Craven and Wahba 1979; Wahba 1990; Golub *et al.* 1999). The simplest method (albeit computationally intensive) is the leave-one-out cross-validation measure

$$\mathrm{CV}(h) = \frac{1}{n}(\mathbf{y} - \hat{\mathbf{y}}_{h,-i})'(\mathbf{y} - \hat{\mathbf{y}}_{h,-i}),$$

where $\hat{\mathbf{y}}_{h,-i}$ is a vector of fitted values resulting from $n$ fits obtained from deleting $y_1, y_2, \ldots, y_n$, respectively. For instance, $\hat{y}_{h,-1}$ is the fitted value of observation 1 using all data other than $y_1$, and so on. The value of $h$ chosen results from minimizing $\mathrm{CV}(h)$ over a grid. Clearly, this is not computationally feasible in most quantitative genetic data sets, where $n$ can range from hundreds to millions of observations. In such a situation, one may wish to carry out a cross-validation involving a less intensive level of folding, *e.g.*, a leave 20%-out assessment. A more appealing (and, on some grounds, theoretically firmer) criterion is the generalized cross-validation

$$\mathrm{GCV}(h) = \frac{(1/n)(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})}{[1 - (1/n)\mathrm{tr}(\mathbf{S}_h)]}.$$

Using (15), the statistic becomes

$$\mathrm{GCV}(h) = \frac{(1/n)(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})}{\{1 - (1/n)[p + [q - \mathrm{tr}((\sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{u}})\mathbf{C}^{uu}\mathbf{A}^{-1})] + [n - \mathrm{tr}((\sigma^2_{\mathrm{e}}/a^{-1})\mathbf{C}^{\alpha\alpha}\mathbf{K}_h)]]\}}.$$

The main difficulty here is the calculation of the inverse matrices under the trace operator. These traces may be approximated (animal breeders have proposed approximations required for REML computations) or estimated via Monte Carlo sampling; *e.g.*, in some Bayesian contexts the diagonal elements of $\mathbf{C}^{uu}$ and $\mathbf{C}^{\alpha\alpha}$ are proportional to posterior variances.

## RKHS CHROMOSOME MIXED MODEL

**The linear model:** Consider again the additive chromosome specification (3), but now in the context of linear model (14). For ease of presentation, let the number of pairs of chromosomes be $C = 2$; generalization is straightforward. The unknown function of SNP genotypes to be inferred is

$$g(\mathbf{x}_{i1}, \mathbf{x}_{i2}) = g_1(\mathbf{x}_{i1}) + g_1(\mathbf{x}_{i2}). \qquad (16)$$

Using the dual formulation, the model can be written as

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Zu} + [\mathbf{K}_{1,h_1} \quad \mathbf{K}_{2,h_2}]\begin{bmatrix}\boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2\end{bmatrix} + \mathbf{e}, \qquad (17)$$

where $\mathbf{K}_{1,h_1}$ is an $n \times n$ matrix with typical element $\exp[-((\mathbf{x}_{i1} - \mathbf{x}_{j1})'(\mathbf{x}_{i1} - \mathbf{x}_{j1})/h_1)]$, $\mathbf{K}_{2,h_2}$ is also $n \times n$ with typical element $\exp[-((\mathbf{x}_{i2} - \mathbf{x}_{j2})'(\mathbf{x}_{i2} - \mathbf{x}_{j2})/h_2)]$, $h_1$ and $h_2$ are the decay parameters corresponding to SNPs in chromosome pairs 1 and 2, respectively, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are each $n \times 1$ vectors of coefficients. As before, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are $p_1 \times 1$ and $p_2 \times 1$ genotype incidence vectors pertaining to the appropriate chromosomes; recall that the number of markers in the two chromosomes is given by $p_1/2$ and $p_2/2$, because two dummy variates are needed for coding the three genotypes uniquely. The counterpart of objective function (12) is

$$\begin{aligned} J(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\alpha} \mid a_1, a_2, h_1, h_2) &= \frac{1}{2\sigma^2_{\mathrm{e}}}(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Zu} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2)' \\ &\quad \times (\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Zu} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2) \\ &\quad + \frac{1}{2\sigma^2_{\mathrm{u}}}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \frac{a_1}{2}\boldsymbol{\alpha}'_1\mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 + \frac{a_2}{2}\boldsymbol{\alpha}'_2\mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2. \qquad (18) \end{aligned}$$

Using the dual formulation, the form of the penalty is equivalent to making the assumption

$$\begin{bmatrix}\boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2\end{bmatrix} \mid a_1, a_2, h_1, h_2 \sim N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}a_1^{-1}\mathbf{K}^{-1}_{1,h_1} & \mathbf{0} \\ \mathbf{0} & a_2^{-1}\mathbf{K}^{-1}_{2,h_2}\end{bmatrix}\right);$$

that is, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are independently and normally distributed with covariance matrices $a_i^{-1}\mathbf{K}^{-1}_{i,h_i}$, $i = 1, 2$. Letting $a_1^{-1} = \sigma^2_{\alpha_1}$, $a_2^{-1} = \sigma^2_{\alpha_2}$, the counterpart of (13) is

$$\begin{bmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{K}_{h_1} & \mathbf{W}'\mathbf{K}_{h_2} \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\mathrm{u}}}\mathbf{A}^{-1} & \mathbf{Z}'\mathbf{K}_{h_1} & \mathbf{Z}'\mathbf{K}_{h_2} \\ \mathbf{K}'_{h_1}\mathbf{W} & \mathbf{K}'_{h_1}\mathbf{Z} & \mathbf{K}'_{h_1}\mathbf{K}_{h_1} + \frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_1}}\mathbf{K}_{h_1} & \mathbf{K}'_{h_1}\mathbf{K}_{h_2} \\ \mathbf{K}'_{h_2}\mathbf{W} & \mathbf{K}'_{h_2}\mathbf{Z} & \mathbf{K}'_{h_2}\mathbf{K}_{h_1} & \mathbf{K}'_{h_2}\mathbf{K}_{h_2} + \frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_2}}\mathbf{K}_{h_2} \end{bmatrix} \times \begin{bmatrix} \hat{\boldsymbol{\beta}}(.) \\ \hat{\mathbf{u}}(.) \\ \hat{\boldsymbol{\alpha}}_1(.) \\ \hat{\boldsymbol{\alpha}}_2(.) \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{K}'_{h_1}\mathbf{y} \\ \mathbf{K}'_{h_2}\mathbf{y} \end{bmatrix}, \qquad (19)$$

where $\hat{\boldsymbol{\beta}}(.)$, $\hat{\mathbf{u}}(.)$, etc., denote that the solution vector in question depends on $(\sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, h_1, h_2)$.

**Implementation:** The procedure can be carried out in a non-Bayesian manner as follows:

Define a grid of $h_1$, $h_2$ values.

For each point in the grid, estimate $\sigma^2_{\mathrm{u}}$, $\sigma^2_{\alpha_1}$, $\sigma^2_{\alpha_2}$, and $\sigma^2_{\mathrm{e}}$ and (19).

For each point in the grid, calculate the fitted values

$$\hat{\mathbf{y}}(.) = \mathbf{W}\hat{\boldsymbol{\beta}}(.) - \mathbf{Z}\hat{\mathbf{u}}(.) - \mathbf{K}_{1,h_1}\hat{\boldsymbol{\alpha}}_1(.) - \mathbf{K}_{2,h_2}\hat{\boldsymbol{\alpha}}_2(.),$$

the smoothing matrix $\mathbf{S}(.) = \mathbf{Q}(.)\mathbf{C}^{-1}(.)\mathbf{Q}'(.)$, and the generalized cross-validation criterion.

