Genet. Res., Camb. (2010), **92**, *pp.* 295–308. © Cambridge University Press 2010
doi:10.1017/S0016672310000285

Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods

GUSTAVO DE LOS CAMPOS[1,2]*†, DANIEL GIANOLA[1], GUILHERME J. M. ROSA[1], KENT A. WEIGEL[1] AND JOSÉ CROSSA[2]
[1] *University of Wisconsin-Madison, 1675 Observatory Drive, WI 53706, USA*
[2] *International Maize and Wheat Improvement Center (CIMMYT), Ap. Postal 6-641, 06600, México DF, México*

(*Received 5 April 2010 and in revised form 11 June 2010*)

Summary

Prediction of genetic values is a central problem in quantitative genetics. Over many decades, such predictions have been successfully accomplished using information on phenotypic records and family structure usually represented with a pedigree. Dense molecular markers are now available in the genome of humans, plants and animals, and this information can be used to enhance the prediction of genetic values. However, the incorporation of dense molecular marker data into models poses many statistical and computational challenges, such as how models can cope with the genetic complexity of multi-factorial traits and with the curse of dimensionality that arises when the number of markers exceeds the number of data points. Reproducing kernel Hilbert spaces regressions can be used to address some of these challenges. The methodology allows regressions on almost any type of prediction sets (covariates, graphs, strings, images, etc.) and has important computational advantages relative to many parametric approaches. Moreover, some parametric models appear as special cases. This article provides an overview of the methodology, a discussion of the problem of kernel choice with a focus on genetic applications, algorithms for kernel selection and an assessment of the proposed methods using a collection of 599 wheat lines evaluated for grain yield in four mega environments.

1. Introduction

Prediction of genetic values is relevant in plant and animal breeding, as well as for assessing the probability of disease in medicine. Standard genetic models view phenotypic outcomes (y_i; $i=1, \ldots, n$) as the sum of a genetic signal (g_i) and of a residual (ε_i), that is: $y_i=g_i+\varepsilon_i$. The statistical learning problem consists of uncovering genetic signal from noisy data, and predictions ($\hat{g}_i$) are constructed using phenotypic records and some type of knowledge about the genetic background of individuals.

Family structure, usually represented as a pedigree, and phenotypic records have been used for the prediction of genetic values in plants and animals over several decades (e.g. Fisher, 1918; Wright, 1921; Henderson, 1975). In pedigree-based models (P), a genealogy is used to derive the expected degree of resemblance between relatives, measured as $\mathrm{Cov}(g_i, g_{i'})$, and this provides a means for smoothing phenotypic records.

Dense molecular marker panels are now available in humans and in many plant and animal species. Unlike pedigree data, genetic markers allow follow-up of Mendelian segregation; a term that in additive models and in the absence of inbreeding accounts for 50% of the genetic variability. However, incorporating molecular markers into models poses several statistical and computational challenges such as how models can cope with the genetic complexity of multi-factorial traits (e.g. Gianola & de los Campos, 2008), and with the curse of dimensionality that arises when a large number of markers is considered. Parametric and semi-parametric methods address these two issues in different ways.

* Corresponding author: 1665 University Boulevard, Ryals Public Health Building 414, AL 35294, USA. e-mail: gcampos@uab.edu
† Now at the Section of Statistical Genetics, Department of Biostatistics, University of Alabama-Birmingham, 1665 University Boulevard, AL 35294, USA.

In parametric regression models for dense molecular markers (e.g. Meuwissen *et al.*, 2001), g_i is a parametric regression on marker covariates, x_{ik} with $k=1, \ldots, p$ indexing markers. The linear model takes the form: $y_i=\sum_{k=1}^{p} x_{ik}\beta_k+\varepsilon_i$, where β_k is the regression of y_i on x_{ik}. Often, $p>>n$ and some shrinkage estimation method such as ridge regression (Hoerl & Kennard, 1970*a*, 1970*b*) or LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996), or their Bayesian counterparts, are used to estimate marker effects. Among the latter, those using marker-specific shrinkage such as the Bayesian LASSO of Park & Casella (2008) or methods BayesA or BayesB of Meuwissen *et al.* (2001) are the most commonly used. In linear regressions, dominance and epistasis may be accommodated by adding appropriate interactions between marker covariates to the model; however, the number of predictor variables is extremely large and modelling interactions is only feasible to a limited degree.

Reproducing kernel Hilbert spaces (RKHS) regressions have been proposed for semi-parametric regression on marker genotypes, e.g. Gianola *et al.* (2006) and Gianola & van Kaam (2008). In RKHS, markers are used to build a covariance structure among genetic values; for example, $\mathrm{Cov}(g_i, g_{i'})\propto K(\mathbf{x}_i, \mathbf{x}_{i'})$, where $\mathbf{x}_i$, $\mathbf{x}_{i'}$ are vectors of marker genotypes and $K(.,.)$, the reproducing kernel (RK), is some positive definite (PD) function (de los Campos *et al.*, 2009*a*). This semi-parametric approach has several attractive features: (a) the methodology can be used with almost any type of information set (e.g. covariates, strings, images and graphs). This is particularly important because techniques for characterizing genomes change rapidly; (b) some parametric methods for genomic selection (GS) appear as special cases and (c) computations are performed in an n-dimensional space. This provides RKHS methods with a great computational advantage relative to some parametric methods, especially when $p>>n$.

This article discusses and evaluates the use of RKHS regressions for genomic-enabled prediction of genetic values of complex traits. Section 2 gives a brief review of RKHS regressions. A special focus is placed on the problem of kernel choice. We discuss cases where a genetic model (e.g. additive infinitesimal) is used to choose the kernel and others where the RK is chosen based on its properties (e.g. predictive ability). Section 3 presents an application to an extensive plant breeding data set where some of the methods discussed in Section 2 are evaluated. Concluding remarks are provided in Section 4.

2. RKHS regression

RKHS methods have been used in many areas of application such as spatial statistics (e.g. 'Kriging'; Cressie, 1993), scatter-plot smoothing (e.g. smoothing splines; Wahba, 1990) and classification problems (e.g. support vector machines; Vapnik, 1998), just to mention a few. Estimates in RKHS regressions can be motivated as solutions to a penalized optimization problem in an RKHS or as posterior modes in a certain class of Bayesian models. A brief description of RKHS estimates in the context of penalized estimation is given first in section 2(i), with its Bayesian interpretation introduced later in section 2(ii). A representation of RKHS regressions that uses orthogonal basis functions is given in section 2(iii). This section ends in 2(iv) with a discussion of the problem of kernel choice.

