doi:10.1017/S0016672308009890 Printed in the United Kingdom

Inferring genetic values for quantitative traits non-parametrically

DANIEL GIANOLA[1,2,3,4]* AND GUSTAVO DE LOS CAMPOS[1]

[1] *Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA*
[2] *Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway*
[3] *Institut National de la Recherche Agronomique, UR631 Station d'Amélioration Génétique des Animaux, BP 52627, 32326 Castanet-Tolosan, France*
[4] *Institut für Tierzucht und Haustiergenetik, Georg-August-Universität, Göttingen, Federal Republic of Germany*

(*Received 8 July 2008 and in revised form 17 September and 2 October 2008*)

Summary

Inferences about genetic values and prediction of phenotypes for a quantitative trait in the presence of complex forms of gene action, issues of importance in animal and plant breeding, and in evolutionary quantitative genetics, are discussed. Current methods for dealing with epistatic variability via variance component models are reviewed. Problems posed by cryptic, non-linear, forms of epistasis are identified and discussed. Alternative statistical procedures are suggested. Non-parametric definitions of additive effects (breeding values), with and without employing molecular information, are proposed, and it is shown how these can be inferred using reproducing kernel Hilbert spaces regression. Two stylized examples are presented to demonstrate the methods numerically. The first example falls in the domain of the infinitesimal model of quantitative genetics, with additive and dominance effects inferred both parametrically and non-parametrically. The second example tackles a non-linear genetic system with two loci, and the predictive ability of several models is evaluated.

1. Introduction

The problem considered here is that of inferring genetic values and of predicting phenotypes for a quantitative trait under complex forms of gene action, an issue of importance in animal and plant breeding, and in evolutionary quantitative genetics (Lynch & Walsh, 1998). The discussion is streamlined as follows. Current methods for dealing with epistatic variability via variance component models are discussed in section 2. Problems posed by cryptic, non-linear, forms of epistasis are identified in section 3. Section 4 proposes non-parametric definitions of additive effects (breeding values), with and without employing molecular information, and shows how these could be inferred using reproducing kernel Hilbert spaces (RKHS) regression models. Sections 5 and 6 present stylized examples to demonstrate the methods. The first example uses the infinitesimal model of quantitative genetics, and the second one tackles a non-linear genetic system. The paper ends with concluding comments.

2. Extant theory

A standard decomposition of phenotypic value in quantitative genetics (Falconer & Mackay, 1996) is

$$y = \mu + a + d + i + e,$$

where a, d and i are additive, dominance and epistatic effects, respectively, and e is a residual, reflecting environmental (residual) variability. This linear decomposition can also be used to describe variability of latent variables, especially if assumed Gaussian (e.g. Dempster & Lerner, 1950; Gianola, 1982). The i effect can be decomposed into additive × additive, additive × dominance, dominance × dominance, etc., deviates. In what has been termed 'statistical epistasis' (Cheverud & Routman, 1995), these deviates are assumed to be random draws from some distributions

* Corresponding author. Tel: +1 6082652054. Fax. +1 608262 5157. e-mail: gianola@ansci.wisc.edu

representing 'interactions' between loci. Under certain assumptions (Cockerham, 1954; Kempthorne, 1954), the deviates are uncorrelated, leading to the standard variance decomposition $\sigma^2=\sigma^2_a+\sigma^2_d+\sigma^2_{aa}+\sigma^2_{ad}+\sigma^2_{dd}+\cdots+\sigma^2_e$, where the σ^2s are variance components, e.g., σ^2_{ad} represents the contribution of additive by dominance effects to variance. The sum of all variance components other than σ^2_a, σ^2_d and σ^2_e is interpreted as variance 'due to epistasis'. For the purposes of discussing problems this representation has, it suffices to assume that the only relevant epistatic effects are those of an additive × additive, additive × dominance and dominance × dominance nature.

(i) *Henderson's methods for predicting epistatic effects*

Using the Cockerham–Kempthorne (hereinafter, CK) assumptions, Henderson (1985) extended their model to the infinitesimal domain, and vectorially, by writing

$$\begin{aligned}\mathbf{y}&=\mathbf{X}\boldsymbol{\beta}+(\mathbf{a}+\mathbf{d}+\mathbf{i}_{aa}+\mathbf{i}_{ad}+\mathbf{i}_{dd})+\mathbf{e}\\&=\mathbf{X}\boldsymbol{\beta}+\mathbf{g}+\mathbf{e},\end{aligned}\tag{1}$$

where $\boldsymbol{\beta}$ is some nuisance location vector (equal to μ if it contains a single element); $\mathbf{X}$ is a known incidence matrix; $\mathbf{a}$ and $\mathbf{d}$ are vectors of additive and dominance effects, respectively; $\mathbf{i}_{aa}$, $\mathbf{i}_{ad}$ and $\mathbf{i}_{dd}$ are epistatic effects, and $\mathbf{g}=\mathbf{a}+\mathbf{d}+\mathbf{i}_{aa}+\mathbf{i}_{ad}+\mathbf{i}_{dd}$ is the 'total' genetic value. Assuming that $\mathbf{g}$ and $\mathbf{e}$ are uncorrelated, the variance–covariance decomposition is

$$\mathbf{V}_y=\mathbf{V}_g+\mathbf{V}_e,\tag{2}$$

where $\mathbf{V}_y$,$\mathbf{V}_g$ and $\mathbf{V}_e$ are the phenotypic, genetic and residual variance–covariance matrices, respectively. Further,

$$\mathbf{V}_g=\mathbf{A}\sigma^2_a+\mathbf{D}\sigma^2_d+(\mathbf{A}\#\mathbf{A})\sigma^2_{aa}+(\mathbf{A}\#\mathbf{D})\sigma^2_{ad}+(\mathbf{D}\#\mathbf{D})\sigma^2_{dd}.\tag{3}$$

Here, $\mathbf{A}$ is the numerator relationship matrix; $\mathbf{D}$ is a matrix due to dominance relationships which can be computed from entries in $\mathbf{A}$, and the remaining matrices involve Hadamard (element by element) products of matrices $\mathbf{A}$ or $\mathbf{D}$. Thus, under CK, all matrices can be computed from elements of $\mathbf{A}$, as noted by Henderson (1985), because absence of inbreeding and of linkage disequilibrium are assumed. Cockerham (1956) and Schnell (1963) gave formulae for covariances between relatives in the presence of linkage, and difficulties posed by linkage disequilibrium are discussed by Gallais (1974).

In CK–Henderson, with the extra assumption that all genetic effects (having null means) and the data (with mean vector $\mathbf{X}\boldsymbol{\beta}$) follow a multivariate normal distribution with known dispersion components, one has

$$E(\mathbf{a}|\mathbf{g})=\mathrm{Cov}(\mathbf{a},\mathbf{g})\mathbf{V}_g^{-1}\mathbf{g}=\sigma^2_a\mathbf{A}\mathbf{V}_g^{-1}\mathbf{g}.\tag{4}$$

The best predictor (Henderson, 1973; Bulmer, 1980) of additive genetic value in the mean-squared error sense is

$$\begin{aligned}E(\mathbf{a}|\mathbf{y})&=E_{g|y}[E(\mathbf{a}|\mathbf{g})]=E_{g|y}[\sigma^2_a\mathbf{A}\mathbf{V}_g^{-1}\mathbf{g}]\\&=\sigma^2_a\mathbf{A}\mathbf{V}_g^{-1}[E_{g|y}(\mathbf{g})],\end{aligned}\tag{5}$$

where $E_{g|y}(\mathbf{g})=\mathbf{V}_g\mathbf{V}_y^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ is the best predictor of $\mathbf{g}$, and the matrix of regression coefficients $\mathbf{V}_g\,\mathbf{V}_y^{-1}$ is the multidimensional counterpart of heritability in the broad sense. With known variance components, $\boldsymbol{\beta}$ is typically estimated by generalized least-squares (equivalently, by maximum likelihood under normality) as $\hat{\boldsymbol{\beta}}$. Then, 'the empirical best predictor' of the vector of additive effects is taken to be

$$\begin{aligned}\hat{E}(\mathbf{a}|\mathbf{y})&=\sigma^2_a\mathbf{A}\mathbf{V}_g^{-1}[\hat{E}_{g|y}(\mathbf{g})]\\&=\sigma^2_a\mathbf{A}\mathbf{V}_g^{-1}\mathbf{V}_g\mathbf{V}_y^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\\&=\sigma^2_a\mathbf{A}\mathbf{V}_y^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}).\end{aligned}\tag{6}$$

This is precisely the best linear unbiased predictor (BLUP) of additive merit when model (1) holds. Likewise, the BLUP of additive × dominance deviations is

$$\hat{E}(\mathbf{i}_{ad}|\mathbf{y})=\sigma^2_{ad}(\mathbf{A}\#\mathbf{D})\mathbf{V}_y^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

and so on.

The BLUPs of $\mathbf{a}$, $\mathbf{d}$, $\mathbf{i}_{aa}$, $\mathbf{i}_{ad}$ and $\mathbf{i}_{dd}$ in (1) can be computed simultaneously using Henderson's mixed model equations, but this requires forming the inverse matrices $\mathbf{A}^{-1}$, $\mathbf{D}^{-1}$, $(\mathbf{A}\#\mathbf{A})^{-1}$, $(\mathbf{A}\#\mathbf{D})^{-1}$ and $(\mathbf{D}\#\mathbf{D})^{-1}$. In a general setting, most of these inverses are impossible to obtain, contrary to that of $\mathbf{A}$, which can be written directly from a genealogy.

(ii) *Reformulation of Henderson's approach*

Equivalently, as in de los Campos *et al.* (2008), one may rewrite (1) as

$$\begin{aligned}\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{A}\mathbf{a}^*+\mathbf{D}\mathbf{d}^*+(\mathbf{A}\#\mathbf{A})\,\mathbf{i}^*_{aa}\\+(\mathbf{A}\#\mathbf{D})\,\mathbf{i}^*_{ad}+(\mathbf{D}\#\mathbf{D})\,\mathbf{i}^*_{dd}+\mathbf{e},\end{aligned}\tag{7}$$

where $\mathbf{a}^*=\mathbf{A}^{-1}\mathbf{a}\sim(\mathbf{0},\,\mathbf{A}^{-1}\sigma^2_a),\ldots,\,\mathbf{i}^*_{dd}=(\mathbf{D}\#\mathbf{D})^{-1}\mathbf{i}_{dd}\sim(\mathbf{0},\,(\mathbf{D}\#\mathbf{D})^{-1}\sigma^2_{dd})$. Then, the BLUP of any of the transformed genetic effects can be found by solving a system of mixed linear model equations that does not involve inverses of any of the genetic variance–covariance matrices. For example, the β-equation is

$$\begin{aligned}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}'\mathbf{A}\hat{\mathbf{a}}^*+\mathbf{X}'\mathbf{D}\hat{\mathbf{d}}^*+\mathbf{X}'(\mathbf{A}\#\mathbf{A})\,\hat{\mathbf{i}}^*_{aa}\\+\mathbf{X}'(\mathbf{A}\#\mathbf{D})\,\hat{\mathbf{i}}_{ad}+\mathbf{X}'(\mathbf{D}\#\mathbf{D})\,\hat{\mathbf{i}}_{dd}=\mathbf{X}'\mathbf{y},\end{aligned}$$