Choose the combination of $h_1$, $h_2$ values optimizing $\mathrm{GCV}(\cdot)$, and predict future observations as outlined previously.

**Bayesian approach:** The procedures described for the "global" and chromosome models (14) and (17), respectively, do not take into account uncertainty about unknown parameters. This can be addressed by adopting a Bayesian perspective; see Mallick *et al.* (2005) and Gianola *et al.* (2006). Here, a Bayesian analysis of the chromosome model (16) using the dual formulation (17) is outlined.

Let the collection of unknowns be $\Psi = (\mathbf{y}^*, \boldsymbol{\theta})$, where, as before, $\mathbf{y}^*$ is a vector of future phenotypic values to be predicted and

$$\boldsymbol{\theta} = [\boldsymbol{\beta}', \mathbf{u}', \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \sigma^2_{\mathrm{u}}, \sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\mathrm{e}}, h_1, h_2]'.$$

Assume the joint prior density has the form

$$\begin{aligned} p(\Psi \mid H) &= N(\mathbf{y}^* \mid \boldsymbol{\theta}) p(\boldsymbol{\beta}) N(\mathbf{u} \mid \mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{u}}) N(\boldsymbol{\alpha}_1 \mid \mathbf{0}, \mathbf{K}^{-1}_{h_1}\sigma^2_{\alpha_1}) \\ &\times N(\boldsymbol{\alpha}_2 \mid \mathbf{0}, \mathbf{K}^{-1}_{h_2}\sigma^2_{\alpha_2}) p(\sigma^2_{\mathrm{u}} \mid \nu_{\mathrm{u}}, S^2_{\mathrm{u}}) p(\sigma^2_{\alpha_1} \mid \nu_\alpha, S^2_\alpha) \\ &\times p(\sigma^2_{\alpha_2} \mid \nu_\alpha, S^2_\alpha) p(\sigma^2_{\mathrm{e}} \mid \nu_{\mathrm{e}}, S^2_{\mathrm{e}}) p(h_1, h_2 \mid h_{\min}, h_{\max}), \end{aligned} \quad (20)$$

where $H$ denotes all hyperparameters (whose values are fixed *a priori*) and $N(.\,|\,.,.)$ indicates a multivariate normal distribution with appropriate mean vector and covariance matrix; the prior $N(\mathbf{y}^* \mid \boldsymbol{\theta})$ is discussed below. The four variance components $\sigma^2_{\mathrm{u}}, \sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\mathrm{e}}$ are assigned independent scaled inverse chi-square prior distributions with degrees of freedom $\nu$ and scale parameters $S^2$, with appropriate subscripts. Assign an improper prior distribution to each of the elements of $\boldsymbol{\beta}$ and, as in Mallick *et al.* (2005), adopt independent uniform priors for $h_1$ and $h_2$ with lower and upper boundaries $h_{\min}$ and $h_{\max}$, respectively.

Given $\boldsymbol{\theta}$, observations are assumed to be conditionally independent, so the distribution of the observed ($\mathbf{y}$) and future ($\mathbf{y}^*$) data is

$$N, \left( \begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \Big| \begin{bmatrix} \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 + \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2 \\ \mathbf{W}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} + \mathbf{K}^*_{1,h_1}\boldsymbol{\alpha}_1 + \mathbf{K}^*_{2,h_2}\boldsymbol{\alpha}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{I}_n\sigma^2_{\mathrm{e}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n*}\sigma^2_{\mathrm{e}} \end{bmatrix} \right), \quad (21)$$

where $n^*$ is the order of the future data vector. Given $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_2$, future observations are independent of past ones. Let $\mathbf{Q}(h_1, h_2) = [\mathbf{W}\,\mathbf{Z}\,\mathbf{K}_{1,h_1}\,\mathbf{K}_{2,h_2}]$ and $\mathbf{Q}^*(h_1, h_2) = [\mathbf{W}^*\,\mathbf{Z}^*\,\mathbf{K}^*_{1,h_1}\,\mathbf{K}^*_{2,h_2}]$.

Given $h_1$, $h_2$ the setting is that of a Bayesian analysis of a mixed linear model, and Markov chain Monte Carlo procedures for this situation are well known (*e.g.*, Wang *et al.* 1993, 1994; Sorensen and Gianola 2002). All conditional posterior distributions are known, except those of $h_1$, $h_2$. A Gibbs–Metropolis sampling scheme can be used in which conditional distributions are used for drawing $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\sigma^2_{\mathrm{u}}$, $\sigma^2_{\alpha_1}$, $\sigma^2_{\alpha_2}$, $\sigma^2_{\mathrm{e}}$, and $\mathbf{y}^*$, and a Metropolis–Hastings update is employed for $h_1$ and $h_2$. The distributions to be sampled from are considered successively.

Draw location effects $\boldsymbol{\beta}$ from a multivariate normal distribution with mean vector

$$\overline{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{y} - \mathbf{Z}\mathbf{u} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2) \quad (22)$$

and covariance matrix $\mathbf{V}_\beta = (\mathbf{W}'\mathbf{W})^{-1}\sigma^2_{\mathrm{e}}$.

Sample additive genetic effects from a normal distribution centered at

$$\overline{\mathbf{u}} = \left( \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\mathrm{u}}} \right)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2) \quad (23)$$

and with covariance matrix $\mathbf{V}_{\mathrm{u}} = (\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}(\sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{u}}))^{-1}\,\sigma^2_{\mathrm{e}}$.

The conditional posterior distributions of each of the coefficients $\boldsymbol{\alpha}$ are multivariate normal as well, with mean vectors

$$\begin{aligned} \overline{\boldsymbol{\alpha}}_1 &= \left( \mathbf{K}'_{1,h_1}\mathbf{K}_{1,h_1} + \mathbf{K}_{h_1}\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_1}} \right)^{-1} \\ &\times \mathbf{K}'_{1,h_1}(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2), \end{aligned} \quad (24)$$

$$\begin{aligned} \overline{\boldsymbol{\alpha}}_2 &= \left( \mathbf{K}'_{2,h_2}\mathbf{K}_{2,h_2} + \mathbf{K}_{h_2}\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_2}} \right)^{-1} \\ &\times \mathbf{K}'_{1,h_2}(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_2), \end{aligned} \quad (25)$$

and variance–covariance matrices

$$\mathbf{V}_{\alpha_1} = \left( \mathbf{K}'_{1,h_1}\mathbf{K}_{1,h_1} + \mathbf{K}_{h_1}\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_1}} \right)^{-1} \sigma^2_{\mathrm{e}}, \quad (26)$$

and

$$\mathbf{V}_{\alpha_2} = \left( \mathbf{K}'_{2,h_1}\mathbf{K}_{2,h_2} + \mathbf{K}_{h_2}\frac{\sigma^2_{\mathrm{e}}}{\sigma^2_{\alpha_2}} \right)^{-1} \sigma^2_{\mathrm{e}}. \quad (27)$$