(i) *Penalized estimation in RKHS*

A standard problem in statistical learning consists of extracting signal from noisy data. The learning task can be described as follows (Vapnik, 1998): given data $\{(y_i, t_i)\}_{i=1}^{n}$, originating from some functional dependency, infer this dependency. The pattern relating input, $t_i\in T$, and output, $y_i\in Y$, variables can be described with an unknown function, g, whose evaluations are $g_i=g(t_i)$. For example, t_i may be a vector of marker genotypes, $t_i=\mathbf{x}_i$ and g may be a function assigning a genetic value to each genotype. Inferring g requires defining a collection (or space) of functions from which an element, $\hat{g}$, will be chosen via a criterion (e.g. a penalized residual sum of squares or a posterior density) for comparing candidate functions. Specifically, in RKHS, estimates are obtained by solving the following optimization problem:

$$\hat{g}=\underset{g\in H}{\arg\min}\{l(g,\mathbf{y})+\lambda\|g\|_H^2\}, \tag{1}$$

where $g\in H$ denotes that the optimization problem is performed within the space of functions H, a RKHS; $l(g, \mathbf{y})$ is a loss function (e.g. some measure of goodness of fit); λ is a parameter controlling trade-offs between goodness of fit and model complexity; and $\|g\|_H^2$ is the square of the norm of g on H, a measure of model complexity. A technical discussion of RKHS of real-valued functions can be found in Wahba (1990); here, we introduce some elements that are needed to understand how $\hat{g}$ is obtained.

Hilbert spaces are complete linear spaces endowed with a norm that is the square root of the inner product in the space. The Hilbert spaces that are relevant for our discussion are RKHS of real-valued functions, here denoted as H. An important result, known as the Moore–Aronszajn theorem (Aronszajn, 1950), states that each RKHS is uniquely associated with a PD function that is a function, $K(t_i, t_{i'})$, satisfying $\sum_i\sum_{i'}\alpha_i\alpha_{i'}K(t_i, t_{i'})>0$ for all sequences, $\{\alpha_i\}$, with $\alpha_i\neq 0$ for some i. This function, $K(t_i, t_{i'})$, also known as the RK, provides basis functions and an inner

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

product (therefore a norm) to H. Therefore, choosing $K(t_i, t_{i'})$ amounts to selecting H; the space of functions where (1) is solved.

Using that duality, Kimeldorf & Wahba (1971) showed that the finite-dimensional solution of (1) admits a linear representation $g(t_i)=\sum_{i'}K(t_i, t_{i'})\alpha_{i'}$, or in matrix notation, $\mathbf{g}=\mathbf{K}\boldsymbol{\alpha}=[g(t_1), \ldots, g(t_n)]'$, where $\mathbf{K}=\{K(t_i, t_{i'})\}$ is an $n\times n$ matrix whose entries are the evaluations of the RK at pairs of pint in T. Further, in this finite-dimensional setting, $\|g\|^2_H=\sum_i\sum_{i'}\alpha_i\alpha_{i'}K(t_i, t_{i'})=\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$. Using this in (1) and setting $l(g, \mathbf{y})$ to be a residual sum of squares, one obtains: $\hat{\mathbf{g}}=\mathbf{K}\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}=(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)'$ is the solution of

$$\hat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}}{\arg\min}\{(\mathbf{y}-\mathbf{K}\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{K}\boldsymbol{\alpha})+\lambda\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}\} \quad (2)$$

and $\mathbf{y}=\{y_i\}$ is a data vector. The first-order conditions of (2) lead to $(\mathbf{K}'\mathbf{K}+\lambda\mathbf{K})\hat{\boldsymbol{\alpha}}=\mathbf{K}'\mathbf{y}$. Further, since $\mathbf{K}=\mathbf{K}'$ and $\mathbf{K}^{-1}$ exists, pre-multiplication by $\mathbf{K}^{-1}$ yields, $[\mathbf{K}+\lambda\mathbf{I}]\hat{\boldsymbol{\alpha}}=\mathbf{y}$. Therefore, the estimated conditional expectation function is $\hat{\mathbf{g}}=\mathbf{K}\hat{\boldsymbol{\alpha}}=\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}=\mathbf{P}(\lambda, K)\mathbf{y}$, where $\mathbf{P}(\lambda, K)=\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}$ is a smoother or influence matrix.

The input information, $t_i\in T$, enters into the objective function and on the solution only through $\mathbf{K}$. This allows using RKHS for regression with any class of information sets (vectors, graphs, images, etc.) where a PD function can be evaluated; the choice of kernel becomes the key element of model specification.

(ii) *Bayesian interpretation*

From a Bayesian perspective, $\hat{\boldsymbol{\alpha}}$ can be viewed as a posterior mode in the following model: $\mathbf{y}=\mathbf{K}\boldsymbol{\alpha}+\boldsymbol{\varepsilon}$; $P(\boldsymbol{\varepsilon}, \boldsymbol{\alpha}|\sigma^2_\varepsilon, \sigma^2_g)=N(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon)N(\boldsymbol{\alpha}|\mathbf{0}, \mathbf{K}^{-1}\sigma^2_g)$. The relationship between RKHS regressions and Gaussian processes was first noted by Kimeldorf & Wahba (1970) and has been revisited by many authors (e.g. Harville, 1983; Speed, 1991). Following de los Campos *et al.* (2009*a*), one can change variables in the above model, with $\mathbf{g}=\mathbf{K}\boldsymbol{\alpha}$, yielding

$$\begin{cases}\mathbf{y}=\mathbf{g}+\boldsymbol{\varepsilon},\\ p(\boldsymbol{\varepsilon}, \mathbf{g}|\sigma^2_\varepsilon, \sigma^2_g)=N(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon)N(\mathbf{g}|\mathbf{0}, \mathbf{K}\sigma^2_g).\end{cases} \quad (3)$$

Thus, from a Bayesian perspective, the evaluations of functions can be viewed as Gaussian processes satisfying $\mathrm{Cov}(g_i, g_{i'})\propto K(t_i, t_{i'})$. The fully Bayesian RKHS regression assumes unknown variance parameters, and the model becomes

$$\begin{cases}\mathbf{y}=\mathbf{g}+\boldsymbol{\varepsilon},\\ p(\boldsymbol{\varepsilon}, \mathbf{g}, \sigma^2_\varepsilon, \sigma^2_g)=N(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon)N(\mathbf{g}|\mathbf{0}, \mathbf{K}\sigma^2_g)p(\sigma^2_\varepsilon, \sigma^2_g),\end{cases} \quad (4)$$

where $p(\sigma^2_\varepsilon, \sigma^2_g)$ is a (proper) prior density assigned to variance parameters.