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

and the $\mathbf{i}_{aa}$-equation is

$$\begin{aligned}&(\mathbf{A}\#\mathbf{A})\,\mathbf{X}\hat{\boldsymbol{\beta}}+(\mathbf{A}\#\mathbf{A})\,\mathbf{A}\hat{\mathbf{a}}^{*}+(\mathbf{A}\#\mathbf{A})\mathbf{D}\hat{\mathbf{d}}^{*}\\&\quad+\left[(\mathbf{A}\#\mathbf{A})^{2}+\frac{\sigma_e^2}{\sigma_{aa}^2}(\mathbf{A}\#\mathbf{A})\right]\hat{\mathbf{i}}_{aa}^{*}+(\mathbf{A}\#\mathbf{A})\,(\mathbf{A}\#\mathbf{D})\,\hat{\mathbf{i}}_{ad}^{*}\\&\quad+(\mathbf{A}\#\mathbf{A})\,(\mathbf{D}\#\mathbf{D})\,\hat{\mathbf{i}}_{dd}^{*}=(\mathbf{A}\#\mathbf{A})\mathbf{y}.\end{aligned}$$

Once the BLUPs of the $(*)$ genetic effects are obtained, linear invariance leads, for example, to $\hat{\mathbf{a}}=\mathbf{A}\hat{\mathbf{a}}^{*}$, and to $\hat{\mathbf{i}}_{dd}=(\mathbf{D}\#\mathbf{D})\,\hat{\mathbf{i}}_{dd}^{*}$. A computational difficulty here is that $\mathbf{A}$ is typically not sparse, so all $\mathbf{A}\#\mathbf{A}$, $\mathbf{A}\#\mathbf{D}$, etc., are not sparse either. Note that the equations for the genetic effects in the 'reparameterized' model are similar to those in Henderson (1984). For example, if in an animal model the a-equation

$$\left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)\hat{\mathbf{a}}=\mathbf{Z}'\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

is premultiplied by $\mathbf{A}$, one obtains

$$\left(\mathbf{A}+\mathbf{I}\frac{\sigma_e^2}{\sigma_a^2}\right)\hat{\mathbf{a}}=\mathbf{A}\mathbf{Z}'\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

(iii) *Bayesian implementation of the reformulated model*

In the standard Bayesian linear model (Wang *et al.*, 1993, 1994; Sorensen & Gianola, 2002), $\hat{\mathbf{a}}^{*}$, $\hat{\mathbf{i}}_{dd}^{*}$, etc., are mean vectors of corresponding conditional posterior distributions, for which sampling procedures are well known. Further, with scaled inverse chi-square priors assigned to variance components, Gibbs sampling is straightforward, and does not require forming inverses either. For example, a draw from the conditional posterior distribution of the dominance × dominance variance component is obtained under the alternative parameterization as

$$\sigma_{dd}^2 \sim [\mathbf{i}_{dd}'^{*}(\mathbf{D}\#\mathbf{D})\,\mathbf{i}_{dd}^{*}+\nu_{dd}S_{dd}^2],$$

where $\mathbf{i}_{dd}^{*}$is a draw from its conditional posterior distribution, given everything else (Sorensen & Gianola, 2002) and ν_{dd} and S_{dd}^2 are hyper-parameters.

While computations may still be formidable, inversion of the genetic covariance matrices is circumvented. Apart from computational issues, an important question, however, is whether or not the CK–Henderson construct can cope with complex genetic systems effectively.

3. Confronting complexity

Dealing with non-additive genetic variability may be much more difficult than what equations (1) and (3) suggest. Theoretically, at least in CK, epistatic variance can be partitioned into 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 or assortative mating (Cockerham, 1954; Kempthorne, 1954). All these assumptions are violated in mature and in breeding programmes. Also, estimation of non-additive components of variance is very difficult, even under standard assumptions (Chang, 1988), leading to imprecise inference. Linkage disequilibrium induces covariances between different types of effects; algebraically heroic attempts to deal with this problem are in Weir & Cockerham (1977) and Wang & Zeng (2006). Gallais (1974) derived expressions aimed to describe the impact of linkage disequilibrium on partition of genetic variance. The number of parameters is large, which makes estimation of the needed dispersion components intractable.

The question of whether or not standard random effects models for quantitative traits can account accurately for non-linear (non-additive) relationships between infinitesimal genotypes and phenotypes remains open. It is argued subsequently that parametric models cannot handle well complexity resulting from interactions between the hundreds or even thousands of genes expected to affect multifactorial traits, such as liability or resistance to disease. This was discussed by Templeton (2000), Gianola *et al.* (2006) and Gianola and van Kaam (2008).

In the standard theory for non-additive gene action, e.g. epistasis, interaction effects enter linearly into the phenotype, that is, the partial derivative of the model with respect to any effect, be it additive or of an interactive nature, is a constant that does not involve any genetic effect. This theory is unrealistic in non-linear systems, such as those used in metabolic control theory (Bost *et al.*, 1999). Arguing from the perspective of 'perturbation in analysis', Feldman & Lewontin (1975) advanced the argument that linear (i.e. analysis of variance) models are 'local', presumably in the following sense. Suppose that the expected value of phenotype y, given some genetic values $\mathbf{G}$ (genotypes are discrete but their effects are continous), is some unknown function of effects of genotypes at L loci represented as $f(G_1, G_2, \ldots, G_L)$. A second-order approximation to the surface of means yields

$$\begin{aligned}E(y|\mathbf{G})&=f(G_1,G_2,\,\ldots,G_L)\\&\approx f(\bar{G}_1,\bar{G}_2,\,\ldots,\bar{G}_L)+\mathbf{f}'\,(\bar{G}_1,\bar{G}_2,\,\ldots,\bar{G}_L)(\mathbf{G}-\bar{\mathbf{G}})\\&\quad+\frac{1}{2}(\mathbf{G}-\bar{\mathbf{G}})'\left\{\frac{\partial^2}{\partial G_i\partial G_j}f(G_1,G_2,\,\ldots,G_L)\right\}_{\mathbf{G}=\bar{\mathbf{G}}}\\&\quad\times(\mathbf{G}-\bar{\mathbf{G}}),\end{aligned}$$

where $\bar{G}_i$ is the mean value of the effect of locus i (typically taken to be 0); $\mathbf{G}=(G_1, G_2, \ldots, G_L)'$ is a column

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

vector and $\bar{\mathbf{G}} = E(\mathbf{G})$; $\mathbf{f}'(\bar{G}_1, \bar{G}_2, \ldots, \bar{G}_L)$ is the row vector of first derivatives of $f(.)$ with respect to **G** and

$$\left\{\frac{\partial^2}{\partial G_i \partial G_j} f(G_1, G_2, \cdots, G_L)\right\}_{\mathbf{G}=\bar{\mathbf{G}}}$$

is the matrix of second derivatives, both evaluated at $\bar{\mathbf{G}}$ in the approximation.

A linear-on-variance-components decomposition of variability may not be enlightening at all when a phenotype results, say, from a sum of sine and cosine waves. To illustrate, suppose that non-linearity (non-additivity) enters as $E\,(y|G_1, G_2, G_3) = G_1 \exp\,(G_2G_3)$. A log-transformation of the expected phenotypic value of individuals with genetic value (G_1, G_2, G_3) would suggest linearity (in a log-scale) with respect to locus 1, and interaction between loci 2 and 3. For this hypothetical model, the first derivatives are

$$\mathbf{f}(G_1, G_2, G_3) = \begin{bmatrix} e^{G_2G_3} \\ G_1G_3e^{G_2G_3} \\ G_1G_2e^{G_2G_3} \end{bmatrix} = \begin{cases} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, & \text{when } \bar{G}_1, \bar{G}_2, \bar{G}_3 = 0, \\ \begin{bmatrix} e \\ e \\ e \end{bmatrix}, & \text{when } \bar{G}_1, \bar{G}_2, \bar{G}_3 = 1, \end{cases}$$

and the matrix of second derivatives is

$$\left\{\frac{\partial^2}{\partial G_i \partial G_j} f(.)\right\} = \begin{bmatrix} 0 & G_3e^{G_2G_3} & G_2e^{G_2G_3} \\ G_3e^{G_2G_3} & G_1G_3^2e^{G_2G_3} & G_1e^{G_2G_3}(G_2G_3+1) \\ G_2e^{G_2G_3} & G_1e^{G_2G_3}(G_2G_3+1) & G_1G_2^2e^{G_2G_3} \end{bmatrix} = \begin{cases} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & \text{when } \bar{G}_1, \bar{G}_2, \bar{G}_3 = 0, \\ \begin{bmatrix} 0 & e & e \\ e & e & 2e \\ e & 2e & e \end{bmatrix}, & \text{when } \bar{G}_1, \bar{G}_2, \bar{G}_3 = 1. \end{cases}$$

In the neighbourhood of 0 for all effects, the local approximation yields

$$E\,(y|G_1, G_2, G_3) \approx G_1. \tag{8}$$

On the other hand, near 1 the local approximation produces

$$E\,(y|G_1, G_2, G_3) \approx e\left(3 - G_1 - 3G_2 - 3G_3 + \frac{1}{2}G_2^2 + \frac{1}{2}G_3^2 + G_1G_2 + G_1G_3 + 2G_2G_3\right), \tag{9}$$

A local approximation near 0 suggests that phenotypes are linear on the effect of locus 1, while an approximation in the neighbourhood of 1 points towards 'dominance' at loci 2 and 3, and at 2 – factor epistasis involving loci (1, 2), (1, 3) and (2, 3). This relates to work by Kojima (1959), who studied conditions for equilibria when a fitness surface was affected by epistasis. He assumed free recombination, random mating and constancy of genotypic values. The arguments leading to (8) and (9) indicate that constancy of genotypic values is not tenable when local approximations are used to study the surface.

To the extent that a linear model provides a 'local' approximation only, it may not be surprising why attaining an understanding of epistasis within the classical paradigm has been elusive. Kempthorne (1978) disagrees with this view, however, although his argument seems more a defence of the technique of the analysis of variance (ANOVA) *per se* than of the lack of ability of a linear model to describe complex, interacting, systems. However, even if a linear model holds at least locally, a standard fixed effects ANOVA of a highly dimensional, multifactorial system is not feasible, because one 'runs out' of degrees of freedom (*df*) (Gianola *et al.*, 2006). Nevertheless, methodologies for dealing with complex epistatic systems are becoming available and these include, for example, machine learning, regularized neural networks (Lee, 2004), neural networks optimized with grammatical evolution computations (Motsinger-Reif *et al.*, 2008) and non-parametric regression (e.g. Gianola *et al.*, 2006; Gianola & van Kaam, 2008). The latter is discussed in what follows.

4. Non-parametric breeding value

There is a wide collection of procedures for non-parametric regression available. In particular, statistical models based on RKHS have been useful, *inter alia*, for regression (Wahba, 1990) and classification (Vapnik, 1998). In many respects, RKHS regression is the 'mother' of non-parametric functional data analysis, since it includes splines, hard and soft classification and even best linear unbiased prediction (de los Campos *et al.*, 2008) as special cases. Contrary to many ad-hoc forms of non-parametric modelling, RKHS is a variational method based on maximizing penalized likelihoods over a rich space of functions defined on a Hilbert space. Also, it uses flexible kernels, which can be adapted to many different circumstances, and allows for varying classes of information inputs, e.g. pedigrees, continuous valued covariates and molecular markers of any type. The Bayesian view of RKHS regression has been used to motivate the methodology using Gaussian processes (Rasmussen & Williams, 2006). In quantitative genetics, Gianola *et al.* (2006) and Gianola & van Kaam

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

(2008) suggested this approach for incorporating information on dense whole-genome markers into models for prediction of genetic value of animals or plants for quantitative traits. The dense molecular data (e.g. SNPs) enter as covariates into a kernel (incidence) square matrix whose dimension is equal to the number of individuals with genotype information available. Thus, the dimensionality of the problem is reduced drastically, from that given by the number of SNPs to the number of individuals genotyped, typically much lower. González-Recio *et al.* (2008*a*, *b*) present applications to mortality rate and feed conversion efficiency in broilers. In the absence of kernels based on substantive theory, genetic interactions are dealt with in RKHS in some form of 'black box'. However, the focus of non-parametric regression is prediction rather than inference, an issue that will be emphasized in this paper. Relationships between RKHS regression and classical models of quantitative genetics are discussed by de los Campos *et al.* (2008). A question that arises naturally in plant and animal breeding is whether or not measures of breeding value can be derived from an RKHS regression model.