The variance components have scaled inverse chi-square conditional posterior distributions and are conditionally independent. The conditional posterior distributions to sample from are as follows, where ELSE denotes all parameters other than those being sampled,

$$\sigma^2_{\mathrm{u}} \mid \mathrm{ELSE} \sim (\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + v_{\mathrm{u}}S^2_{\mathrm{u}})\chi^{-2}_{q+v_{\mathrm{u}}}, \quad (28)$$

$$\sigma^2_{\alpha_1} \mid \mathrm{ELSE} \sim (\boldsymbol{\alpha}_1'\mathbf{K}_{h_1}\boldsymbol{\alpha}_1 + \nu_\alpha S^2_\alpha)\chi^{-2}_{n+v_\alpha}, \quad (29)$$

$$\sigma^2_{\alpha_2} \mid \mathrm{ELSE} \sim (\boldsymbol{\alpha}_2'\mathbf{K}_{h_2}\boldsymbol{\alpha}_2 + \nu_\alpha S^2_\alpha)\chi^{-2}_{n+v_\alpha}, \quad (30)$$

and

$$\sigma^2_{\mathrm{e}} \mid \mathrm{ELSE} \sim (\mathbf{r}'\mathbf{r} + \nu_{\mathrm{e}}S^2_{\mathrm{e}})\chi^{-2}_{n_1+v_{\mathrm{e}}}, \quad (31)$$

where $\mathbf{r} = \mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{K}_{1,h_1}\boldsymbol{\alpha}_1 - \mathbf{K}_{2,h_2}\boldsymbol{\alpha}_2$ is the vector of residuals evaluated at the current sample values of the location effects.

The most difficult parameters to sample are $h_1$ and $h_2$. However, if the kernels do not involve $h$'s this step is omitted from the sampling process. If uniform (bounded) priors are adopted, their conditional posterior density is

$$p(h_1,\ h_2 \,|\, \mathrm{ELSE}) \propto \exp\left[-\frac{\mathbf{r}'_{h_1,h_2}\mathbf{r}_{h_1,h_2}}{2\sigma^2_{\mathrm{e}}}\right] I(h_{\min},\ h_{\max}), \quad (32)$$

where $I(h_{\min},\ h_{\max})$ is an indicator function taking the value 1 if both $h$ parameters are between the bounds and 0 otherwise. Further, the residual vector $\mathbf{r}_{h_1,h_2}$ has as its $i$th element

$$r_{i(h_1,h_2)} = y_i - \mathbf{w}'_i\boldsymbol{\beta} - \mathbf{z}'_i\mathbf{u} - \sum_{j=1}^{n} k_{h_1}(\mathbf{x}_{i1},\ \mathbf{x}_{j1})\alpha_{j1} - \sum_{j=1}^{n} k_{h_2}(\mathbf{x}_{i2},\ \mathbf{x}_{j2})\alpha_{j2},$$

recalling that

$$k_{h_1}(\mathbf{x}_{i1},\ \mathbf{x}_{j1}) = \exp\left[-\frac{(\mathbf{x}_{i1}-\mathbf{x}_{j1})'(\mathbf{x}_{i1}-\mathbf{x}_{j1})}{h_1}\right]$$

and

$$k_{h_2}(\mathbf{x}_{i2},\ \mathbf{x}_{j2}) = \exp\left[-\frac{(\mathbf{x}_{i2}-\mathbf{x}_{j2})'(\mathbf{x}_{i2}-\mathbf{x}_{j2})}{h_2}\right].$$

Density (32) is not in a recognizable form. However, a Metropolis algorithm (Metropolis *et al.* 1953), as suggested by Mallick *et al.* (2005) and Gianola *et al.* (2006), can be tailored for obtaining samples from the distribution $[h_1,\ h_2 \,|\, \mathrm{ELSE}]$. Let the Markov chain be at state $h_1^{[t]}$, $h_2^{[t]}$ and draw proposal values $h_1^*$ and $h_2^*$ from some symmetric candidate-generating distribution. The proposed values are accepted with probability

$$\gamma = \min\left[\frac{p(h_1^*,\ h_2^* \,|\, \mathrm{ELSE})}{p(h_1^{[t]},\ h_2^{[t]} \,|\, \mathrm{ELSE})},\ 1\right].$$

If the proposal is accepted, then set $h_1^{[t+1]}(h_2^{[t+1]}) = h_1^*(h_2^*)$; otherwise, the chain stays at $h_1^{[t]}(h_2^{[t]})$.

Finally, the vector of yet to be observed phenotypes is inferred from samples drawn from the conditional distribution

$$[\mathbf{y}^* \,|\, \mathrm{ELSE}] = N(\mathbf{W}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} + \mathbf{K}^*_{1,h_1}\boldsymbol{\alpha}_1 + \mathbf{K}^*_{2,h_2}\boldsymbol{\alpha}_2,\ \mathbf{I}_{n*}\sigma^2_{\mathrm{e}}), \quad (33)$$

with the values of $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\sigma^2_{\mathrm{e}}$, $h_1$, and $h_2$ evaluated at the corresponding realizations from the current round of Markov chain Monte Carlo sampling. When the algorithm converges to the equilibrium distribution, the samples of $\mathbf{y}^*$ drawn are from the predictive distribution $[\mathbf{y}^* \,|\, \mathbf{y},\ H]$, which fully takes into account the uncertainty about all unknown model parameters.

## DISCUSSION AND EXTENSIONS

This article presents a procedure for quantitative genetic analysis using information on whole-genome markers, such as SNPs, and phenotypic measurements for some complex candidate trait. Focus is on theory and methods based on RKHS regression, arguably the state of the art in functional data analysis (Wahba 1990; Gu 2002; Wood 2006). The model contains a parametric component, represented by the classical additive genetic model of quantitative genetics, and an unknown function or set of functions of SNP genotypes that is dealt with nonparametrically, as in the generalized additive models of Hastie and Tibshirani (1990). The number of nonparametric functions employed is a model choice issue, and many alternative specifications can be formulated. Here, a global function was considered, expected to reflect all relevant "genetic signals," *e.g.*, dominance and various forms of epistasis, as well as a sum of chromosome-specific functions, each of which is expected to capture dominance as well as epistasis involving linked loci at the corresponding chromosome.

The parametric component includes additive genetic effects only. Dominance and (some) epistatic effects can be handled parametrically. However, a standard treatment requires constructing and inverting large and possibly dense matrices such as, *e.g.*, **A** # **A** # **D** # **D** if an additive × additive × dominance × dominance variance component were to enter into the parameterization; # denotes Hadamard product. Further, the Cockerham–Kempthorne decomposition machinery available for dealing with epistatic variance collapses under inbreeding and selection; *e.g.*, all sorts of covariances between genetic effects crop up, rendering the parametric approach invalid. A related point is related to limitations of the orthodox view of nonadditive genetic effects in quantitative genetics. The classical definitions of epistatis pertain to a model in which effects enter linearly when forming the genotype. However, one could argue that biology is far from linear. For instance, it may not be enlightening to think in terms of variance components in situations in which a phenotype results from a sum of sine and cosine waves or when nonadditivity enters via terms such as $a_1\exp(a_2a_3/a_4)$, where the $a$'s are effects of alleles at some different loci.

Gianola *et al.* (2006) discuss how a "nonparametric analog of breeding value" can be derived via a Taylor series expansion; this may have merit in genomic selection contexts in which the objective is to increase (or decrease) additive genetic value for some quantitative trait. Caveats and generalizations of the procedures are discussed next.