(iii) *Representation using orthogonal random variables*

Representing model (4) with orthogonal random variables simplifies computations greatly and provides additional insights into the nature of the RKHS regressions. To this end, we make use of the eigenvalue (EV) decomposition (e.g. Golub & Van Loan, 1996) of the kernel matrix $\mathbf{K}=\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}$ is a matrix of eigenvectors satisfying $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}=\mathbf{I}$; $\boldsymbol{\Psi}=\mathrm{Diag}\{\Psi_j\}$, $\Psi_1\geqslant\Psi_2\geqslant\ldots\geqslant\Psi_n>0$, is a diagonal matrix whose non-zero entries are the EVs of $\mathbf{K}$; and $j=1, \ldots, n$, indexes eigenvectors (i.e. columns of $\boldsymbol{\Lambda}$) and the associated EV. Using these, (4) becomes

$$\begin{cases}\mathbf{y}=\boldsymbol{\Lambda}\boldsymbol{\delta}+\boldsymbol{\varepsilon},\\ p(\boldsymbol{\varepsilon}, \boldsymbol{\delta}, \sigma^2_\varepsilon, \sigma^2_g)\propto N(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon)N(\boldsymbol{\delta}|\mathbf{0}, \boldsymbol{\Psi}\sigma^2_g)p(\sigma^2_\varepsilon, \sigma^2_g).\end{cases} \quad (5)$$

To see the equivalence of (4) and (5), note that $\boldsymbol{\Lambda}\boldsymbol{\delta}$ is multivariate normal because so is $\boldsymbol{\delta}$. Further, $E(\boldsymbol{\Lambda}\boldsymbol{\delta})=\boldsymbol{\Lambda}E(\boldsymbol{\delta})=\mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\Lambda}\boldsymbol{\delta})=\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'\sigma^2_g=\mathbf{K}\sigma^2_g$. Therefore, equations (4) and (5) are two parameterizations of the same probability model. However, equation (5) is much more computationally convenient, as discussed next.

The joint posterior distribution of (5) does not have a closed form; however, draws can be obtained using a Gibbs sampler. Sampling regression coefficients from the corresponding fully conditional distribution, $p(\boldsymbol{\delta}|\mathbf{y}, \sigma^2_\varepsilon, \sigma^2_g)$, is usually the most computationally demanding step. From standard results of Bayesian linear models, one can show that $p(\boldsymbol{\delta}|\mathrm{ELSE})=N(\hat{\boldsymbol{\delta}}, \sigma^2_\varepsilon\mathbf{C}^{-1})$, where ELSE denotes everything else other than $\boldsymbol{\delta}$, $\mathbf{C}=[\boldsymbol{\Lambda}'\boldsymbol{\Lambda}+\sigma^2_\varepsilon\sigma_g^{-2}\boldsymbol{\Psi}^{-1}]=\mathrm{Diag}\{1+\sigma^2_\varepsilon\sigma_g^{-2}\Psi_j^{-1}\}$ and $\hat{\boldsymbol{\delta}}=\mathbf{C}^{-1}\boldsymbol{\Lambda}'\mathbf{y}$. This simplification occurs because $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}=\mathbf{I}$ and $\boldsymbol{\Psi}=\mathrm{Diag}\{\Psi_j\}$. The fully conditional distribution of $\boldsymbol{\delta}$ is multivariate normal, and the (co)variance matrix, $\sigma^2_\varepsilon\mathbf{C}^{-1}$, is diagonal; therefore $p(\boldsymbol{\delta}|\mathrm{ELSE})=\prod_{j=1}^n p(\delta_j|\mathrm{ELSE})$. Moreover, $p(\delta_j|\mathrm{ELSE})$ is normal, centred at $[1+\sigma^2_\varepsilon\sigma_g^{-2}\Psi_j^{-1}]^{-1}y_{.j}$ and with variance $\sigma^2_\varepsilon[1+\sigma^2_\varepsilon\sigma_g^{-2}\Psi_j^{-1}]^{-1}$. Here, $y_{.j}=\boldsymbol{\lambda}'_j\mathbf{y}$, where $\boldsymbol{\lambda}_j$ is the jth eigenvector (i.e. the jth column of $\boldsymbol{\Lambda}$). Note that model unknowns are not required for computing $y_{.j}$, implying that these quantities remain constant across iterations of a sampler. The only quantities that need to be updated are $[1+\sigma^2_\varepsilon\sigma_g^{-2}\Psi_j^{-1}]$ and $\sigma^2_\varepsilon[1+\sigma^2_\varepsilon\sigma_g^{-2}\Psi_j^{-1}]$. If model (5) is extended to include other effects (e.g. an intercept or some fixed effects), the right-hand side of the mixed model equations associated to $p(\boldsymbol{\delta}|\mathrm{ELSE})$ will need to be updated at each iteration of the sampler; however, the matrix of coefficients remains diagonal and this simplifies computations greatly (see Appendix).

In equation (5), the conditional expectation function is a linear combination of eigenvectors: $\mathbf{g}=\boldsymbol{\Lambda}\boldsymbol{\delta}=\sum_j\boldsymbol{\lambda}_j\delta_j$. The EV are usually sorted such that $\Psi_1\geqslant\Psi_2\geqslant\ldots\geqslant\Psi_n>0$. The prior precision variance of regression coefficients is proportional to the EV, that

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

is, $\text{Var}(\delta_j) \propto \Psi_j$. Therefore, the extent of shrinkage increases as j does. For most RKs, the decay of the EV will be such that for the first EV $[1+\sigeps^2\sigma_g^{-2}\Psi_j^{-1}]$ is close to one, yielding negligible shrinkage of the corresponding regression coefficients. Therefore, linear combinations of the first eigenvectors can then be seen as components of g that are (essentially) not penalized.

(iv) *Choosing the RK*

The RK is a central element of model specification in RKHS. Kernels can be chosen so as to represent a parametric model, or based on their ability of predicting future observations. Examples of these two approaches are discussed next.