Briefly, a RKHS regression can be represented in terms of the linear model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{K_h}\boldsymbol{\alpha}+\mathbf{e}, \tag{10}$$

where $\mathbf{X}\boldsymbol{\beta}$ is as before; $\mathbf{K_h}=\{k(\mathbf{x}_i,\mathbf{x}_j,\mathbf{h})\}$ is an $n\times n$ symmetric positive-definite matrix of kernel entries dependent on covariates (e.g. SNPs) $\mathbf{x}_i$ and $\mathbf{x}_j$ (subscripts i and j define the individual whose phenotype is considered and some other individual in the sample, respectively), and possibly on a set of bandwith parameters $\mathbf{h}$ which must be tuned in some manner, as noted below. González-Recio *et al.* (2008*a*) discuss kernels that do not involve any $\mathbf{h}$. Further, $\boldsymbol{\alpha}$ in (10) is a set of non-parametric regression coefficients, one for each individual in the sample of data. For instance, if SNP genotypes enter into $k(\mathbf{x}_i,\mathbf{x}_j,\mathbf{h})$, then $\mathbf{K_h}\boldsymbol{\alpha}$ can be construed as a vector of genetic effects marked by SNPs.

The choice of kernel $k(\mathbf{x}_i,\mathbf{x}_j,\mathbf{h})$ is absolutely critical for attaining good predictions in RKHS regression, and it can be addressed via some suitable model comparison, e.g. by cross-validation. Each kernel is associated with a space of functions, and de los Campos *et al.* (2008) describe conditions under which a kernel may be expected to work well. For example, if a Hilbert space of functions associated with a given kernel spans functions of additive and non-additive genetic effects, the model would be expected to capture such effects. Otherwise, predictions may be very poor, even worse than those attained with a linear model (which, in some cases will also be a RKHS regression).

It can be shown that the solution to the RKHS regression problem is obtained by assuming that $\boldsymbol{\alpha}\sim\mathrm{N}(\mathbf{0},\mathbf{K_h}^{-1}\sigma^2_\alpha)$, where σ^2_α is a variance component entering into the smoothing parameter $\frac{\sigma^2_e}{\sigma^2_\alpha}$, and by solving

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'K_h}\\ \mathbf{K_h'X} & \mathbf{K_h'K_h}+\dfrac{\sigma^2_e}{\sigma^2_\alpha}\mathbf{K_h}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\alpha}}\end{bmatrix}=\begin{bmatrix}\mathbf{X'y}\\ \mathbf{K_h'y}\end{bmatrix}. \tag{11}$$

The method can be implemented in either a BLUP – residual maximum likelihood (REML) context, given $\mathbf{h}$ assessed previously by cross-validation or generalized cross-validation (Craven & Wahba, 1979), or in a fully Bayesian manner, with all parameters assigned prior distributions.

The predicted genetic or genomic value is then $\mathbf{K_h}\hat{\boldsymbol{\alpha}}$, as illustrated by González-Recio *et al.* (2008*a*) in an analysis of broiler mortality in which sires had been genotyped for thousands of SNPs, and by González-Recio *et al.* (2008*b*) in a similar study of food conversion ratio conducted in the same population. Their results suggest that RKHS regression using SNP information can produce more reliable prediction of current and future (offspring) phenotypes than the standard parametric additive model used in animal breeding. Also, albeit not unambiguously, some RKHS specifications with filtered SNPs tended to outperform parametric Bayesian regression models in which additive effects of all SNPs had been fitted.

In what follows, the RKHS machinery is used to develop non-parametric measures of breeding value.

(i) *Infinitesimal non-parametric breeding value*

Assume that some kernel matrix $\mathbf{K_h}$ has been found satisfactory, in the sense mentioned above. Given the context, vector $\mathbf{K_h}\boldsymbol{\alpha}$ in (10) is a molecularly marked counterpart of $\mathbf{g}=\mathbf{a}+\mathbf{d}+\mathbf{i}_{aa}+\mathbf{i}_{ad}+\mathbf{i}_{dd}$ in (1). Consider now the positive-definite kernel matrix $\mathbf{K}=\mathbf{A}+\mathbf{D}+(\mathbf{A}\#\mathbf{A})+(\mathbf{A}\#\mathbf{D})+(\mathbf{D}\#\mathbf{D})$. This is positive-definite, because it consists of the sum of positive-definite matrices, and leads to the RKHS model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+[\mathbf{A}+\mathbf{D}+(\mathbf{A}\#\mathbf{A})+(\mathbf{A}\#\mathbf{D})+(\mathbf{D}\#\mathbf{D})]\boldsymbol{\alpha}+\mathbf{e}, \tag{12}$$

with $\boldsymbol{\alpha}\sim\mathrm{N}(\mathbf{0},\mathbf{K}^{-1}\sigma^2_\alpha)$. Here, $\mathbf{K}$ does not depend on any bandwidth parameters, which simplifies matters greatly, relative to the parametric specification (7), in which 6 variance components enter into the problem. To obtain the RKHS predictor of genetic value (and with only 2 variance components intervening), note that solving (11) is equivalent to solving

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'K}\\ \mathbf{X} & \mathbf{K}+\dfrac{\sigma^2_e}{\sigma^2_\alpha}\mathbf{I}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\alpha}}\end{bmatrix}=\begin{bmatrix}\mathbf{X'y}\\ \mathbf{y}\end{bmatrix},$$

and the genetic value is predicted as

$$\begin{aligned}\hat{\mathbf{g}}&=[\mathbf{A}+\mathbf{D}+(\mathbf{A}\#\mathbf{A})+(\mathbf{A}\#\mathbf{D})+(\mathbf{D}\#\mathbf{D})]\hat{\boldsymbol{\alpha}}\\ &=\mathbf{A}\hat{\boldsymbol{\alpha}}+\mathbf{D}\hat{\boldsymbol{\alpha}}+(\mathbf{A}\#\mathbf{A})\hat{\boldsymbol{\alpha}}+(\mathbf{A}\#\mathbf{D})\hat{\boldsymbol{\alpha}}+(\mathbf{D}\#\mathbf{D})\hat{\boldsymbol{\alpha}}.\end{aligned} \tag{13}$$

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

Here, the 'non-parametric' breeding value would be $\mathbf{A}\boldsymbol{\alpha}$, and its predictor is $\mathbf{A}\hat{\boldsymbol{\alpha}}$. The posterior distribution of $\mathbf{A}\boldsymbol{\alpha}$ can be arrived at by drawing samples from the posterior distribution of $\boldsymbol{\alpha}$.

(ii) *SNP-based non-parametric breeding value*

Let now $\mathbf{K_h}$ be a kernel matrix whose entries depend on a string $\mathbf{x}$ of SNP genotypes or haplotypes available for a set of individuals having phenotypic records. For example, Gianola & van Kaam (2008) consider the Gaussian kernel

$$k_{\mathbf{h}}(\mathbf{x}_i,\mathbf{x}_j)=\exp\left[-\frac{(\mathbf{x}_i-\mathbf{x}_j)'(\mathbf{x}_i-\mathbf{x}_j)}{h}\right]. \tag{14}$$

Alternatively, one could use as kernel

$$k_{\mathbf{h}}(\mathbf{x}_i,\mathbf{x}_j)=\prod_{k=1}^{\text{number of chromosome pairs}}\exp\left[-\frac{(\mathbf{x}_{ik}-\mathbf{x}_{jk})'(\mathbf{x}_{ik}-\mathbf{x}_{jk})}{h_k}\right],$$

where $\mathbf{x}_{jk}$ is the SNP string for chromosome k in individual j, and $h_1, h_2, \ldots, h_k$ are chromosome-specific positive bandwidth parameters.

Irrespective of the kernel adopted, let now

$$\mathbf{K}\boldsymbol{\alpha}=\mathbf{a}_K+\boldsymbol{\gamma},$$

where $\mathbf{a}_K$ is defined as non-parametric breeding value, and $\boldsymbol{\gamma}$ is independent of $\mathbf{a}_K$. With $\mathbf{A}$ being the numerator relationship matrix between the individuals having SNP information, use of the reasoning leading to (4) produces

$$\begin{aligned}\mathbf{a}_K&=\sigma_a^2\mathbf{A}[\mathrm{Var}(\mathbf{K}\boldsymbol{\alpha})]^{-1}\mathbf{K}\boldsymbol{\alpha}\\&=\sigma_a^2\mathbf{A}[\mathbf{K}\mathbf{K}^{-1}\sigma_\alpha^2\mathbf{K}]^{-1}\mathbf{K}\boldsymbol{\alpha}=\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}\alpha,\end{aligned} \tag{15}$$

where σ_a^2 is some estimate of additive genetic variance for the trait in question. Thus, the non-parametric breeding value of individual i in the sample has the form

$$a_{K,i}=\frac{\sigma_a^2}{\sigma_\alpha^2}\sum_{j=1}^{n}a_{ij}\alpha_j,$$

so that, if individuals are unrelated, $a_{K,i}=\frac{\sigma_a^2}{\sigma_\alpha^2}\alpha_i$, with variance $V(a_{K,i})=\sigma_a^2\left(\frac{\sigma_a^2}{\sigma_\alpha^2}k^{ii}\right)$, where k^{ii} is the ith diagonal element of $\mathbf{K}^{-1}$.

Note that

$$\mathrm{Var}(\mathbf{a}_K)=\mathbf{A}\sigma_a^2\left(\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{K}^{-1}\mathbf{A}\right). \tag{16}$$

This is equal to $\mathbf{A}\sigma_a^2$ only if $\mathbf{K}=\mathbf{A}$ and $\sigma_\alpha^2=\sigma_a^2$; in this case, no molecular information is used at all. Suppose now that $\mathbf{A}=\mathbf{I}$, so that individuals are genetically unrelated. However, non-parametric breeding values turn out to be correlated, since

$$\mathrm{Var}(\mathbf{a}_K)=\sigma_a^2\left(\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{K}^{-1}\right). \tag{17}$$

This can be interpreted in the following manner, using a Gaussian kernel to illustrate. Here, the entries of the kernel matrix have the form

$$\begin{aligned}k_{\mathbf{h}}(\mathbf{x}_i,\mathbf{x}_j)&=\exp\left[-\frac{(\mathbf{x}_i-\mathbf{x}_j)'(\mathbf{x}_i-\mathbf{x}_j)}{h}\right]\\&=\prod_{k=1}^{\text{number of markers}}\exp\left[-\frac{(x_{ik}-x_{jk})^2}{h}\right],\end{aligned}$$

and $\exp\left[-\frac{(x_{ik}-x_{jk})^2}{h}\right]$ is maximum when $x_{ik}=x_{jk}$, that is, when individuals i and j have the same genotype at marker k. This means that the elements of the kernel matrix (taking values between 0 and 1) will be larger for pairs of individuals that are more 'molecularly alike', even if unrelated by line of descent. This molecular similarity is then propagated into (16) and (17).