**Filtering SNPs:** Availability of a massive number of SNPs does not necessarily imply that all markers should be included in a prediction model. Apart from standard preprocessing based on minimum allele frequency or information content, it may well be that predictive ability is enhanced by reducing the dimension of the features used as input in a model. For example, Long *et al.* (2007) described a machine-learning technique based on filtering (using entropy reduction) and wrapping (Bayesian classification performance), to process >5000 SNPs genotyped in broiler families. Predictive ability of bird mortality was increased when the top (based on information gain) 50 SNPs were downsized to 24. More research is needed in regard to the strategic use of markers, *e.g.*, using a few ones *vs.* all, or on the assignment of different window-width parameters to genomic regions in a whole-genome treatment.

**Model choice:** A natural question is that of the model to be used for predicting phenotypes. For instance, should a chromosome model be adopted, instead of a global specification? Comparison of models is a complex issue, as some specifications are better for describing observed data, while others may have a superior predictive ability. See Sorensen and Waagepetersen (2003) for a case study using several criteria for Bayesian model comparison. Some non-Bayesian techniques are discussed by Hastie and Tibshirani (1990) and Wood (2006). The latter include likelihood ratios evaluated at the penalized likelihood estimates and differences in deviances using approximations to the model degrees of freedom; none takes into account the uncertainty associated with the estimates of the smoothing parameters. A different approach, called BRUTO (Hastie and Tibshirani 1990) is based on iterative minimization of a modification of the generalized cross-validation statistic GCV(.) discussed earlier. For example, in a model with additive functions for each of $C$ chromosomes, there would be $2C$ tuning parameters $(h_1, h_2, \ldots, h_C, \sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \ldots, \sigma^2_{\alpha_C})$ involved in the BRUTO iteration. Bayesian methods of model comparison have a stronger formal justification.

**Incomplete genotyping:** In animal breeding, it is not feasible to genotype all individuals for the SNPs. For instance, poultry breeding and cattle artificial insemination companies typically genotype sires only. The number of animals with phenotypic information can be in the order of hundreds of thousands, and genotyping is selective, so animals with SNPs are not a random sample from the population. Methods for dealing with incomplete molecular information are discussed by Gianola *et al.* (2006), and some include sampling of genotypes. Here, the problem is revisited in the light of RKHS regression, primarily to illustrate difficulties.

Assume a global model with a single $h$ parameter. The vector of phenotypic values is partitioned as $\mathbf{y} = [\mathbf{y}_1' \, \mathbf{y}_2']'$, where $\mathbf{y}_1$ $(n_1 \times 1)$ consists of records of individuals lacking SNP data, and $\mathbf{y}_2$ $(n_2 \times 1)$ includes phenotypic data of genotyped individuals. In animal breeding $n_1 > p >> n_2$. Write the model

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{bmatrix}\mathbf{u} + \begin{bmatrix}\mathbf{K}_{11}^{\text{miss}}(h) & \mathbf{K}_{12}^{\text{miss}}(h)\\ \mathbf{K}_{21}^{\text{miss}}(h) & \mathbf{K}_{22}(h)\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix},$$

where $\mathbf{K}_{11}^{\text{miss}}(h)$ is an unobserved $n_1 \times n_1$ matrix of kernels, $\mathbf{K}_{12}^{\text{miss}}(h) = \mathbf{K}_{21}^{\text{miss}}(h)'$ is also an unobserved $n_1 \times n_2$ matrix, $\mathbf{K}_{22}(h)$ is the $n_2 \times n_2$ matrix of observed kernels, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are $n_1 \times 1$ and $n_2 \times 1$ vectors of coefficients. Specifically

$$\mathbf{K}_{11}^{\text{miss}}(h) = \left\{ k_h(\mathbf{x}_i^{\text{miss}}, \mathbf{x}_j^{\text{miss}}) = \exp\left[-\frac{(\mathbf{x}_i^{\text{miss}} - \mathbf{x}_j^{\text{miss}})'(\mathbf{x}_i^{\text{miss}} - \mathbf{x}_j^{\text{miss}})}{h}\right]\right\},$$

$$\mathbf{K}_{12}^{\text{miss}}(h) = \left\{ k_h(\mathbf{x}_i^{\text{miss}}, \mathbf{x}_j) = \exp\left[-\frac{(\mathbf{x}_i^{\text{miss}} - \mathbf{x}_j)'(\mathbf{x}_i^{\text{miss}} - \mathbf{x}_j)}{h}\right]\right\},$$

$$\mathbf{K}_{22}(h) = \left\{ k_h(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h}\right]\right\},$$

where $\mathbf{x}_i^{\text{miss}}$ denotes the vector of unobserved SNP genotypes in individuals with phenotypes $\mathbf{y}_1$. Assign the multivariate normal distribution (suppressing dependence on $h$ in the notation)

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\alpha}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{A}\sigma^2_{\text{u}} & \mathbf{0}\\ \mathbf{0} & \begin{pmatrix}\mathbf{K}_{11}^{\text{miss}} & \mathbf{K}_{12}^{\text{miss}}\\ \mathbf{K}_{21}^{\text{miss}} & \mathbf{K}_{22}\end{pmatrix}^{-1}\sigma^2_{\alpha}\end{bmatrix}\right),$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2]'$. Given $\mathbf{x}_i^{\text{miss}} (i = 1, 2, \ldots, n_1)$, $h$, $\sigma^2_{\text{u}}$, $\sigma^2_{\alpha}$, $\sigma^2_{\text{e}}$, and the phenotypes $\mathbf{y}_1$ and $\mathbf{y}_2$, the best linear unbiased predictor (conditional posterior mean) of $\mathbf{u}$ and $\boldsymbol{\alpha}$ is the solution to

$$\begin{bmatrix}
\sum_{i=1}^{2}\mathbf{W}_i'\mathbf{W}_i & \sum_{i=1}^{2}\mathbf{W}_i'\mathbf{Z}_i & \sum_{i=1}^{2}\mathbf{W}_i'\mathbf{K}_{i1}^{\text{miss}} & \sum_{i=1}^{2}\mathbf{W}_i'\mathbf{K}_{i2}^{\text{miss}}\\
\sum_{i=1}^{2}\mathbf{Z}_i'\mathbf{W}_i & \sum_{i=1}^{2}\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{A}^{-1}\frac{\sigma^2_{\text{e}}}{\sigma^2_{\text{u}}} & \sum_{i=1}^{2}\mathbf{Z}_i'\mathbf{K}_{i1}^{\text{miss}} & \sum_{i=1}^{2}\mathbf{Z}_i'\mathbf{K}_{i2}^{\text{miss}}\\
\sum_{i=1}^{2}\mathbf{K}_{1i}^{\text{miss}}\mathbf{W}_i & \sum_{i=1}^{2}\mathbf{K}_{1i}^{\text{miss}}\mathbf{Z}_i & \Phi_{11} & \Phi_{12}\\
\sum_{i=1}^{2}\mathbf{K}_{2i}^{\text{miss}}\mathbf{W}_i & \sum_{i=1}^{2}\mathbf{K}_{2i}^{\text{miss}}\mathbf{Z}_i & \Phi_{21} & \Phi_{22}
\end{bmatrix}
\times \begin{bmatrix}\hat{\boldsymbol{\beta}}^{\text{miss}}\\ \hat{\mathbf{u}}^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_1^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_2^{\text{miss}}\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^{2}\mathbf{W}_i'\mathbf{y}_i\\ \sum_{i=1}^{2}\mathbf{Z}_i'\mathbf{y}\\ \mathbf{K}_{11}^{\text{miss}}\mathbf{y}_1 + \mathbf{K}_{12}^{\text{miss}}\mathbf{y}_2\\ \mathbf{K}_{21}^{\text{miss}}\mathbf{y}_1 + \mathbf{K}_{22}\mathbf{y}_2\end{bmatrix}, \tag{34}$$