The standard additive infinitesimal model of quantitative genetics (e.g. Fisher, 1918; Henderson, 1975), is an example of a model-driven kernel (e.g. de los Campos *et al.*, 2009*a*). Here, the information set (a pedigree) consists of a directed acyclic graph and $K(t_i, t_{i'})$ gives the expected degree of resemblance between relatives under an additive infinitesimal model. Another example of an RKHS regression with a model-derived kernel is the case where $\mathbf{K}$ is chosen to be a marker-based estimate of a kinship matrix (usually denoted as $\mathbf{G}$, cf, Ritland, 1996; Lynch & Ritland, 1999; Eding & Meuwissen, 2001; Van Raden, 2007; Hayes & Goddard, 2008). An example of a (co)variance structure derived from a quantitative trait locus (QTL)-model is given in Fernando & Grossman (1989).

Ridge regression and its Bayesian counterpart (Bayesian ridge regression (BRR)) can also be represented using (4) or (5). A BRR is defined by $\mathbf{y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$ and $p(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigeps^2, \sigma_\beta^2) = N(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigeps^2)N(\boldsymbol{\beta}|\mathbf{0}, \mathbf{I}\sigma_\beta^2) \times p(\sigeps^2, \sigma_\beta^2)$. To see how a BRR constitutes a special case of (5), one can make use of the singular value decomposition (e.g. Golub & Van Loan, 1996) of $\mathbf{X}=\mathbf{UDV}'$. Here, $\mathbf{U}$ $(n\times n)$ and $\mathbf{V}$ $(p\times n)$ are matrices whose columns are orthogonal, and $\mathbf{D}=\text{Diag}\{\xi_j\}$ is a diagonal matrix whose non-null entries are the singular values of $\mathbf{X}$. Using this in the data equation, we obtain $\mathbf{y}=\mathbf{UDV}'\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{U}\boldsymbol{\delta}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\delta}=\mathbf{DV}'\boldsymbol{\beta}$. The distribution of $\boldsymbol{\delta}$ is multivariate normal because so is that of $\boldsymbol{\beta}$. Further, $E(\boldsymbol{\delta})=\mathbf{DV}'E(\boldsymbol{\beta})=\mathbf{0}$ and $\text{Cov}(\boldsymbol{\delta})=\mathbf{DV}'\mathbf{VD}'\sigma_\beta^2=\mathbf{DD}'\sigma_\beta^2$; thus, $\boldsymbol{\delta}\sim N[\mathbf{0}, \text{Diag}\{\xi_j^2\}\sigma_\beta^2]$. Therefore, a BRR can be equivalently represented using (5) with $\boldsymbol{\Lambda}=\mathbf{U}$ and $\boldsymbol{\Psi}=\text{Diag}\{\xi_j^2\}$. Note that using $\boldsymbol{\Lambda}=\mathbf{U}$ and $\boldsymbol{\Psi}=\text{Diag}\{\xi_j^2\}$ in (5) implies $\mathbf{K}=\mathbf{UDD}'\mathbf{U}'=\mathbf{UDV}'\mathbf{VD}'\mathbf{U}'=\mathbf{XX}'$ in (4). Habier Fernando & Dekkers (2009) argue that as the number of markers increases, $\mathbf{XX}'$ approaches the numerator relationship matrix, $\mathbf{A}$. From this perspective, $\mathbf{XX}'$ can also be viewed just as another choice for an estimate of a kinship matrix. However, the derivation of the argument follows the standard treatment of quantitative genetic models where genotypes are random and marker effects are fixed, whereas in BRR, the opposite is true (see Gianola *et al.*, 2009 for further discussion).

In the examples given above, the RK was defined in such a manner that it represents a parametric model. An appeal of using parametric models is that estimates can be interpreted in terms of the theory used for deriving $\mathbf{K}$. For example, if $\mathbf{K}=\mathbf{A}$ then σ_g^2 is interpretable as an additive genetic variance and $\sigma_g^2(\sigma_g^2+\sigeps^2)^{-1}$ can be interpreted as the heritability of the trait. However, these models may not be optimal from a predictive perspective. Another approach (e.g. Shawe-Taylor & Cristianini, 2004) views RKs as smoothers, with the choice of kernel based on their predictive ability or some other criterion. Moreover, the choice of the kernel may become a task of the algorithm.

For example, one can index a Gaussian kernel with a bandwidth parameter, θ, so that $K(t_i,t_{i'}|\theta)=\exp\{-\theta d(t_i,t_{i'})\}$. Here, $d(t_i,t_{i'})$ is some distance function and θ controls how fast the covariance function drops as points get further apart as measured by $d(t_i, t_{i'})$. The bandwidth parameter may be chosen by cross-validation (CV) or with Bayesian methods (e.g. Mallick *et al.*, 2005). However, when θ is treated as uncertain in a Bayesian model with Markov chain Monte Carlo (MCMC) methods, the computational burden increases markedly because the RK must be computed every time that a new sample of θ becomes available. It is computationally easier to evaluate model performance over a grid of values of θ; this is illustrated in section 3.

The (co)variance structure implied by a Gaussian kernel is not derived from any mechanistic consideration; therefore, no specific interpretation can be attached to the bandwidth parameter. However, using results for infinitesimal models under epistasis one could argue that a high degree of epistatic interaction between additive infinitesimal effects may induce a highly local (co)variance pattern in the same way that large values of θ do. This argument is revisited later in this section.

The decay of the EV controls, to a certain extent, the shrinkage of estimates of $\boldsymbol{\delta}$ and, with this, the trade-offs between goodness of fit and model complexity. Transformations of EV (indexed with unknown parameters) can also be used to generate a family of kernels. One such example is the diffusion kernel $\mathbf{K}_\alpha=\boldsymbol{\Lambda}\text{Diag}\{\exp(\alpha\Psi_j)\}\boldsymbol{\Lambda}'$ (e.g. Kondor & Lafferty, 2002). Here, $\alpha>0$ is used to control the decay of EV. In this case, the bandwidth parameter can be interpreted as a quantity characterizing the diffusion of signal (e.g. heat) along edges of a graph, with smaller values being associated with more diffusion.