Given a point estimate of $\boldsymbol{\alpha}(\hat{\boldsymbol{\alpha}})$, e.g. the posterior mean, the non-parametric estimate of breeding value is

$$\hat{\mathbf{a}}_K=\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}\hat{\boldsymbol{\alpha}}. \tag{18}$$

If computations are carried out in a fully Bayesian Monte Carlo context (with m samples drawn from the posterior distribution), the posterior mean estimate is

$$\hat{\mathbf{a}}_K=E\left(\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}\boldsymbol{\alpha}|\mathrm{DATA}\right)\approx\frac{\mathbf{A}\sigma_a^2}{m}\sum_{i=1}^{m}\frac{\boldsymbol{\alpha}^{(i)}}{\sigma_\alpha^{2(i)}},$$

and the uncertainty measure (akin to prediction error variance–covariance in BLUP) is

$$\begin{aligned}&\mathrm{Var}\left(\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}\boldsymbol{\alpha}|\mathrm{DATA}\right)\\&\quad\approx\sigma_a^4\mathbf{A}\left\{\frac{1}{m}\sum_{i=1}^{m}\frac{\boldsymbol{\alpha}^{(i)}\boldsymbol{\alpha}^{(i)\prime}}{\left(\sigma_\alpha^{2\,(i)}\right)^2}-\left(\frac{1}{m}\right)^2\left[\sum_{i=1}^{m}\frac{\boldsymbol{\alpha}^{(i)}}{\left(\sigma_\alpha^{2\,(i)}\right)}\right]\right.\\&\qquad\left.\times\left[\sum_{i=1}^{m}\frac{\boldsymbol{\alpha}^{(i)\prime}}{\left(\sigma_\alpha^{2\,(i)}\right)}\right]\right\}\mathbf{A}.\end{aligned}$$

Non-parametric breeding values of individuals that are not genotyped can be inferred as follows. Let $\mathbf{A}_{(-,+)}$ be the additive relationship matrix between individuals that are not genotyped (–) and those which are genotyped (+). The non-parametric breeding value of non-genotyped individuals would be

$$\mathbf{a}_{(-,K)}=\frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}_{(-,+)}\boldsymbol{\alpha}$$

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

with its point estimate being

$$\hat{\mathbf{a}}_{(-,K)} = \frac{\sigma_a^2}{\hat{\sigma}_\alpha^2}\mathbf{A}_{(-,+)}\hat{\boldsymbol{\alpha}}.$$

Likewise, non-parametric breeding values of yet-to-be genotyped and phenotyped progeny would be

$$\mathbf{a}_{(\text{prog},K)} = \frac{\sigma_a^2}{\sigma_\alpha^2}\mathbf{A}_{(\text{prog},K)}\boldsymbol{\alpha},$$

where $\mathbf{A}(_{\text{prog},+})$ is the additive relationship matrix between progenies and individuals with data.

(iii) *Semi-parametric breeding value*

Another possibility consists of treating additive genetic effects in a parametric manner and non-additive effects in the RKHS framework, as suggested by Gianola *et al.* (2006). Let now the model be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}, \qquad (19)$$

where $\mathbf{u} \sim (\mathbf{0}, \mathbf{A}\sigma_a^2)$ and $\mathbf{K}\boldsymbol{\alpha}$ as before, but with

$$\mathbf{K} = [\mathbf{D}\#\mathbf{K}_h + (\mathbf{A}\#\mathbf{A}\#\mathbf{K}_h) + (\mathbf{A}\#\mathbf{D}\#\mathbf{K}_h) + (\mathbf{D}\#\mathbf{D}\#\mathbf{K}_h)],$$

where $\mathbf{K}_h$ is a positive-definite matrix of Gaussian kernels with entries as in (14), so that SNP information is used. Since $\mathbf{K}$ is the sum of Hadamard products of positive-definite matrices, it is positive-definite as well and, hence, it is a valid kernel for RKHS regression. The estimating equations (can be rendered symmetric by premultiplying the $\boldsymbol{\alpha}$-set of equations by $\mathbf{K}$) take now the form

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}' & \mathbf{X}'\mathbf{K} \\ \mathbf{X} & \mathbf{I} + \dfrac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{A}^{-1} & \mathbf{K} \\ \mathbf{X} & \mathbf{I} & \mathbf{K} + \dfrac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \end{bmatrix},$$

where $\hat{\mathbf{u}}$ is the predicted breeding value, and

$$\mathbf{K}\hat{\boldsymbol{\alpha}} = (\mathbf{D}\#\mathbf{K}_h)\hat{\boldsymbol{\alpha}} + (\mathbf{A}\#\mathbf{A}\#\mathbf{K}_h)\hat{\boldsymbol{\alpha}} + (\mathbf{A}\#\mathbf{D}\#\mathbf{K}_h)\hat{\boldsymbol{\alpha}} + (\mathbf{D}\#\mathbf{D}\#K_h)\hat{\boldsymbol{\alpha}}$$

is the predicted (SNP-based) non-additive genetic value. For instance, $(\mathbf{D}\#\mathbf{K}_h)\hat{\boldsymbol{\alpha}}$ is interpretable as a predicted dominance value, and so on. A natural extension of this consists of removing $\mathbf{u}$ from (19) and then taking as kernel matrix

$$\mathbf{K} = [\mathbf{A} + \mathbf{D} + (\mathbf{A}\#\mathbf{A}) + (\mathbf{A}\#D) + (\mathbf{D}\#\mathbf{D})]\#\mathbf{K}_h,$$

so that the predicted additive breeding value would now be $(\mathbf{A}\#\mathbf{K}_h)\hat{\boldsymbol{\alpha}}$, much along the lines of (18). This would give a completely non-parametric treatment of prediction of additive and non-additive genetic effects using dense molecular information.

5. Illustration: infinitesimal model

The toy example in Henderson (1985) is considered. The problem is to infer additive and dominance genetic effects of five individuals using data consisting of phenotypic records for only four (subject 1 lacks a record). The data and model are:

$$\begin{bmatrix} y_2 = 5 \\ y_3 = 3 \\ y_4 = 7 \\ y_5 = 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\times \left(\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} + \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix} \right) + \begin{bmatrix} e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e}.$$

Although 1 does not have a record, all additive and dominance effects are included in the model, zeroing out the incidence of a_1 and d_1 via a column of 0s in the $\mathbf{Z}$ incidence matrix. Henderson (1985) assumed that the additive, dominance and residual variance components were $\sigma_a^2 = 5$, $\sigma_d^2 = 4$ and $\sigma_e^2 = 20$, so that the phenotypic variance was 29, and used the additive and dominance relationship matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Application of BLUP leads to the unique solutions (rounded at the third decimal)

$$\hat{\boldsymbol{\beta}}' = [5{\cdot}145 \quad 0{\cdot}241],$$

$$\hat{\mathbf{a}}' = [0{\cdot}045 \quad -0{\cdot}192 \quad -0{\cdot}343 \quad 0{\cdot}096 \quad 0{\cdot}242],$$

$$\hat{\mathbf{d}}' = [0 \quad -0{\cdot}073 \quad -0{\cdot}365 \quad 0{\cdot}162 \quad 0{\cdot}234].$$

The predicted total genetic value is $\hat{\mathbf{g}} = \hat{\mathbf{a}} + \hat{\mathbf{d}}$, yielding

$$\hat{\mathbf{g}} = [0{\cdot}045 \quad -0{\cdot}265 \quad -0{\cdot}708 \quad 0{\cdot}259 \quad 0{\cdot}477].$$

Alternatively, a RKHS representation as in (12) is used now, with $\mathbf{K} = \mathbf{A} + \mathbf{D}$ as kernel matrix, which does not involve any bandwidth parameters. One has

$$\mathbf{K} = \begin{bmatrix} 2 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix}.$$

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

The RKHS model (using the part of $\mathbf{K}$ pertaining to individual with records) is

$$\begin{bmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\alpha} + \mathrm{e}.$$

The value adopted for the sole smoothing parameter is $\sigma_\alpha^2 = \sigma_a^2 + \sigma_d^2 = 9$, and the estimating equation (11) and the solution to the RKHS mixed model equations is

$$[\hat{\beta}_0 = 5{\cdot}289 \quad \hat{\beta}_1 = 0{\cdot}200 \quad \hat{\alpha}_2 = -0{\cdot}128$$

$$\hat{\alpha}_3 = -0{\cdot}781 \quad \hat{\alpha}_4 = 0{\cdot}487 \quad \hat{\alpha}_5 = 0{\cdot}422].$$

The non-parametric additive genetic effects are inferred using formula (18)

$$\begin{bmatrix} \hat{a}_{K,1} \\ \hat{a}_{K,2} \\ \hat{a}_{K,3} \\ \hat{a}_{K,4} \\ \hat{a}_{K,5} \end{bmatrix} = \frac{\sigma_a^2}{\sigma_\alpha^2} \mathbf{A}_{all,+} \hat{\boldsymbol{\alpha}} = \frac{5}{9} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} -0{\cdot}128 \\ -0{\cdot}781 \\ 0{\cdot}487 \\ 0{\cdot}422 \end{bmatrix}$$

$$= \begin{bmatrix} 0{\cdot}036 \\ -0{\cdot}153 \\ -0{\cdot}276 \\ 0{\cdot}076 \\ 0{\cdot}194 \end{bmatrix},$$

and the non-parametric estimates of dominance effects are

$$\begin{bmatrix} \hat{d}_{K,1} \\ \hat{d}_{K,2} \\ \hat{d}_{K,3} \\ \hat{d}_{K,4} \\ \hat{d}_{K,5} \end{bmatrix} = \frac{\sigma_d^2}{\sigma_\alpha^2} \mathbf{D}_{all,+} \hat{\boldsymbol{\alpha}} = \frac{4}{9} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -0{\cdot}128 \\ -0{\cdot}781 \\ 0{\cdot}487 \\ 0{\cdot}422 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ -0{\cdot}057 \\ -0{\cdot}293 \\ 0{\cdot}130 \\ 0{\cdot}188 \end{bmatrix}.$$

The total genetic value is estimated at

$$\begin{bmatrix} \hat{g}_{K,1} \\ \hat{g}_{K,2} \\ \hat{g}_{K,3} \\ \hat{g}_{K,4} \\ \hat{g}_{K,5} \end{bmatrix} = \begin{bmatrix} \hat{a}_{K,1} \\ \hat{a}_{K,2} \\ \hat{a}_{K,3} \\ \hat{a}_{K,4} \\ \hat{a}_{K,5} \end{bmatrix} + \begin{bmatrix} \hat{d}_{K,1} \\ \hat{d}_{K,2} \\ \hat{d}_{K,3} \\ \hat{d}_{K,4} \\ \hat{d}_{K,5} \end{bmatrix} = \begin{bmatrix} 0{\cdot}036 \\ -0{\cdot}210 \\ -0{\cdot}569 \\ 0{\cdot}206 \\ 0{\cdot}382 \end{bmatrix}.$$

Suppose now that one wishes to predict future performance of these five individuals, and assume that the record of 1 will be made under the same conditions as those for the record of 2, with all conditions remaining the same. The model for future records f is then