where

$$\begin{bmatrix}\boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12}\\ \boldsymbol{\Phi}_{21} & \boldsymbol{\Phi}_{22}\end{bmatrix} = \begin{bmatrix} (\mathbf{K}_{11}^{\text{miss}})^2 + \mathbf{K}_{12}^{\text{miss}}\mathbf{K}_{21}^{\text{miss}} & \mathbf{K}_{11}^{\text{miss}}\mathbf{K}_{12}^{\text{miss}} + \mathbf{K}_{12}^{\text{miss}}\mathbf{K}_{22} \\ +\frac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{K}_{11}^{\text{miss}} & +\frac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{K}_{12}^{\text{miss}} \\ \mathbf{K}_{21}^{\text{miss}}\mathbf{K}_{11}^{\text{miss}} + \mathbf{K}_{22}\mathbf{K}_{21}^{\text{miss}} & \mathbf{K}_{21}^{\text{miss}}\mathbf{K}_{12}^{\text{miss}} + (\mathbf{K}_{22})^2 \\ +\frac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{K}_{21}^{\text{miss}} & +\frac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{K}_{22} \end{bmatrix}.$$

As shown in (34) both the coefficient matrix and the vector of right-hand sides depend on $\mathbf{x}_i^{\text{miss}}(i = 1, 2, \ldots, n_1)$. From a Bayesian perspective under Gaussian assumptions,

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}^{\text{miss}}\\ \hat{\mathbf{u}}^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_1^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_2^{\text{miss}}\end{bmatrix} = E\left(\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\\ \boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{bmatrix} \mid \mathbf{y}_1, \mathbf{y}_2, \sigma_e^2, \sigma_\alpha^2, h, \mathbf{x}^{\text{miss}}, \mathbf{x}\right),$$

where $\mathbf{x}^{\text{miss}}$ is the collection of unobserved SNPs $\mathbf{x}_i^{\text{miss}}(i = 1, 2, \ldots, n_1)$ and $\mathbf{x}$ denotes the SNPs of all genotyped individuals. Assuming $\sigma_e^2$, $\sigma_\alpha^2$, and $h$ are known, for simplicity, one is interested in arriving at the unconditional expectation

$$E\left(\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\\ \boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{bmatrix} \mid \mathbf{y}_1, \mathbf{y}_2, \sigma_e^2, \sigma_\alpha^2, h, \mathbf{x}\right) = E_{\mathbf{x}^{\text{miss}} \mid \mathbf{y}_1, \mathbf{y}_2, \sigma_e^2, \sigma_\alpha^2, h, \mathbf{x}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}^{\text{miss}}\\ \hat{\mathbf{u}}^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_1^{\text{miss}}\\ \hat{\boldsymbol{\alpha}}_2^{\text{miss}}\end{bmatrix}\right), \quad (35)$$

so the Bayesian solution requires averaging over the conditional distribution $[\mathbf{x}^{\text{miss}} \mid \mathbf{y}_1, \mathbf{y}_2, \sigma_e^2, \sigma_\alpha^2, h, \mathbf{x}]$. This is a formidable probabilistic imputation, although some simplification is possible. It would seem reasonable to approximate this distribution by $[\mathbf{x}^{\text{miss}} \mid \mathbf{x}]$, arguing that, conditionally on $\mathbf{x}$, phenotypic values do not provide much additional information about SNP genotypes. Subsequently, form the Bayesian classifier

$$\Pr(\mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}} \mid \mathbf{x}) = \frac{\Pr(\mathbf{x} \mid \mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}})\Pr(\mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}})}{\Pr(\mathbf{x})},$$

where $\Pr(\mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}})$ is the prior probability of observing genotype configuration $\tilde{\mathbf{x}}^{\text{miss}}$ in the population. The prior probability can be estimated using models with various degrees of refinement; *e.g.*, one can assume linkage equilibrium and estimate the joint prior probability from the product of the marginal distributions at individual SNP loci. Likewise, $\Pr(\mathbf{x} \mid \mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}})$ can be approximated by some naïve probability calculation, *e.g.*, assuming independence between individuals and loci. The denominator can also be approximated as

$$\Pr(\mathbf{x}) \approx \sum_{\tilde{\mathbf{x}}^{\text{miss}} \in S} \Pr(\mathbf{x} \mid \mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}})\Pr(\mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}}),$$

where $S$ is a set of missing values having relatively high plausibility. Naïve Bayesian classifiers have been enormously successful in the machine-learning literature (*e.g.*, Elkan 1997) and may prove to be competitive against some involved genotype sampling procedures that have been suggested (Van Arendonk *et al.* 1989; Fernando *et al.* 1993; Sheehan and Thomas 1993; Jensen *et al.* 1995; Kerr and Kinghorn 1996; Jensen and Kong 1999; Fernández *et al.* 2002; Stricker *et al.* 2002). Once an approximation to $\Pr(\mathbf{x}^{\text{miss}} = \tilde{\mathbf{x}}^{\text{miss}} \mid \mathbf{x})$ is arrived at, this can be used as mixing distribution in (35). Alternatives are discussed in Gianola *et al.* (2006), including fitting a bivariate model.

**Choice of kernel and discreteness of genotypes:** Following Hastie and Tibshirani (1990) and Wood (2006), the developments carry when the distribution of $y$ is a member of the exponential family. As in density estimation, many candidate kernels are available, and attaining a good predictive behavior is critically dependent on the choice of kernel. As noted, some kernels do not involve tuning parameters; for instance, see González-Recio *et al.* (2008).

The theory of RKHS regression holds for a continuously valued $\mathbf{x}$. It is unknown if the procedures are robust with respect to using a Gaussian kernel function when, in fact, SNP genotypes or haplotypes are discrete. Silverman (1986) discussed univariate density estimation and concluded that various kernels differed little in mean squared error. It is unknown if this robustness argument holds for RKHS regression (clearly the inner product arguments are valid for the continuous case), but a discrete approximation may work.