A third way of generating families of kernels is to use closure properties of PD functions (Shawe-Taylor &

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

Cristianini, 2004). For example, linear combinations of PD functions, $\tilde{K}(t_i, t_{i'}) = \sigma^2_{g_1} K_1(t_i, t_{i'}) + \sigma^2_{g_2} K_2(t_i, t_{i'})$, with $\sigma^2_{g.} \geqslant 0$, are PD as well. From a Bayesian perspective, $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are interpretable as variance parameters. To see this, consider extending (4) to two random effects so that: $\mathbf{g} = \mathbf{g_1} + \mathbf{g_2}$ and, $p(\mathbf{g_1}, \mathbf{g_2}|\sigma^2_{g_1}, \sigma^2_{g_2}) = N(\mathbf{g_1}|\mathbf{0}, \mathbf{K_1}\sigma^2_{g_1}) N(\mathbf{g_2}|\mathbf{0}, \mathbf{K_2}\sigma^2_{g_2})$. It follows that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{K_1}\sigma^2_{g_1} + \mathbf{K_2}\sigma^2_{g_2})$, or equivalently $\mathbf{g} \sim N(\mathbf{0}, \tilde{\mathbf{K}}\tilde{\sigma}^2_g)$, where $\tilde{\sigma}^2_g = (\sigma^2_{g_1} + \sigma^2_{g_2})$ and $\tilde{\mathbf{K}} = \mathbf{K_1}\sigma^2_{g_1}\tilde{\sigma}_g^{-2} + \mathbf{K_2}\sigma^2_{g_2}\tilde{\sigma}_g^{-2}$. Therefore, fitting an RKHS with two random effects is equivalent to using $\tilde{\mathbf{K}}$ in (4). Extending this argument to r kernels one obtains: $\tilde{\mathbf{K}} = \sum_r \mathbf{K}_r \sigma^2_{g_r} \tilde{\sigma}_{g_r}^{-2}$, where $\tilde{\sigma}^2_g = \sum_r \tilde{\sigma}^2_{g_r}$. For example, one can obtain a sequence of kernels, $\{\mathbf{K}_r\}$, by evaluating a Gaussian kernel over a grid of values of a bandwidth parameter $\{\theta_r\}$. The variance parameters, $\{\sigma^2_{g_r}\}$, associated with each kernel in the sequence can be viewed as weights. Inferring these variances amounts to inferring a kernel, $\tilde{\mathbf{K}}$, which can be seen as an approximation to an optimal kernel. We refer to this approach as kernel selection via kernel averaging (KA); an example of this is given in section 3.

The Haddamard (or Schur) product of PD functions is also PD, that is, if $K_1(t_i, t_{i'})$ and $K_2(t_i, t_{i'})$ are PD, so is $K(t_i, t_{i'}) = K_1(t_i, t_{i'}) K_2(t_i, t_{i'})$; in matrix notation, this is usually denoted as $\mathbf{K} = \mathbf{K_1} \# \mathbf{K_2}$. From a genetic perspective, this formulation can be used to accommodate non-additive infinitesimal effects (e.g. Cockerham, 1954; Kempthorne, 1954). For example, under random mating, linkage equilibrium and in the absence of selection, $\mathbf{K} = \mathbf{A} \# \mathbf{A} = \{a(i, i')^2\}$ gives the expected degree of resemblance between relatives under an infinitesimal model for additive × additive interactions. For epistatic interaction between infinitesimal additive effects of qth order, the expected (co)variance structure is, $\mathbf{K} = \{a(i, i')^{q+1}\}$. Therefore, for $q \geqslant 1$ and $i \neq i'$, the prior correlation,

$$0 < \frac{a(i,i')^{q-1}}{\sqrt{a(i,i)^{q-1} a(i',i')^{q-1}}} = \left[\frac{a(i,i')}{\sqrt{a(i,i)a(i',i')}}\right]^{q-1} < \frac{a(i,i')}{\sqrt{a(i,i)a(i',i')}} < 1,$$

decreases, i.e. the kernel becomes increasingly local, as the degree of epistatic interaction increases, producing an effect similar to that of a bandwidth parameter of a Gaussian kernel.

3. Application to plant breeding data

Some of the methods discussed in the previous section were evaluated using a data set consisting of a collection of historical wheat lines from the Global Wheat Breeding Programme of CIMMYT (International Maize and Wheat Improvement Center). In plant breeding programmes, lines are selected based on their expected performance and collecting phenotypic records is expensive. An important question is whether phenotypes collected on ancestor lines, together with pedigrees and markers, can be used to predict performance of lines for which phenotypic records are not available yet. If so, breeding programmes could perform several rounds of selection based on marker data only; with phenotypes measured every few generations. The reduction in generation interval attainable by selection based on markers may increase the rate of genetic progress and, at the same time, the cost of phenotyping would be reduced (e.g. Bernardo & Yu, 2007; Heffner *et al*., 2009). Thus, assessing the ability of a model to predict future outcomes is central in breeding programmes.

The study presented in this section attempted to evaluate: (*a*) how much could be gained in predictive ability by incorporating marker information into a pedigree-based model, (*b*) how sensitive these results are with respect to the choice of kernel, (*c*) whether or not Bayesian KA is effective for selecting kernels and (*d*) how RKHS performs relative to a parametric regression model, the Bayesian LASSO (BL; Park & Casella, 2008).

(i) *Materials and methods*

The data comprise family, marker and phenotypic information of 599 wheat lines that were evaluated for grain yield (GY) in four environments. Single-trait models were fitted to data from each environment. Marker information consisted of genotypes for 1447 Diversity Array Technology (DArT) markers, generated by Triticarte Pty. Ltd (Canberra, Australia; http://www.triticarte.com.au). Pedigree information was used to compute additive relationships between lines (i.e. twice the kinship coefficient; Wright, 1921) using the Browse application of the International Crop Information System, as described in McLaren *et al*. (2005).

A sequence of models was fitted to the entire data set and in a CV setting. Figure 1 gives a summary of the models considered. In all environments, phenotypes were represented using equation $y_i = \mu + g_i + \varepsilon_i$, where y_i ($i = 1, \ldots, 599$) is the phenotype of the ith line; μ is an effect common to all lines; g_i is the genetic value of the ith line; and ε_i is a line-specific residual. Phenotypes were standardized to a unit variance in each of the environments. Residuals were assumed to follow a normal distribution $\varepsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2_\varepsilon)$, where σ^2_ε is the residual variance. The conditional distribution of the data was $p(\mathbf{y}|\mu, \mathbf{g}, \sigma^2_\varepsilon) = \prod_{i=1}^n N(y_i|\mu + g_i, \sigma^2_\varepsilon)$, where $\mathbf{g} = (g_1, \ldots, g_n)'$. Models differed on how g_i was modelled.