$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} + \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix} + \begin{bmatrix} e_1^f \\ e_2^f \\ e_3^f \\ e_4^f \\ e_5^f \end{bmatrix}$$

$$= \mathbf{M}_P \boldsymbol{\theta}_P + \mathbf{e}^f,$$

for the parametric model, whereas that for the RKHS treatment is

$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} + \begin{bmatrix} e_1^f \\ e_2^f \\ e_3^f \\ e_4^f \\ e_5^f \end{bmatrix}$$

$$= \mathbf{M}_K \boldsymbol{\theta}_K + \mathbf{e}^f.$$

Above $\mathbf{M}_{\cdot}$ and $\boldsymbol{\theta}_{\cdot}$ denote incidence matrices and regression coefficients, respectively; P and K indicate parametric and RKHS treatments, respectively. Using a standard Bayesian argument (Sorensen & Gianola, 2002), the mean vector and covariance matrix of the predictive distributions are

$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} \left| \begin{bmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \right., \text{ dispersion (smoothing) parameters}$$

$$\sim (\mathbf{M}_{\cdot}\hat{\boldsymbol{\theta}}_{\cdot}, (\mathrm{M}_{\cdot}\mathrm{C}_{\cdot}^{-1}\mathrm{M}_{\cdot}' + \mathbf{I}_f)\sigma_e^2), \qquad (20)$$

where $\mathbf{C}$ is the corresponding coefficient matrix for each of the procedures; the 5×5 matrix $\mathbf{I}_f \sigma_e^2$ conveys uncertainty stemming from the fact that the future records (f) have not been realized yet. For the two procedures, means and standard deviations of the

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

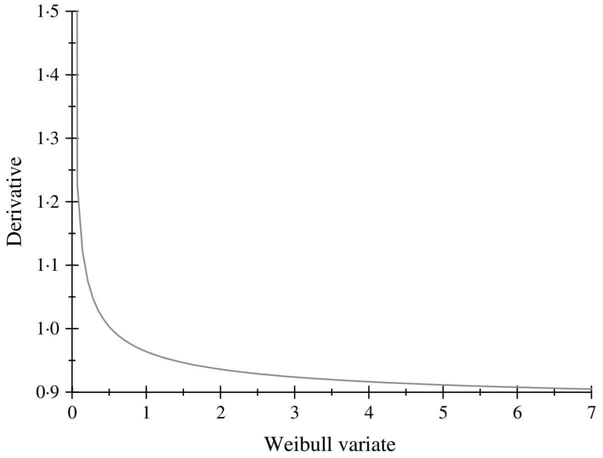

Fig. 1. Rate of change of the expected phenotypic value $\left(\frac{\partial E(.)}{\partial \beta_j}=\beta_i+\frac{1}{2}\alpha_i\alpha_j\sqrt{\frac{\beta_i}{\beta_j}}\right)$ with respect to the Weibull variable β_j.

predictive distributions are

$$P=\begin{bmatrix}5{\cdot}674\pm6{\cdot}020\\5{\cdot}364\pm5{\cdot}460\\5{\cdot}162\pm5{\cdot}353\\5{\cdot}646\pm5{\cdot}834\\6{\cdot}828\pm6{\cdot}115\end{bmatrix};\quad K=\begin{bmatrix}5{\cdot}754\pm5{\cdot}576\\5{\cdot}286\pm5{\cdot}659\\4{\cdot}735\pm5{\cdot}561\\5{\cdot}919\pm5{\cdot}940\\7{\cdot}061\pm6{\cdot}157\end{bmatrix}.$$

6. Illustration: non-linear system

(i) *Two-locus model*

A hypothetical system with two biallelic loci was simulated. It was assumed that phenotypes were generated according to the rule

$$E(y|\alpha_i,\alpha_j,\beta_i,\beta_j)=\alpha_i+\alpha_j+\beta_i\beta_j+\alpha_i\alpha_j\sqrt{\beta_i\beta_j},\qquad(21)$$

where α_i (β_i) and α_j (β_j) are effects of alleles i and j at the α (β) locus. The system is nonlinear on allelic effects, as indicated by the first derivatives of the conditional expectation function with respect to the αs or βs. For instance,

$$\frac{\partial E(.)}{\partial \alpha_j}=1+\alpha_i\sqrt{\beta_i\beta_j};\qquad \frac{\partial E(.)}{\partial \beta_j}=\beta_i+\frac{1}{2}\alpha_i\alpha_j\sqrt{\frac{\beta_i}{\beta_j}}.$$

The α-effects were two random draws from an exponential distribution with mean value equal to 2; the first draw was assigned to allele A and the second to allele a. The βs were drawn from a Weibull (2,1) distribution, having median, mean and mode equal to 0·347, 0·887 and 0·25; the first (second) deviate was the effect of allele B (b). The non-linearity of the system is illustrated in Fig. 1, where the derivative of the model for the expected phenotype with respect to the effect of Weibull allele b is plotted, with all other alleles evaluated at the values drawn. Variation in values of b produces drastic modifications near the origin, but phenotypes are essentially insensitive for $b>1{\cdot}5$, even though these values are plausible in the Weibull process hypothesized.

Residuals were drawn from the normal distribution N (0, 20), and added to (21) to form phenotypes. The resulting phenotypic distribution is unknown, because y is a non-linear function of exponential and Weibull variates, plus an additive normally distributed residual. There were 5 individuals with records for each of the $AABB$, $AABb$, $AAbb$ genotypes; 20 for each of $AaBB$, $AaBb$ and $Aabb$ and 5 for each of $aaBB$, $aaBb$ and $aabb$. Thus, there were 90 individuals with phenotypic records, in total.

Since there are nine distinct mean genotypic values, variation among their average values can be explained completely with a linear model on 9 *df*; in the standard treatment, these *df* correspond to an overall mean, additive effects (2 *df*), dominance (2 *df*) and epistasis (4 *df*). Due to non-linearity, it is not straightforward to assess the proportion of variance ‘due to genetic effects’ using a random effects model based on the Weibull and exponential distributions, although an approximation can be arrived at. A linear approximation of the phenotype of an individual yields

$$\begin{aligned}y\approx{}&2\bar{\alpha}+\bar{\beta}^2+\bar{\alpha}^2\bar{\beta}+(2\bar{\alpha}\bar{\beta}+2)(\alpha-\bar{\alpha})\\&+(\bar{\alpha}^2+2\bar{\beta})(\beta-\bar{\beta})+e\end{aligned},$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the means of the exponential and Weibull processes, respectively, and $e\sim N(0,\sigma^2)$. An approximation to heritability is

$$h^2=\frac{(2\bar{\alpha}\bar{\beta}+2)^2\mathrm{Var}(\alpha)+(\bar{\alpha}^2+2\bar{\beta})^2\mathrm{Var}(\beta)}{(2\bar{\alpha}\bar{\beta}+2)^2\mathrm{Var}(\alpha)+(\bar{\alpha}^2+2\bar{\beta})^2\mathrm{Var}(\beta)+\sigma^2}.\qquad(22)$$

For the situation simulated, $\bar{\alpha}=2$, $\bar{\beta}=0{\cdot}887$, $\mathrm{Var}(\alpha)=4$, $\mathrm{Var}(\beta)=0{\cdot}785$, so that for $\sigma^2=1000$, 100, 20, 10 one obtains $h^2\approx0{\cdot}13$, 0·60, 0·88 and 0·94, respectively. The simulation produced a high penetrance trait, that is, one with heritability near 0·88, which provides a challenge to any non-parametric treatment.

(ii) *RKHS modelling*

Among the many possible candidate kernels available, an arbitrarily chosen Gaussian kernel was adopted for the RKHS regression implementations, using as covariate a 2×1 vector consisting of the number of alleles at each of the two loci, e.g. $x_{AA}=2$, $x_{Aa}=1$ and $x_{aa}=0$. For example, the kernel entry for

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

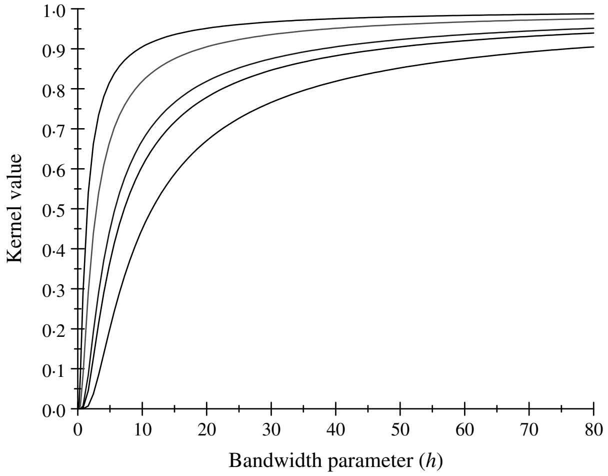

Fig. 2. Kernel value $k(.,.;h)=\exp\left(-\frac{S}{h}\right)$ against the bandwidth parameter h. Curves, from top to bottom, correspond to $S=1$, 2, 4, 5, 8.

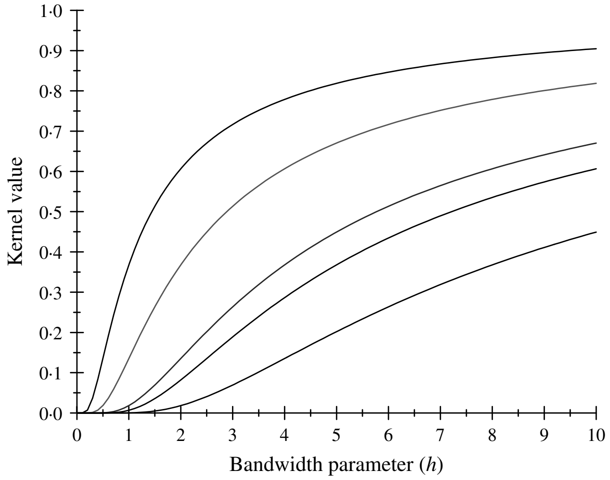

Fig. 3. Kernel value $k(.,.;h)=\exp\left(-\frac{S}{h}\right)$ against the bandwidth parameter h. Curves, from top to bottom, correspond to $S=1$, 2, 4, 5, 8.

genotypes *AABB* and *AAbb* is

$$\begin{aligned} k(\mathbf{x}_{AABB},\mathbf{x}_{AAbb},h) &= \exp\left[-\frac{(2-2)^2+(2-0)^2}{h}\right] \\ &= \exp\left[-\frac{4}{h}\right], \end{aligned}$$

and the 9×9 kernel matrix for all possible genotypes (labels for genotypes are included, to facilitate understanding of entries) is

$$\mathbf{K}_h=\left[\begin{array}{cccccccccc} & AABB & AABb & AAbb & AaBB & AaBb & Aabb & aaBB & aaBb & aabb \\ AABB & 1 & e^{-1/h} & e^{-4/h} & e^{-1/h} & e^{-2/h} & e^{-5/h} & e^{-4/h} & e^{-5/h} & e^{-8/h} \\ AABb & e^{-1/h} & 1 & e^{-1/h} & e^{-2/h} & e^{-1/h} & e^{-2/h} & e^{-5/h} & e^{-4/h} & e^{-5/h} \\ AAbb & e^{-4/h} & e^{-1/h} & 1 & e^{-5/h} & e^{-2/h} & e^{-1/h} & e^{-8/h} & e^{-5/h} & e^{-4/h} \\ AaBB & e^{-1/h} & e^{-2/h} & e^{-5/h} & 1 & e^{-1/h} & e^{-4/h} & e^{-1/h} & e^{-2/h} & e^{-5/h} \\ AaBb & e^{-2/h} & e^{-1/h} & e^{-2/h} & e^{-1/h} & 1 & e^{-1/h} & e^{-2/h} & e^{-1/h} & e^{-2/h} \\ Aabb & e^{-5/h} & e^{-2/h} & e^{-1/h} & e^{-4/h} & e^{-1/h} & 1 & e^{-5/h} & e^{-2/h} & e^{-1/h} \\ aaBB & e^{-4/h} & e^{-5/h} & e^{-8/h} & e^{-1/h} & e^{-2/h} & e^{-5/h} & 1 & e^{-1/h} & e^{-4/h} \\ aaBb & e^{-5/h} & e^{-4/h} & e^{-5/h} & e^{-2/h} & e^{-1/h} & e^{-2/h} & e^{-1/h} & 1 & e^{-1/h} \\ aabb & e^{-8/h} & e^{-5/h} & e^{-4/h} & e^{-5/h} & e^{-2/h} & e^{-1/h} & e^{-4/h} & e^{-1/h} & 1 \end{array}\right].$$