A kernel suitable for discrete covariates is proposed here. For a biallelic SNP, there are three possible genotypes at each "locus." Suppose the elements of $\mathbf{x}$ are coded as 0, 1, 2, to denote the appropriate genotypes. For an $\mathbf{x}$ vector with $p$ coordinates, its statistical distribution is given by the probabilities of each of the $3^p$ outcomes. With SNPs, $p$ can be very large (possibly much larger than $n$), so it is hopeless to estimate the probability distribution of genotypes accurately from observed relative frequencies, and smoothing is required (Silverman 1986). The number of disagreements between a focal $\mathbf{x}$ and the observed $\mathbf{x}_i$ in subject $i$ is given by

$$d(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}_i - \mathbf{x})'(\mathbf{x}_i - \mathbf{x}),$$

where $d(.)$ takes values between 0 and $4p$. As an illustration, if "genotype" *AABbccDd* is the focal point and individual $i$ in the sample is *aaBbccDD*, then, $d(\mathbf{x}, \mathbf{x}_i) = 5$. Following Silverman (1986), one could use the "binomial" kernel

$$k(\mathbf{x}, \mathbf{x}_i, h) = h^{4p - d(\mathbf{x}, \mathbf{x}_i)}(1 - h)^{d(\mathbf{x}, \mathbf{x}_i)},$$

with $\frac{1}{2} \leq h \leq 1$; alternative forms of the kernel function are discussed by Aitchison and Aitken (1976) and Racine and Li (2004). The (pseudo-)RKHS dual formulation would take the form

$$y_i = \mathbf{w}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + \sum_{j=1}^{n} h^{4p - d(\mathbf{x}_i, \mathbf{x}_j)}(1 - h)^{d(\mathbf{x}_i, \mathbf{x}_j)}\alpha_{j1}.$$

Alternatively, incidence of the three genotypes at a locus can be described with two "free" dummy variates, as in standard ANOVA. There are $p$ predictor variates, where $p/2$ is the number of markers. Let $x_{1k}$ and $x_{2k}$ be the two dummy variates at locus $k$ ($k = 1, 2, \ldots, p/2$). Define

$$d_{1ik}(\mathbf{x}_1, \mathbf{x}_{i1}) = (x_{i1k} - x_{1k})^2, \quad d_{2ik}(\mathbf{x}_2, \mathbf{x}_{i2}) = (x_{i2k} - x_{2k})^2,$$

which give the number of disagreements between a focal $x_{1k}$ ($x_{2k}$) and the observed $x_{i1k}(x_{i2k})$ in subject $i$. Each of the $d$'s varies between 0 and 1. Then, let

$$d_{1ij} = \sum_{k=1}^{p/2}(x_{i1k} - x_{j1k})^2, \quad d_{2ij} = \sum_{k=1}^{p/2}(x_{i2k} - x_{j2k})^2,$$

each varying between 0 and $p/2$. Subsequently, consider the "trinomial" kernel

$$\begin{aligned} k'(\mathbf{x}_i, \mathbf{x}_j, h_1, h_2) &= h_1^{d_{1ij}} h_2^{d_{2ij}} (1 - h_1 - h_2)^{p - d_{1ij} - d_{2ij}} \\ &= h_1^{\sum_{k=1}^{p/2}(x_{i1k}-x_{j1k})^2} h_2^{\sum_{k=1}^{p/2}(x_{i2k}-x_{j2k})^2} \\ &\quad \times (1 - h_1 - h_2)^{p - \sum_{k=1}^{p/2}\left[(x_{i1k}-x_{j1k})^2 + (x_{i2k}-x_{j2k})^2\right]}, \end{aligned}$$

where $\mathbf{x}_i$ is the "focal" $p \times 1$ vector of covariates and $\mathbf{x}_j$ is the observed value in individual $j$. If each of the $h_1$, $h_2$ parameters takes values in 0–1, such that $0 < h_1 + h_2 < 1$, then $k$ takes values in 0–1 and is a suitable candidate as kernel, because the matrix $\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j, h_1, h_2)\}$ would be positive definite. The dual (pseudo-)RKHS representation would be

$$y_i = \mathbf{w}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + \sum_{j=1}^{n} k'(\mathbf{x}_i, \mathbf{x}_j, h_1, h_2)\alpha_j + e_i.$$

Then, procedures for inferring the variance components and $h$ parameters outlined in this article would be followed. Research is needed for assessing the adequacy of these approximations. González-Recio *et al.* (2008) present an application of the trinomial kernel.

**Conclusion:** The methods presented here take the whole-genome view of those of Meuwissen *et al.* (2001), Gianola *et al.* (2003), Xu (2003), Yi *et al.* (2003), Ter Braak *et al.* (2005), Wang *et al.* (2005), and Zhang and Xu (2005). The main difference is the attempt to capture unknown forms of interaction between many loci that, arguably, parametric models are not able to explore properly, due to either violation of assumptions (for decomposition of epistatic variance) or inadequate statistical machinery for understanding high-level "physiological epistasis."

An application of the theory presented here is in González-Recio *et al.* (2008). Using mortality rates observed in 200 families of paternal half-sib broilers, these authors compared the predictive ability of a standard parametric mixed model against that from linear regression (fitting additive effects of 24 SNPs), kernel regression, RKHS, and the Bayesian regression model of Xu (2003). For RHKS, the kernel consisted of a similarity score between any two SNP sequences. The five models were contrasted in terms of a fivefold predictive cross-validation. Results indicated an advantage of RKHS, which had a global "accuracy" that was twice as large as the one from the mixed model, was 2.5 times larger than the one attained with the linear regression specification, and exceeded that attained with the procedure of Xu (2003) by 25%. However, predictive cross-validation accuracy was not large, probably due to the very low heritability of the trait used for the case study, chick mortality.

It should be noted that a kernel function explores commonalities in some sense; *e.g.*, in a chromosome model markers in a contiguous position borrow information. In spirit, this is similar to the use of relationship or identity-by-descent matrices between individuals or cultivars in animal and plant breeding, respectively.

Recent developments in nonparametric statistics and machine learning offer exciting avenues for whole-genome analysis of quantitative traits and perhaps suggest a change in analytical paradigms. This theoretical article intends to make a contribution in this direction.

Gustavo de los Campos and Oscar González-Recio are thanked for useful discussions. Support by the Wisconsin Agriculture Experiment Station and by National Science Foundation (NSF) grant DMS-NSF DMS-044371 to D. Gianola is acknowledged. Research was completed while D. Gianola had a Senior Researcher position within a Marie Curie European Transfer of Knowledge–Development project with contract no. MTKD/I-CT-2004-14412. J. B. C. H. M. van Kaam acknowledges support by grant D.M. 305/7303/06 of the Ministero delle Politiche Agricole e Forestali.

## LITERATURE CITED

Aitchison, J., and C. G. G. Aitken, 1976 Multivariate binary discrimination by the kernel method. Biometrika **63:** 413–420.

Aronszajn, N., 1950 Theory of reproducing kernels. Trans. Am. Math. Soc. **68:** 337–404.

Balding, D. J., 2006 A tutorial on statistical methods for population association studies. Nat. Rev. Genet. **7:** 781–791.

Chang, H. L. A., 1988 Studies on estimation of genetic variances under nonadditive gene action. Ph.D. Thesis, University of Illinois, Urbana-Champaign, IL.

Cheverud, J. M., and E. J. Routman, 1995 Epistasis and its contribution to genetic variance components. Genetics **139:** 1455–1461.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics **39:** 859–882.

Craven, P., and G. Wahba, 1979 Smoothing noisy data with spline functions. Num. Math. **31:** 377–403.

Dekkers, J. C. M., and F. Hospital, 2002 The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3:** 22–32.

Elkan, C., 1997 Boosting and naïve Bayesian learning. Technical Report. University of California, San Diego.

Fenster, C. B., and L. F. Galloway, 2000 Population differentiation in an annual legume: genetic architecture. Evolution **54:** 1157–1172.

Fernández, S. A., R. L. Fernando, B. Gulbrandtsen, C. Stricker, M. Schelling *et al.*, 2002 Irreducibility and efficiency of ESIP to sample marker genotypes in large pedigrees with loops. Genet. Sel. Evol. **34:** 537–555.

Fernando, R. L., and M. Grossman, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Fernando, R. L., C. Stricker and R. C. Elston, 1993 An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. Theor. Appl. Genet. **87:** 89–93.