In a standard infinitesimal additive model (P, standing for pedigree-model), genetic values are $\mathbf{g} = \mathbf{a}$ with $p(\mathbf{a}|\sigma^2_a) = N(\mathbf{0}, \mathbf{A}\sigma^2_a)$, where σ^2_a is the additive

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

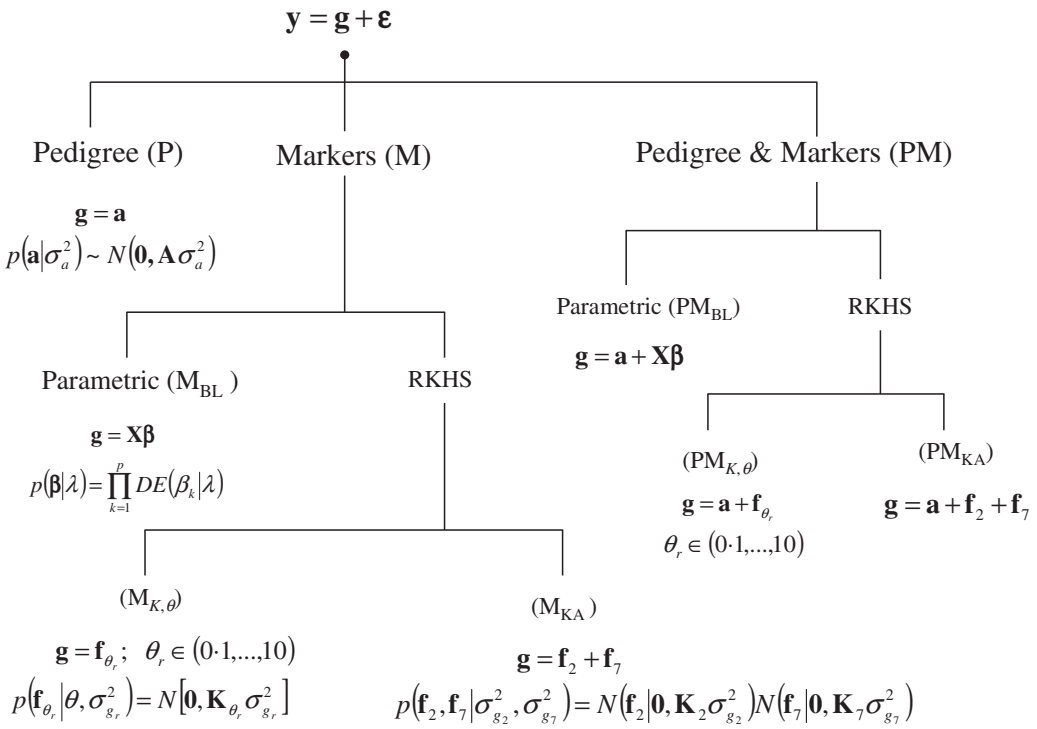

Fig. 1. Alternative models for prediction of genetic values. Phenotypic records (**y**) were always the sum of a genetic signal (**g**) and a vector of Gaussian residuals (**ε**). Models differed on how **g** was represented, as described in the figure. BL, Bayesian LASSO; RKHS, reproducing kernel Hilbert spaces regression; λ, LASSO regularization parameter; θ, RKHS bandwidth parameter; σ^2, variance parameter; KA, kernel averaging; N(., .), normal density; DE(.), double-exponential density.

genetic variance and $\mathbf{A}=\{a(i,i')\}$, as before, is the numerator relationship matrix among lines computed from the pedigree. This is a RKHS with $\mathbf{K}=\mathbf{A}$.

For marker-based models (M), two alternatives were considered: BL and RKHS regression. In the BL, genetic values were a linear function of marker covariates, $\mathbf{g}=\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an incidence matrix with marker genotypes codes and $\boldsymbol{\beta}=(\beta_1, \ldots, \beta_p)'$, the vector of regression coefficients, was inferred using the BL of Park & Casella (2008). Following de los Campos *et al*. (2009*b*), the prior density of the regularization parameter of the BL, here denoted as $\tilde{\lambda}$, was $p(\tilde{\lambda}) \propto \text{Beta}(\tilde{\lambda}/150|\tilde{\alpha}_1=1{\cdot}2, \tilde{\alpha}_2=1{\cdot}2)$, which is flat over a fairly wide range. This model is denoted as M_{BL}.

In marker-based RKHS regressions (M_{K}) $\mathbf{g}=\mathbf{f}_\theta$, where $\mathbf{f}_\theta=(f_{\theta,1}, \ldots, f_{\theta,n})'$ was assigned a Gaussian prior with null mean and (co)variance matrix $\text{Cov}(\mathbf{f}_\theta)\propto\mathbf{K}_\theta=\{\exp(-\theta k^{-1}d_{ii'})\}$. Here, θ is a bandwidth parameter, $d_{ii'}=\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2$ is the square Euclidean distance between marker codes $\mathbf{x}_i=(x_{i1}, \ldots, x_{ip})'$ and $\mathbf{x}_{i'}=(x_{i'1}, \ldots, x_{i'p})'$, and $k=\max_{(i,i')}\{\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2\}$. Models were fitted over a grid of values of θ and are denoted as $\text{M}_{k,\theta}$. The optimal value of the bandwidth parameter is expected to change with many factors such as: (a) distance function; (b) number of markers, allelic frequency and coding of markers, all factors affecting the distribution of observed distances and (c) genetic architecture of the trait, a factor affecting the expected prior correlation of genetic values (see section 2(iv)). We generated a grid of values, $\theta\in\{0{\cdot}1, 0{\cdot}25, 0{\cdot}5, 0{\cdot}75, 1, 2, 3, 5, 7, 10\}$, that for this data set allowed exploring a wide variety of kernels. Figure 2 gives a histogram of the evaluations of the kernel for two extreme values of the bandwidth parameter; $\theta=0{\cdot}25$ gives very high prior correlations, while $\theta=7$ gives a kernel matrix with very low correlations in the off-diagonal.