There are only five distinct entries, a result of the measure of 'allelic disimilarity' adopted, e.g. $e^{-8/h}$ stems from disimilarity in 2 alleles at each of the *A* and *B* loci (*AABB* versus *aabb*). Figures 2 and 3 depict values of the kernel as a function of the bandwidth parameter h at different levels of S, which enters into $\exp(-S/h)$. Values of h larger than 10 (Fig. 2) produce strong 'prior correlations' between genotypes; also, the kernel matrix becomes more poorly conditioned as h increases. After evaluating the kernel matrix at $h=6$, 4, 2 and 1·75, it was decided to adopt $h=1{\cdot}75$ as the bandwidth parameter, producing 6 unique entries in the **K** matrix: 1·0 (diagonal elements, the two individuals have identical genotypes); 0·565 (3 alleles in common in a pair of individuals); 0·319 (2 alleles in common, 1 per locus) or 0·102 (2 alleles in common at only one locus); 0·06 (1 allele in common) and 0·01 (no alleles shared).

(iii) *Fitting the RKHS model to the means*

A 'means' RKHS regression was fitted, including an intercept (β). The model for the 9×1 vector of averages $\bar{\mathbf{y}}=\{\bar{y}_i\}$ was

$$\begin{bmatrix} \bar{y}_{AABB} \\ \bar{y}_{AABb} \\ \bar{y}_{AAbb} \\ \bar{y}_{AaBB} \\ \bar{y}_{AaBb} \\ \bar{y}_{Aabb} \\ \bar{y}_{aaBB} \\ \bar{y}_{aaBb} \\ \bar{y}_{aabb} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\beta+\mathbf{K}_{1\cdot 75}\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{bmatrix}+\begin{bmatrix} \bar{e}_{AABB} \\ \bar{e}_{AABb} \\ \bar{e}_{AAbb} \\ \bar{e}_{AaBB} \\ \bar{e}_{AaBb} \\ \bar{e}_{Aabb} \\ \bar{e}_{aaBB} \\ \bar{e}_{aaBb} \\ \bar{e}_{aabb} \end{bmatrix},$$

where the αs are the non-parametric regression coefficients; $\mathbf{K}_{1\cdot 75}$ is the 9×9 kernel matrix with $h=1{\cdot}75$

and $\bar{e}_i$ is the mean of the residuals pertaining to observations of individuals with genotype i. As before, the assumptions were $\boldsymbol{\alpha}|\sigma^2_\alpha \sim N(\mathbf{0}, \mathbf{K}^{-1}_{1\cdot75}\sigma^2_\alpha)$, and, for the 9×1 vector of average residuals, $\bar{\mathbf{e}}|\sigma^2_e \sim N(\mathbf{0}, \mathbf{N}^{-1}\sigma^2_e)$. Here, $\mathbf{N}=\text{Diag}\ \{5, 5, 5, 20, 20, 20, 5, 5, 5\}$ is a 9×9 diagonal matrix.

The solution to the RKHS regression problem was obtained as

$$\begin{bmatrix}\hat{\beta}\\ \hat{\boldsymbol{\alpha}}\end{bmatrix}=\begin{bmatrix}\mathbf{1}'\mathbf{N}\mathbf{1} & \mathbf{1}'\mathbf{N}\mathbf{K}_{1\cdot75}\\ \mathbf{K}'_{1\cdot75}\mathbf{N}\mathbf{1} & \mathbf{K}'_{1\cdot75}\mathbf{N}\mathbf{K}_{1\cdot75}+\dfrac{\sigma^2_e}{\sigma^2_\alpha}\mathbf{K}_{1\cdot75}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{1}'\mathbf{N}\bar{\mathbf{y}}\\ \mathbf{K}'_{1\cdot75}\mathbf{N}\bar{\mathbf{y}}\end{bmatrix}.$$

The model was fitted for each of the following values of the shrinkage ratio $\lambda=\frac{\sigma^2_e}{\sigma^2_\alpha}$: 100 (strong shrinkage towards 0), 15, 1, $\frac{1}{15}$ and $\frac{1}{100}$; with $\lambda=0$ there is no shrinkage of solutions at all. For each of these values, the residual sum of squares

$$\text{SSR}_\lambda=\left(\bar{\mathbf{y}}-\mathbf{1}\hat{\beta}-\mathbf{K}_{1\cdot75}\hat{\boldsymbol{\alpha}}\right)'\left(\bar{\mathbf{y}}-\mathbf{1}\hat{\beta}-\mathbf{K}_{1\cdot75}\hat{\boldsymbol{\alpha}}\right),$$

and the weighted residual sum of squares

$$\text{WSSR}_\lambda=\left(\bar{\mathbf{y}}-\mathbf{1}\hat{\beta}-\mathbf{K}_{1\cdot75}\hat{\boldsymbol{\alpha}}\right)'\mathbf{N}\left(\bar{\mathbf{y}}-\mathbf{1}\hat{\beta}-\mathbf{K}_{1\cdot75}\hat{\boldsymbol{\alpha}}\right),$$

were computed, where $\mathbf{1}$ is a 9×1 vector of ones. Also, the effective number of parameters, or model df (Ruppert *et al.*, 2003), was assessed as

$$df_\lambda=\text{Tr}(\mathbf{W}^*\mathbf{C}^{-1}\mathbf{W}^{*\prime}),$$

where

$$\mathbf{W}^*=\mathbf{N}^{1/2}[\mathbf{1}\quad \mathbf{K}_{1\cdot75}],$$

and $\mathbf{N}^{1/2}$ is a diagonal matrix containing the square roots of the entries of $\mathbf{N}$. Note that, when $\frac{\sigma^2_e}{\sigma^2_\alpha}=0$, the estimating equations have an infinite number of solutions, as only 9 parameters are estimable. In this case, a solution is obtained by using a generalized inverse of the coefficient matrix, yielding

$$\begin{bmatrix}\hat{\beta}^0\\ \hat{\boldsymbol{\alpha}}^0\end{bmatrix}=\begin{bmatrix}0 & \mathbf{0}\\ \mathbf{0} & \mathbf{K}'_{1\cdot75}\mathbf{N}\mathbf{K}_{1\cdot75}\end{bmatrix}^{-}\begin{bmatrix}\mathbf{1}'\mathbf{N}\bar{\mathbf{y}}\\ \mathbf{K}'_{1\cdot75}\mathbf{N}\bar{\mathbf{y}}\end{bmatrix}$$
$$=\begin{bmatrix}0 & \mathbf{0}\\ \mathbf{0} & (\mathbf{K}'_{1\cdot75}\mathbf{N}\mathbf{K}_{1\cdot75})^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{1}'\mathbf{N}\bar{\mathbf{y}}\\ \mathbf{K}'_{1\cdot75}\mathbf{N}\bar{\mathbf{y}}\end{bmatrix}=\begin{bmatrix}0\\ \mathbf{K}^{-1}_{1\cdot75}\bar{\mathbf{y}}\end{bmatrix}.$$

Here, the fitted values are $\hat{\bar{\mathbf{y}}}=\mathbf{1}\times0+\mathbf{K}_{1\cdot75}\mathbf{K}^{-1}_{1\cdot75}\bar{\mathbf{y}}=\bar{\mathbf{y}}$ so the model 'copies' the data, and the fit is perfect. Note that $\bar{y}_i$ is the least-squares estimate of $\mu_{ij}=\gamma_i+\delta_j+(\gamma\delta)_{ij}$ where γ_i ($i=AA$, Aa, aa) and δ_j ($j=BB$, Bb, bb) are main effects of genotypes at loci A and B, respectively; this model accounts for 8 df 'due to' additive and dominance effects at each of the two loci, and additive × additive, additive × dominance, dominance × additive and dominance × dominance interactions.

For the sake of comparison, the following fixed effects model was fitted as well

$$y=\mu+\gamma_i+\delta_j+\bar{e}_i,$$

where, as before, γ_i ($i=AA$, Aa, aa) are main effects of genotypes at the A-*locus*, and δ_j ($j=BB$, Bb, bb) are the counterparts at the locus B. The linear model was parameterized as

$$\begin{bmatrix}\bar{y}_{AABB}\\ \bar{y}_{AABb}\\ \bar{y}_{AAbb}\\ \bar{y}_{AaBB}\\ \bar{y}_{AaBb}\\ \bar{y}_{Aabb}\\ \bar{y}_{aaBB}\\ \bar{y}_{aaBb}\\ \bar{y}_{aabb}\end{bmatrix}=\begin{bmatrix}1&1&0&0&1&0&0\\ 1&1&0&0&0&1&0\\ 1&1&0&0&0&0&1\\ 1&0&1&0&1&0&0\\ 1&0&1&0&0&1&0\\ 1&0&1&0&0&0&1\\ 1&0&0&1&1&0&0\\ 1&0&0&1&0&1&0\\ 1&0&0&1&0&0&1\end{bmatrix}\begin{bmatrix}\mu\\ \gamma_1\\ \gamma_2\\ \gamma_3\\ \delta_1\\ \delta_2\\ \delta_3\end{bmatrix}+\begin{bmatrix}\bar{e}_{AABB}\\ \bar{e}_{AABb}\\ \bar{e}_{AAbb}\\ \bar{e}_{AaBB}\\ \bar{e}_{AaBb}\\ \bar{e}_{Aabb}\\ \bar{e}_{aaBB}\\ \bar{e}_{aaBb}\\ \bar{e}_{aabb}\end{bmatrix}. \qquad (23)$$

This model has 5 free parameters, interpretable as an 'overall mean' plus additive and dominance effects at each of the 2 loci. Estimates of estimable functions of fixed effects were obtained from a weighted least-squares approach with solution vector $\mathbf{b}^0=(\mathbf{X}'\mathbf{N}\mathbf{X})^-\times\mathbf{X}'\mathbf{N}\bar{\mathbf{y}}$, where $(\mathbf{X}'\mathbf{N}\mathbf{X})^-$ is a generalized inverse of $\mathbf{X}'\mathbf{N}\mathbf{X}$ and $\mathbf{X}$ is the 9×7 incidence matrix given above. Weighted residual sums of squares were computed as $(\bar{\mathbf{y}}-\mathbf{X}\mathbf{b}^0)'\mathbf{N}(\bar{\mathbf{y}}-\mathbf{X}\mathbf{b}^0)$. Mean values of A-locus and B-locus genotypes were estimated as

$$\begin{bmatrix}\hat{\mu}_{AA}\\ \hat{\mu}_{Aa}\\ \hat{\mu}_{aa}\\ \hat{\mu}_{BB}\\ \hat{\mu}_{Bb}\\ \hat{\mu}_{bb}\end{bmatrix}=\begin{bmatrix}1&1&0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ 1&0&1&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ 1&0&0&1&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ 1&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&1&0&0\\ 1&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0&1&0\\ 1&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0&0&1\end{bmatrix}\mathbf{b}^0=\begin{bmatrix}8\cdot37\\ 3\cdot97\\ 1\cdot39\\ 4\cdot87\\ 5\cdot73\\ 3\cdot12\end{bmatrix},$$

and dominance effects were inferred as

$$\begin{bmatrix}\hat{d}_A\text{-locus}\\ \hat{d}_B\text{-locus}\end{bmatrix}=\begin{bmatrix}0&-\frac{1}{2}&1&-\frac{1}{2}&0&0&0\\ 0&0&0&0&-\frac{1}{2}&1&-\frac{1}{2}\end{bmatrix}\mathbf{b}^0$$
$$=\begin{bmatrix}-0\cdot91\\ 1\cdot73\end{bmatrix}.$$