Fox, J., 2005 Introduction to nonparametric regression. Lecture Notes. http://socserv.mcmaster.ca/jfox/Courses/Oxford.

Franklin, I., and R. C. Lewontin, 1970 Is the gene the unit of selection? Genetics **65:** 707–734.

Gianola, D., M. Perez-Enciso and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. Genetics **163:** 347–365.

Gianola, D., R. L. Fernando and A. Stella, 2006 Genomic assisted prediction of genetic value with semi-parametric procedures. Genetics **173:** 1761–1776.

Golub, T. R., D. Slonim, P. Tamayo, C. Huard, M. Gasenbeek *et al.*, 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286:** 531–537.

González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa *et al.*, 2008 Non-parametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics **178:** 2305–2313.

Gu, C., 2002 *Smoothing Spline ANOVA Models*. Springer, New York.

Hartl, D. L., and E. W. Jones, 2005 *Genetics: Analysis of Genes and Genomes*, Ed. 6. Jones & Bartlett, Boston.

Hastie, T. J., and R. J. Tibshirani, 1990 *Generalized Additive Models*. Chapman & Hall, London.

Hayes, B., J. Laerdahl, D. Lien, A. Adzhubei and B. Høyheim, 2004 Large scale discovery of single nucleotide polymorphism (SNP) markers in Atlantic Salmon (Salmo salar). AKVAFORSK, Institute of Aquaculture Research, Aas, Norway. www.mabit.no/pdf/hayes.pdf.

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky, 1999 Bayesian model averaging: a tutorial. Stat. Sci. **14:** 382–417.

Jensen, C. S., and A. Kong, 1999 Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. Am. J. Hum. Genet. **65:** 885–901.

Jensen, C. S., A. Kong and U. Kjaerulff, 1995 Blocking Gibbs sampling in very large probabilistic expert systems. Int. J. Hum. Comp. Stud. **42:** 647–666.

Kempthorne, O., 1954 The correlation between relatives in a random mating population. Proc. R. Soc. Lond. Ser. B **143:** 103–113.

Kerr, R. J., and B. P. Kinghorn, 1996 An efficient algorithm for segregation analysis in large populations. J. Anim. Breed. Genet. **113:** 457–469.

Kimeldorf, G., and G. Wahba, 1971 Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33:** 82–95.

Kimura, M., 1965 Attainment of quasi-linkage equilibrium when gene frequencies are changing. Genetics **52:** 875–890.

Long, N., D. Gianola, G. J. M. Rosa, K. Weigel and S. Avendaño, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. **124:** 377–389.

Mallick, B. K., D. Ghosh and M. Ghosh, 2005 Bayesian classification of tumours by using gene expression data**.** J. R. Stat. Soc. B **67:** 219–234.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Is it possible to predict the total genetic merit under a very dense marker map? Genetics **157:** 1819–1829.

O'Sullivan, F., B. S. Yandell and W. J. Raynor, 1986 Automatic smoothing of regression functions in generalized linear models. J. Am. Stat. Assoc. **81:** 96–103.

Quaas, R. L., and E. J. Pollak, 1980 Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. **51:** 1277–1287.

Racine, J., and Q. Li, 2004 Nonparametric estimation of regression functions with both categorical and continuous data. J. Econometrics **119:** 99–130.

Rasmussen, C. E., and C. K. I. Williams, 2006 *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Sheehan, N., and A. Thomas, 1993 On the irreducibility of a Markov chain defined on a space of genotype configurations by a sample scheme. Biometrics **49:** 163–175.

Silverman, B. W., 1986 *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Soller, M., and J. Beckmann, 1982 Restriction fragment length polymorphisms and genetic improvement, pp. 396–404 in *Proceedings of the 2nd World Congress on Genetics Applied to Livestock Production*, Vol. 6. Editorial Garsi, Madrid.

Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Sorensen, D., and R. Waagepetersen, 2003 Normal linear models with genetically structured residual variance heterogeneity: a case study. Genet. Res. **82:** 207–222.

Stricker, C., M. Schelling, F. Du, I. Hoeschele, S. A. Fernandez *et al.*, 2002 A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. Proceedings of 7th World Congress on Genetics Applied to Livestock Production, INRA, Castanet-Tolosan, France, CD-ROM Communication no. 21–12.

Takezawa, K., 2005 *Introduction to Non-Parametric Regression*. Wiley-Interscience, Hoboken, NJ.

Ter Braak, C. J. F., M. Boer and M. Bink, 2005 Extending Xu's (2003) Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics **170:** 1435–1438.

Van Arendonk, J. A. M., C. Smith and B. W. Kennedy, 1989 Method to estimate genotype probabilities at individual loci in farm livestock. Theor. Appl. Genet. **78:** 735–740.

Wahba, G., 1990 *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G., 1999 Support vector machines, reproducing kernel Hilbert spaces and the randomized GAVC, pp. 68–88 in *Advances in Kernel Methods*, edited by B. Schölkopf, C. Burges and A. Smola. MIT Press, Cambridge, MA.

Wahba, G., 2002 Soft and hard classification by reproducing kernel Hilbert spaces methods. Proc. Natl. Acad. Sci. USA **99:** 16524–16530.

Wang, C. S., J. J. Rutledge and D. Gianola, 1993 Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. **25:** 41–62.

Wang, C. S., J. J. Rutledge and D. Gianola, 1994 Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. **26:** 91–115.

Wang, H., Y. M. Zhang, X. Li, G. Masinde, S. Mohan *et al.*, 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics **170:** 465–480.

Wong, G. K., B. Liu, J. Wang, Y. Zhang, X. Yang *et al.*, 2004 A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature **432:** 717–722.

Wood, S. N., 2006 *Generalized Additive Models*. Chapman & Hall/CRC Press, Boca Raton, FL.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics **163:** 789–801.

Yi, N., G. Varghese and D. A. Allison, 2003 Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics **164:** 1129–1138.

Zhang, Y., and S. Xu, 2005 A penalized maximum-likelihood method for estimating epistatic effects of QTL. Heredity **95:** 96–104.

Communicating editor: J. B. Walsh

## APPENDIX

In an Euclidean space of dimension $n$, the dot product between vectors $\mathbf{v}$ and $\mathbf{w}$ is $\sum_{i=1}^{n} v_i w_i$, and the norm is $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$. The inner product generalizes the dot product to vectors of infinite dimension. For instance, in a vector space of real functions with domain $[a, b]$, the inner product is

$$\langle g_1, g_2 \rangle = \int_a^b g_1(x) g_2(x) dx,$$

and the norm is $\|g_1\| = \sqrt{\int_a^b g_1(x)^2 dx}$. If $x$ is a continuous random variable with probability density function $p(x)$, the inner product (Hastie and Tibshirani 1990) is

$$\langle g_1, g_2 \rangle = \int_a^b g_1(x) g_2(x) p(x) dx = E[g_1(x) g_2(x)]. \qquad \text{(A1)}$$

Consider now the choice of basis functions for $\mathbf{x}$, that is, a transformation of the input (SNP) space to be used as regressors in the nonparametric regression. A kernel $k(\alpha, x)$ is a function that maps inputs $\alpha$, $x$ into some space, and the kernel is said to be symmetric if $k(\alpha, x) = k(x, \alpha)$ (Rasmussen and Williams 2006). A kernel $k(\mathbf{x}, \mathbf{t})$ involving random vectors $\mathbf{x}$, $\mathbf{t}$ is positive definite if

$$\int k(\mathbf{x}, \mathbf{t}) g(\mathbf{x}) g(\mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} > 0,$$

for all functions $g$, where $p(\mathbf{x}, \mathbf{t})$ is a joint density. An eigenfunction of positive-definite kernel $k$ with eigenvalue $\lambda$ satisfies the equation

$$\int k(\mathbf{x}, \mathbf{x}^*) \phi(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda \phi(\mathbf{x}^*).$$