A model where **g** was the sum of two components: $\mathbf{g}=\mathbf{f}_{0{\cdot}25}+\mathbf{f}_7$, with $p(\mathbf{f}_{0{\cdot}25}, \mathbf{f}_7|\sigma^2_{g_{0{\cdot}25}}, \sigma^2_{g_7})=N(\mathbf{f}_{0{\cdot}25}|\mathbf{0}, \mathbf{K}_{0{\cdot}25}\sigma^2_{g_{0{\cdot}25}})N(\mathbf{f}_7|\mathbf{0}, \mathbf{K}_7\sigma^2_{g_7})$ was fitted as well. This model is referred to as M_{KA}, standing for marker-based model with 'kernel-averaging'. Note that $\mathbf{K}_{0{\cdot}25}$ and $\mathbf{K}_7$ provide very different kernels (see Fig. 2). With more extreme values of the bandwidth parameter, marker information is virtually lost. Indeed, choosing $\theta=0$ gives a kernel matrix full of ones and $\theta\to\infty$ gives $\mathbf{K}_\theta\to\mathbf{I}$, and averaging these two kernels gives a resulting (co)variance structure that does not use marker information at all.

Finally, a sequence of models including pedigree and marker data (PM) was obtained by setting $\mathbf{g}=\mathbf{a}+\mathbf{X}\boldsymbol{\beta}$, denoted as PM_{BL}; $\mathbf{g}=\mathbf{a}+\mathbf{f}_\theta$, $\theta=\{0{\cdot}1, 0{\cdot}25, 0{\cdot}5, 0{\cdot}75, 1, 2, 3, 5, 7, 10\}$, denoted as $\text{PM}_{k,\theta}$; and, $\mathbf{g}=\mathbf{a}+\mathbf{f}_{0{\cdot}25}+\mathbf{f}_7$, denoted as PM_{KA}.

In all models, variance parameters were treated as unknown and assigned identical independent scaled inverse chi-square prior distributions with three

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

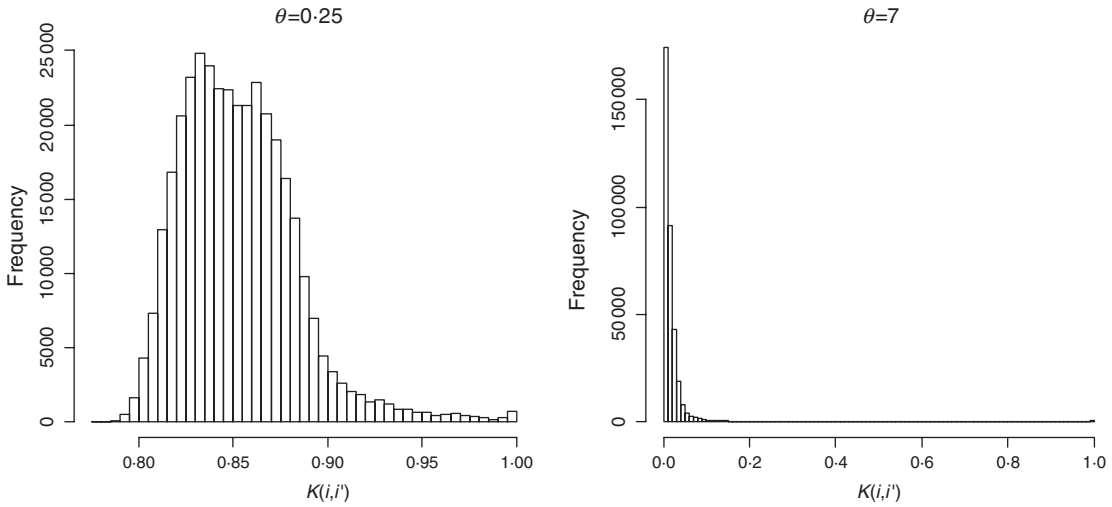

Fig. 2. Histogram of the evaluations of Gaussian kernel $K(i,i')=\exp\{-\theta k^{-1}d_{ii'}\}$ by value of the bandwidth parameter ($\theta=0{\cdot}25$ left and $\theta=7$, right). Here, $d_{ii'}=\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2$ is the squared Euclidean distance between marker codes $\mathbf{x}_i=(x_{i1}, \ldots, x_{ip})'$ and $\mathbf{x}_{i'}=(x_{i'1}, \ldots, x_{i'p})'$, and $k=\max_{(i,i')}\{\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2\}$.

degrees of freedom and scale parameters equal to 1, $p(\sigma^2)=\chi^{-2}(\sigma^2|\text{df}=3,\ S=1)$. Samples from posterior distributions for each of the models were obtained with a Gibbs sampler (see de los Campos *et al*., 2009*b*, for the case of M_{BL} and PM_{BL}, and the Appendix for RKHS models). Inferences were based on all 35 000 samples obtained after discarding 2000 samples as burn-in. The distribution of prediction errors was estimated using a 10-fold CV (e.g. Hastie *et al*., 2009).

(ii) *Results*

Figure 3 shows the posterior means of the residual variance in $\text{M}_{k,\theta}$ and $\text{PM}_{k,\theta}$ versus values of the bandwidth parameter θ obtained when models were fitted to the entire data. Each panel in Fig. 3 corresponds to one environment and the horizontal lines give the posterior means of the residual variance from *P* and PM_{KA}. Table A1 of the Appendix gives estimates of the posterior means and of the posterior standard deviations of the residual variance from each of the 25 models, by environment. The posterior means of the residual variances indicate that models M and PM fitted the data better than *P*, and PM_{KA} gave almost always better fit than $\text{M}_{k,\theta}$ and $\text{PM}_{k,\theta}$. In all environments, the posterior mean of the residual variance decreased monotonically with θ; this was expected because $\mathbf{K}_\theta$ becomes increasingly local as the bandwidth parameter increases. In environments 2, 3 and 4, the slopes of the curves relating the posterior mean of residual variance to θ were gentler for $\text{PM}_{k,\theta}$ than for $\text{M}_{k,\theta}$. This occurs, because in $\text{PM}_{k,\theta}$, the regression function has two components, one of which, the regression on the pedigree, is not a function of the bandwidth parameter. Models M_{BL} and PM_{BL} did not fit the training data as well as most of the RKHS counterparts, with a posterior mean of the residual variance that was close to that of $\text{M}_{k,\mathbf{0{\cdot}1}}$ and $\text{PM}_{k,\mathbf{0{\cdot}5}}$, respectively (see Table A1 of the Appendix).