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

Table 1. *Estimates of the intercept (β) and of non-parametric regression coefficients (α_i) for each of the values of the variance ratio ($\lambda = \sigma_e^2/\sigma_\alpha^2$) employed. RSS and WRSS are the residual and weighted residual sums of squares, respectively; df gives the model df, or effective number of parameters fitted*

Item	$\lambda = 10^2$	$\lambda = 15$	$\lambda = 1$	$\lambda = 15^{-1}$	$\lambda = 10^{-2}$	$\lambda = 0$
β	4·26	4·25	4·30	4·22	4·21	0
α_1	0·27	1·27	2·24	−1·50	−2·34	0·94
α_2	0·31	1·589	7·501	14·543	15·808	16·219
α_3	$1{\cdot}4 \times 10^{-3}$	$-6{\cdot}5 \times 10^{-2}$	−0·90	0·35	0·77	4·78
α_4	$6{\cdot}2 \times 10^{-2}$	0·11	1·56	7·49	8·77	9·28
α_5	$2{\cdot}2 \times 10^{-2}$	−0·21	−4·53	−13·16	−14·90	−16·00
α_6	−0·27	−1·01	−3·72	−9·15	−10·31	−11·22
α_7	−0·21	−1·07	−4·51	−9·00	−9·88	−6·61
α_8	−0·10	−0·35	1·96	7·70	8·84	9·22
α_9	$-8{\cdot}2 \times 10^{-2}$	−0·27	0·41	2·73	3·25	7·29
RSS	91·83	49·80	3·55	0·07	$1{\cdot}9 \times 10^{-3}$	0
WRSS	487·18	258·01	19·12	0·39	$1{\cdot}1 \times 10^{-2}$	0
df	1·46	2·96	6·75	8·71	8·95	9

This analysis would create the illusion of non-additivity and overdominance at the *A* and *B* loci, respectively, but without bringing light with respect to the non-linearities of (21). While this model does not have mechanistic relevance, it has predictive value, an issue which is illustrated below.

Table 1 gives estimates of the intercept β and of the RKHS regression coefficients α for each of the shrinkage ratio values employed, using $h = 1{\cdot}75$ as bandwidth parameter in all cases. The residual sum of squares (weighted and unweighted) and the effective number of parameters, or model *df*, are presented as well. As λ decreased from 10^2 to 10^{-2}, model fit improved, but the efective number of parameters increased from 1·46 to 8·95, near the maximum of 9. The implementations with $\lambda = 15^{-1}$ and $\lambda = 10^{-2}$ essentially produced a 'saturated' model, that is, one that fits to the means perfectly (as it is the case when $\lambda = 0$). Large values of the variance ratio (15, 100) produced excessive shrinkage, as indicated by the small values of the RKHS regression coefficients α, and small values of the variance ratio tended to overfit, as noted. On the other hand, the linear 2-locus model on additive effects, with 5 parameters, had residual sum of squares, SSR = 19·98, and weighted residual sum of squares, SSR = 99·88. These values are more than 5 times those attained with the RKHS regression implementation with a variance ratio of 1 (SSR = 3·55, WSSR = 19·12), which are shown in Table 1.

A more important issue, at least from the perspective taken in this paper, is 'out of sample' predictive ability. To examine this, 3 new (independent) samples of phenotypes were generated, assuming the residual distribution *N* (0, 20) , as before, and with 5 individuals per genotype, i.e. there were 45 subjects in each sample. The predictive residual sums of squares (unweighted and weighted, using 5 as weight) were calculated, using the fitted values from the training sample employed to compute statistics in Table 1, and the 'new sample' phenotypes. The 2-locus additive model was also involved in the comparison. Results are shown in Table 2, where entries are the average, minimum and maximum values of the predictive sum of squares over the 3 new samples. As expected, predictive residual sum of squares were much larger than those observed in the training sample, notably for implementations in which overfitting to the training data was obvious ($\lambda = \frac{\sigma_e^2}{\sigma_\alpha^2} = 15^{-1}, 10^{-2}, 0$; see Table 1). The specifications producing strong shrinkage towards 0 in the training sample had the worse predictive performance, whereas that with $\frac{\sigma_e^2}{\sigma_e^2} = 1$ had the best performance, on average, albeit close to the 2-locus additive model. It is not surprising that small variance ratios led to reasonable predictive performance, because the simulation mimicked high penetrance,

Table 2. *Predictive (weighted and unweighted by the number of individuals per geno-type) residual sums of squares for each of the variance ratios ($\lambda = \sigma_e^2/\sigma_\alpha^2$) employed in the non-parametric regression implementation, and for the two-locus model with main effects of genotypes at each of the loci. Entries are average (boldface) from three predictive samples, with minimum and maximum values over samples in parentheses*

Item	Sum of squares	Weighted sum of squares
$\lambda = 10^2$	**136·6** (96·6, 198·9)	**682·8** (483·1, 994·3)
$\lambda = 15$	**92·5** (61·8, 136·6)	**462·4** (308·9, 682·9)
$\lambda = 1$	**51·3** (39·3, 58·51)	**256·7** (196·7, 292·6)
$\lambda = 15^{-1}$	**58·8** (54·2, 66·6)	**293·9** (271·1, 333·2)
$\lambda = 10^{-2}$	**60·6** (56·1, 68·8)	**303·0** (280·4, 343·8)
$\lambda = 0$	**61·0** (56·2, 69·2)	**304·9** (281·0, 346·0)
Two-locus additive model	**53·9** (48·8, 62·2)	**269·7** (244·1, 311·0)

i.e. phenotypes are very informative about genotypes, this being so because the approximate heritability (22) for a residual variance of 20 is about 0·88, as already noted.

For this example, the 2-locus additive model is untenable, at least mechanistically, yet it had a reasonable predictive performance. Apart from genetic considerations discussed later, this is because any of the models considered here can be associated with a linear smoother, with predictions having the form

$$\hat{\mathbf{y}} = \mathbf{L}\mathbf{y} = \{\mathbf{l}_i'\mathbf{y}\},$$

for some matrix $\mathbf{L}$ and where $\mathbf{l}_i'$ is its ith row. For the 2-locus model,

$$\mathbf{L} = \mathbf{X}(\mathbf{X}'\mathbf{N}\mathbf{X})^{-}\mathbf{X}'\mathbf{N},$$

whereas for any of the RKHS specifications

$$\mathbf{L} = [\mathbf{1} \quad \mathbf{K}_{1\cdot75}] \begin{bmatrix} \mathbf{1}'\mathbf{N}\mathbf{1} & \mathbf{1}'\mathbf{N}\mathbf{K}_{1\cdot75} \\ \mathbf{K}_{1\cdot75}'\mathbf{N}\mathbf{1} & \mathbf{K}_{1\cdot75}'\mathbf{N}\mathbf{K}_{1\cdot75} + \frac{\sigma_e^2}{\sigma_\alpha^2}\mathbf{K}_{1\cdot75} \end{bmatrix}^{-1} \times [\mathbf{1} \quad \mathbf{K}_{1\cdot75}]'.$$

Hence, all predictors can be viewed as consisting of different forms of averaging observations, with the RKHS-based averages having optimality properties, in some well defined sense (Kimeldorf & Wahba, 1971; Wahba, 1990; Gianola & van Kaam, 2008; de los Campos *et al*., 2008). Since the data consisted of phenotypic averages, i.e. a non-parametric averaging method where a 'bin' is a given genotype, this represented a challenge for any of the smoothers, including the 2-locus additive model. However, some of the smoothers (e.g. RKHS with $\lambda = 1$ and the 2-locus additive model) met the challenge satisfactorily.

(iv) *Fitting the RKHS model to individual observations*

The RKHS regression model (using the Gaussian kernel with $h = 1{\cdot}75$) was fitted again to the 90 data points in a newly simulated sample used to estimate (train) the intercept and the nine non-parametric coefficients. In addition, the following parametric specifications were fitted: additive model (3 location parameters: intercept and additive effects of the A and B loci), additive + dominance model (2 extra parameters corresponding to dominance effects at the two loci), and additive + dominance + epistasis (4 additional *df* pertaining to additive × additive, additive × dominance, dominance × additive and dominance × dominance interactions). The corresponding regression coefficients for the parametric models were 'shrunken' using a common variance ratio, corresponding to each of the 15 λ values employed in the RKHS fitting. Subsequently, 100 predictive samples of size 45 each were simulated, and the realized values were compared against the predictions obtained from the sample used to train either the RKHS regression or the three parametric models.

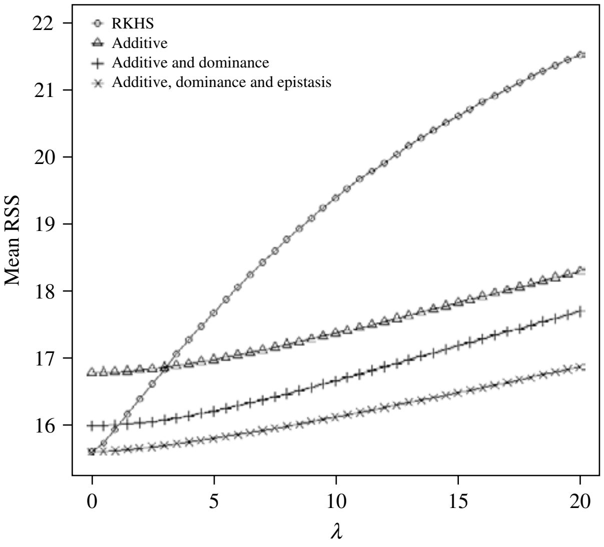

Fig. 4. Average (over 90 data points) squared residual for four models plotted to the training sample (RKHS = RKHS regression with Gaussian kernel and $h = 1{\cdot}75$) for each value of the smoothing parameter λ.

Figure 4 displays the average squared residual (over the 90 data points in the training sample) for each of the four models fitted. As expected, the RHKS and the additive + dominance + epistatic models fitted the data best at $\lambda = 0$ (no shrinkage), because of having a larger effective number of parameters (9) than the additive (3) and additive + dominance (5) models. When effects were gradually shrunken (λ increased from 0 to 20), the parametric models maintained their relative standings, whereas RKHS voyaged through all three models, eventually giving a very poor fit, due to oversmoothing. The trajectory of the effective *df* is shown in Fig. 5, where the ability of RKHS to explore models of different degrees of complexity becomes clear. The out-of-sample predictive performance of the four models is shown in Fig. 6. The simplest model, that is, one with additive effects at each of the two loci fitted, had the best predictive performance, even better than the two additional parametric specifications although the simulated gene action was not additive at all! RKHS regression was competitive, but its predictive ability deteriorated markedly when λ was greater than 5 in the training sample.