There are an infinite number of eigenfunctions $\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots$ with an ordering corresponding to $\lambda_1 \geq \lambda_2 \ldots$. The eigenfunctions are orthogonal and can be normalized to satisfy

$$\begin{aligned} &\int \sqrt{\lambda_i} \phi_i(\mathbf{x}) \sqrt{\lambda_j} \phi_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = 0 \quad \text{for } i \neq j; \\ &\int \sqrt{\lambda_i} \phi_i(\mathbf{x}) \sqrt{\lambda_i} \phi_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = 1 \quad \text{for } i = 1, 2, \ldots, \infty. \qquad \text{(A2)} \end{aligned}$$

Mercer's theorem (Rasmussen and Williams 2006) enables expressing a kernel in terms of its eigendecomposition such that

$$k(\mathbf{x}, \mathbf{x}^*) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}^*). \qquad \text{(A3)}$$

If all eigenvalues are positive, the sum is infinite. If, on the other hand, the sum terminates at some value $p$, say, this yields a degenerate kernel of rank $p$. For example, in a linear random-regression model with coefficients $\mathbf{b} \sim (\mathbf{0}, \mathbf{B})$ in which the $\mathbf{x}$ variables are transformed into orthonormal basis functions $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_p(\mathbf{x})]'$, the regression function $\boldsymbol{\phi}(\mathbf{x})'\mathbf{b}$ evaluated at points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ generates the covariance matrix $\mathbf{K}_{n \times n} = \{k(i, j) = \boldsymbol{\phi}(\mathbf{x}_i)'\mathbf{B}\boldsymbol{\phi}(\mathbf{x}_j)\}$ of rank $n$; recall that, in our situation, $n < < p$. Here, the eigendecomposition becomes that of a covariance matrix of finite dimension. In general, $k(i, j)$ is called a covariance function, and $\mathbf{K}$ is an infinite-dimensional generalization of a covariance matrix. The kernel is clearly symmetric, as $k(i, j) = k(j, i)$.

The notation $k(\mathbf{x}, .)$ and $g(.)$ will denote that, at some fixed point $\mathbf{x}^*$, these functions take values $k(\mathbf{x}, \mathbf{x}^*)$ and $g(\mathbf{x}^*)$, respectively. A space of functions $\mathcal{H}$ is a RKHS with kernel $k$ if the two following conditions are met (Wahba 2002; Mallick *et al.* 2005; Rasmussen and Williams 2006):

For every $\mathbf{x}$, $k(\mathbf{x}, .)$ is in the Hilbert space $\mathcal{H}$.
For all $\mathbf{x}$ and for every $g$ in $\mathcal{H}$ the inner product $\langle k(\mathbf{x}, .), g(.) \rangle = g(\mathbf{x})$ holds; this is a reproducing property, in some sense.

Consider now a Hilbert space constituted by linear combinations of the orthonormal eigenfunctions; *e.g.*, $f(\mathbf{x}) = \sum_{i=1}^{p} f_i \phi_i(\mathbf{x})$ and $v(\mathbf{x}) = \sum_{i=1}^{p} v_i \phi_i(\mathbf{x})$, where $f_i$ and $v_i$ are loadings or regression coefficients on the eigenfunctions. Using definition (A1) and orthonormality condition (A2), the inner product is

$$\begin{aligned} \langle f, v \rangle_{\mathcal{H}} &= \int \left( \sum_{i=1}^{p} f_i \phi_i(\mathbf{x}) \right) \left( \sum_{j=1}^{p} v_j \boldsymbol{\phi}_j(\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x} \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} f_i v_j \int \phi_i(\mathbf{x}) \boldsymbol{\phi}_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^{p} \frac{f_i v_i}{\lambda_i}. \end{aligned}$$

Because $\sqrt{\langle f, f \rangle_{\mathcal{H}}} = \sqrt{\sum_{i=1}^{p} (f_i^2 / \lambda_i)}$, the Hilbert space has a norm. In its infinite-dimensional form $(p \to \infty)$, this implies that the sequence of $f_i$ coefficients must decay quickly, which imposes smoothness conditions (Rasmussen and Williams 2006).

Next, examine if $k(\mathbf{x}, .)$ is in the Hilbert space spanned by functions such as $f$ and $v$ above. First, recall the eigendecomposition of kernel $k(\mathbf{x}, \mathbf{x}^*) = \sum_{j=1}^{p} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}^*)$. Using again the orthogonality property, the inner product

$$\begin{aligned} &\langle f(.), k(\mathbf{x}, .) \rangle_{\mathcal{H}} \\ &= \int \left( \sum_{i=1}^{p} f_i \phi_i(\mathbf{x}^*) \right) \left( \sum_{j=1}^{p} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}^*) \right) p(\mathbf{x}^*) d\mathbf{x}^* \\ &= \sum_{i=1}^{p} f_i \phi_i(\mathbf{x}) \lambda_i \int \phi_i(\mathbf{x}^*) \phi_i(\mathbf{x}^*) p(\mathbf{x}^*) d\mathbf{x}^* \\ &= \sum_{i=1}^{p} f_i \phi_i(\mathbf{x}) = f(\mathbf{x}). \qquad \text{(A4)} \end{aligned}$$

This shows that the inner product between the function and the kernel reproduces the function. Also

$$\begin{aligned}\langle k(\mathbf{x}, .),\ k(\mathbf{x}^*, .)\rangle_{\mathcal{H}} \\ &= \int\left(\sum_{i=1}^{p}\lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{t})\sum_{j=1}^{p}\lambda_j\phi_j(\mathbf{x}^*)\phi_j(\mathbf{t})\right)p(\mathbf{t})\,d\mathbf{t} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_i\lambda_j\int\phi_i(\mathbf{x})\phi_i(\mathbf{t})\phi_j(\mathbf{x}^*)\phi_j(\mathbf{t})p(\mathbf{t})\,d\mathbf{t} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_i\phi_i(\mathbf{x})\lambda_j\phi_j(\mathbf{x}^*)\int\phi_i(\mathbf{t})\phi_j(\mathbf{t})p(\mathbf{t})\,d\mathbf{t}.\end{aligned}$$

Because the eigenfunctions are orthogonal, terms where $i \neq j$ vanish, and recall from (A2) that $\int\phi_i(\mathbf{t})\phi_i(\mathbf{t})p(\mathbf{t}) = \lambda_i^{-1}$, so

$$\langle k(\mathbf{x}, .),\ k(\mathbf{x}^*, .)\rangle_{\mathcal{H}} = \sum_{i=1}^{p}\lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{x}^*) = k(\mathbf{x}, \mathbf{x}^*). \tag{A5}$$

Hence, the inner product of kernels $k(\mathbf{x}, .)$, $k(\mathbf{x}^*, .)$ produces kernel $k(\mathbf{x}, \mathbf{x}^*)$. This demonstrates that the Hilbert space constituted by linear combinations of the eigenfunctions of $k$ has the reproducing kernel properties.