How does one explain the paradox that a simple additive model had better predictive performance when gene action was non-linear, as simulated here? In order to address this question, consider the 'true' mean value of the 9 genotypes simulated:

	BB	*Bb*	*bb*
AA	11·933	8·000	6·417
Aa	3·626	2·919	2·757
Aa	0·916	0·304	0·185

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

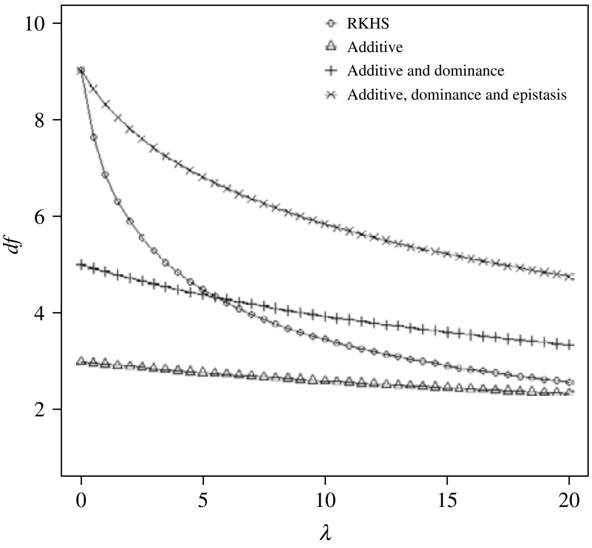

Fig. 5. Effective *df* for four models plotted to the training sample (RKHS = RKHS regression with Gaussian kernel and $h = 1{\cdot}75$) at each value of the smoothing parameter λ.

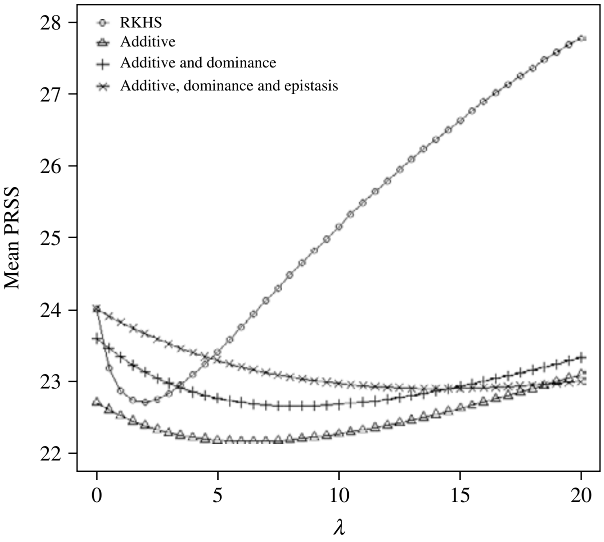

Fig. 6. Average (over 100 samples with 45 realized observations in each) squared prediction error for four models plotted to the predictive sample (RKHS = RKHS regression with Gaussian kernel and $h = 1{\cdot}75$) for each value of the smoothing parameter λ.

The 'corrected' sum of squares among these means is 125·23. A fixed effects ANOVA of these 'true' values (assuming genotypes were equally frequent) gives the following partition of sequential sum of squares, apart from rounding errors: (i) additive effect of locus *A*: 82·8%; (ii) additive effect of locus *B* after accounting for *A*: 7·06%; (iii) dominance effects of loci *A* and *B*: 4·2%, and (iv) epistasis: 6·2%. Thus, even though the genetic system was non-linear, most of the variation among genotypic means can be accounted for with a linear model on additive effects. The additive model had the worst fit to the data (even worse than the models that assume dominance and epistasis) and, yet, it had the best predictive ability, followed by RKHS for (roughly) $0{\cdot}5 < \lambda < 3$.

The simulation was also carried out at larger values of the residual variance, $\sigma_e^2 = 100$ and 500. Again, the purely additive model had the best predictive performance, but the difference between models essentially disappeared for $\lambda > 50$. In particular, the average squared prediction error of RKHS was larger than that for the additive model by 5, 3 and 1% for $\sigma_e^2 = 20$, 100 and 500, respectively, when evaluated at the 'optimal' λ in each case.

It should be noted that the RKHS implementation used here was completely arbitrary. For example, the kernel chosen was not the result of any formal model comparison, so predictive performance could be enhanced by a more judicious choice of kernel. As noted earlier, the choice of a good kernel is critical in this form of non-parametric modelling.

7. Conclusion

Inference about genotypes and future phenotypes for a complex quantitative trait was discussed in this paper. In particular, it was argued that the Kempthorne–Cockerham theory for partioning variance into additive, dominance and epistatic components has doubtful usefulness, because practically all assumptions required are violated in artificial and natural populations. This theory is probably illusory when genetic systems are complex and non-linear, in agreement with views in Feldman & Lewontin (1975) and Karlin *et al.* (1983). Further, an ANOVA-type decomposition is inadequate for a non-linear system (because the ANOVA model is linear in the parameters), and unfeasible in a highly dimensional and interactive genetic system involving hundreds or thousands of genes. As a minimum, the ANOVA treatment would encounter a huge number of empty cells, extreme lack of orthogonality, and high-order interactions would be extremely difficult to interpret, in the usual sense. Last but not the least, the ANOVA model would require more *df* than the number of data points available for analysis.

For these reasons, a predictive approach was advanced in this paper, focused on the use of non-parametric methods, especifically RKHS regression. Use of this methodogy in conjunction with standard theory of quantitative genetics leads to non-parametric estimates of additive, dominance and epistatic effects. These ideas were illustrated using a stylyzed example in Henderson (1985), and it was shown how additive, dominance and total genetic values can be predicted using a single smoothing parameter (in addition to the residual variance) coupled with kernels based on substantive theory, a point that is also made in de los

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

Campos *et al.* (2008). A non-linear 2-locus system was simulated as well, to illustrate the RKHS approach, which was found to have a better out-of-sample predictive performance of means than the standard 2-locus fixed effects model (with and without epistasis).

On the other hand, a 2-locus model with estimates of additive effects shrunken to different degrees had the best performance when predicting future individual observations. This was explained by the observation that most of the variation among genotypic means could be accounted for by 'additive effects'. This is consistent with theoretical and empirical results presented by Hill *et al.* (2008), who gave evidence that, even in the presence of non-additive genetic action, most variance is of an additive type. Even though molecular geneticists view the additive model as irritatingly reductive, our results give reassurance to a common practice in animal breeding, i.e. predict genetic values using additive theory only. It is unknown, however, to what extent these results (from a predictive perspective) carry to more complex systems, difficult to be described suitably with naive linear models. In such situations, RKHS regression could be valuable, because of its ability to navigate through models of different degrees of complexity and, as shown here, it can be very competitive when the smoothing parameters are tuned properly.

In conclusion, it is felt that the non-parametric methods discussed here coupled with machine learning procedures, such as in Long *et al.* (2007) and Long *et al.* (2008), could play an important role in quantitative genetics in the post-genomic era.

Useful comments made by two anonymous reviewers and by Professor William G. Hill and Dr Andres Legarra are acknowledged. Professor Daniel Sorensen is thanked for suggesting that a 2-locus additive model with shrinkage of effects could have good predictive value, which was confirmed. Part of this work was carried out while the senior author was a Visiting Professor at Georg-August-Universität, Göttingen (Alexander von Humboldt Foundation Senior Researcher Award), and Visiting Scientist at the Station d'Amélioration Génétique des Animaux, Centre de Recherche de Toulouse (Chaire D'Excellence Pierre de Fermat, Agence Innovation, Midi-Pyreneés). Support by the Wisconsin Agriculture Experiment Station and by NSF DMS-NSF grant DMS-044371 to the first author is acknowledged.

References

Bost, B., Dillmann, C. & De Vienne, D. (1999). Fluxes and metabolic pools as model traits for quantitative genetics. *Genetics* **153**, 2001–2012.

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford: Clarendon Press.

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

Cheverud, J. M. & Routman, E. J. (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.

Cockerham, C. C. (1956). Effect of linkage on the covariances between relatives. *Genetics* **41**, 138–141.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

de los Campos, G., Gianola, D. & Rosa, G. J. M. (2008). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* (accepted).

Dempster, E. R. & Lerner, I. M. (1950). Heritability of threshold characters. *Genetics* **35**, 212–236.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. New York: Longman.

Feldman, M. W. & Lewontin, R. C. (1975). The heritability hang-up. *Science* **190**, 1163–1168.

Gallais, A. (1974). Covariances between arbitrary relatives with linkage and epistasis in the case of linkage disequilibrium. *Biometrics* **30**, 429–446.

Gianola, D. (1982). Theory and analysis of threshold characters. *Journal of Animal Science* **54**, 1079–1096.

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

Gianola, D. & van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces methods for genomic assisted prediction of quantitative traits. *Genetics* **178**, 2289–2303.

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M. & Avendaño, S. (2008*a*). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* **178**, 2305–2313.

Gonzalez-Recio, O., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Avendaño, S. (2008*b*). Genome-assisted prediction of a quantitative trait in parents and progeny: application to food conversion rate in chickens. *Genetics Selection Evolution*, submitted.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding Genetics Symposium in Honour of J. L. Lush*, pp. 10–41. Champaign, IL: American Society of Animal Science and American Dairy Science Association.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, ON: University of Guelph.

Henderson, C. R. (1985). Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *Journal of Animal Science* **60**, 111–117.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics* **4**, el000008.

Karlin, S., Cameron, E. C. & Chakraborty, R. (1983). Path analysis in genetic epidemiology: a critique. *American Journal of Human Genetics* **35**, 695–732.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London, Series B* **143**, 103–113.

Kempthorne, O. (1978). Logical, epistemological and statistical aspects of nature-nurture data interpretation. *Biometrics* **34**, 1–23.

Kimeldorf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **35**, 82–95.

Kojima, K. I. (1959). Role of epistasis and overdominance in stability of equilibria with selection. *Proceedings of*

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press

the National Academy of Sciences of the USA **45**, 984–989.

Lee, H. K. H. (2004). *Bayesian Nonparametrics Via Neural Networks*. Philadelphia, PA: ASA-SIAM.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. & Avendaño, S. (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics* **124**, 377–389.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Avendaño, S. (2008). Marker-assisted assessment of genotype by environment interaction: a case study of SNP–mortality association in broilers in two hygiene environments. *Journal of Animal Science* (in press).

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. (2008). Comparison of approaches for machine learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genetic Epidemiology* **32**, 325–340.

Schnell, F. W. (1963). The covariance between relatives in the presence of linkage. In *Statistical Genetics and Plant Breeding*. (ed. W. D. Hanson & H. F. Robinson), pp. 463–483. Washington, DC: National Academy of Sciences – National Research Council.

Sorensen, D. & Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. New York: Springer.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Templeton, A. R. (2000). Epistasis and complex traits. In *Epistasis and the Evolutionary Process* (ed. J. B. Wolf, E. D. Brodie III and M. J. Wade), pp. 41–57. New York: Oxford University Press.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

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

Wang, C. S., Rutledge, J. J. & Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution* **25**, 41–62.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics Selection Evolution* **26**, 91–115.

Wang, T. & Zeng, Z.-B. (2006). Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium. *BMC Genetics* **7**, 9.

Weir, B. S. & Cockerham, C. C. (1977). Two-locus theory in quantitative genetics. In *Proceedings of the International Conference on Quantitative Genetics*, pp. 247–269. Ames, IA: Iowa State University Press.

https://doi.org/10.1017/S0016672308009890 Published online by Cambridge University Press