The contribution of **a**, that is, the regression on the pedigree, to the conditional expectation function, **g**, can be assessed via the posterior mean of σ_a^2 (see Fig. A1 in the Appendix). The posterior mean of σ_a^2 was larger in P models than in their PM counterparts; this was expected, because in P the regression on the pedigree is the only component of the conditional expectation function that contributes to phenotypic variance. Within $\text{PM}_{k,\theta}$, the posterior mean of σ_a^2 was minimum at intermediate values of the bandwidth parameters. At extreme values of θ, the RK may not represent the types of patterns present in the data and, thus, the estimated conditional expectation function would depend more strongly on the regression on the pedigree (large values of σ_a^2).

Plots in Fig. 4 give the estimated mean-squared error (MSE) between CV predictions and observations versus values of the bandwidth parameter (*x*-axis), by environment and model. The predictive MSE of the *P* and PM_{KA} models are displayed as horizontal dashed lines, and values of those for the BL (both in M_{BL} and PM_{BL}) are shown at the bottom of the panels. Table A2 in the Appendix gives the estimated MSE by model and environment, respectively.

Overall, models including marker information had better predictive ability than pedigree-based models. For example, relative to P, using PM_{KA} yielded decreases in MSE between CV predictions of observations of 20·4, 8·8, 7·0 and 11·0% for E1 through E4, respectively (Table A2 in the Appendix). Thus, it appears that sizable gains in predictive ability can be attained by considering markers and pedigrees jointly,

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

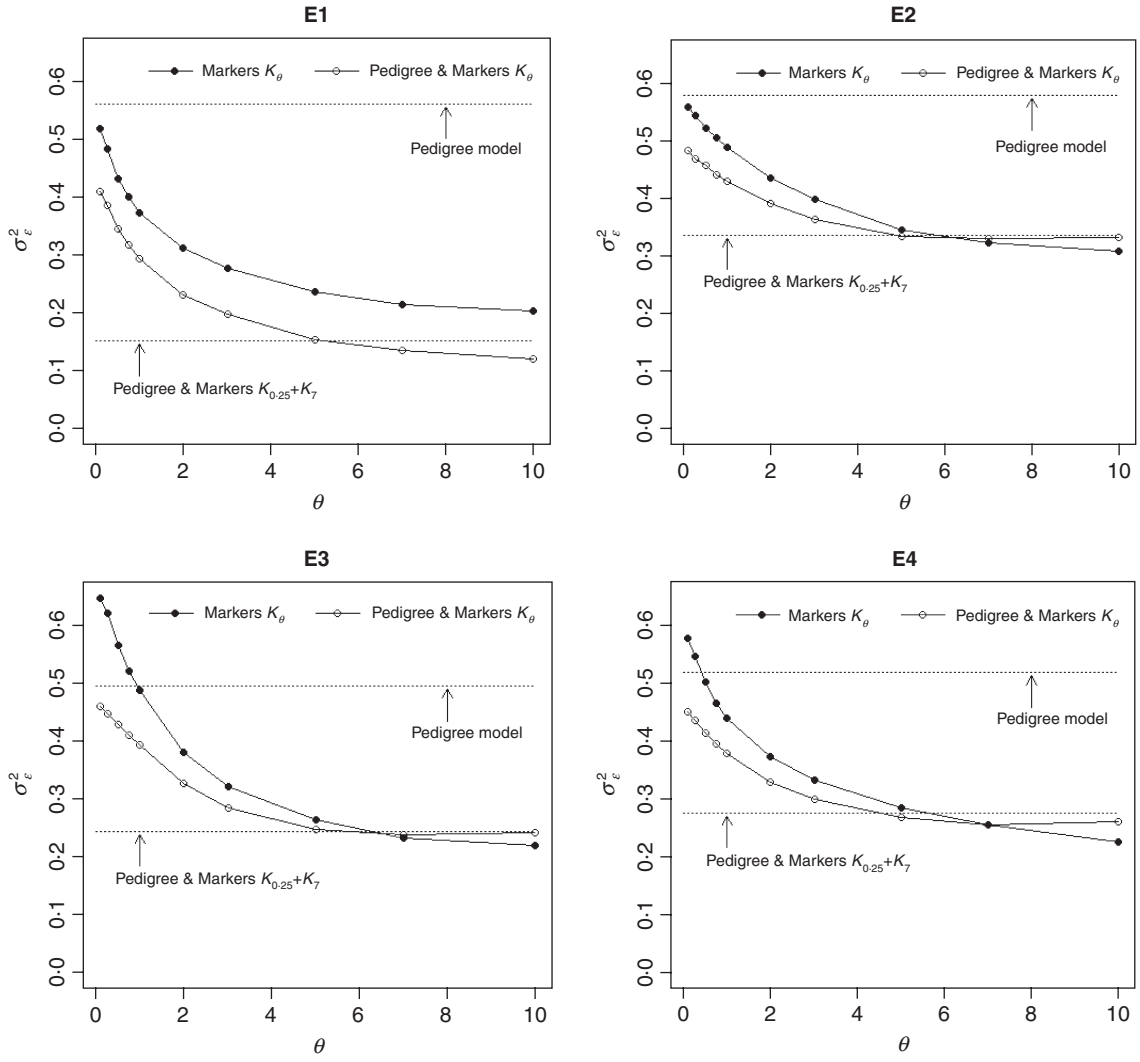

Fig. 3. Estimated posterior mean of the residual variance versus values of the bandwidth parameter, θ, by environment and model. K_θ is a marker-based RKHS regression with bandwidth parameter θ; Pedigree & Markers K_θ uses pedigree and markers, here, θ is the value of the bandwidth parameter for markers. Pedigree & Markers $K_{0.25}+K_7$ uses pedigree and markers with kernel averaging (KA) E1–E4: environments where the lines were evaluated.

as in PM_{KA}. These results are in agreement with some empirical studies (e.g. Corrada Bravo *et al*., 2009; de los Campos *et al*., 2009*b*) that provided evidence of a gain in predictive ability by jointly considering markers and pedigree information. However, marker density in this study was relatively low; as marker density increases it is expected that the relative importance of considering pedigree information will decrease (e.g. Calus & Veerkamp, 2007).

As shown in Fig. 4, the value of the bandwidth parameter that gave the best predictive ability was in the range (2,4), except for environment E2 in which values of θ near one performed slightly better. The value of the bandwidth parameter that was optimal from the perspective of predictive ability was similar in M and PM models (Fig. 4 and Table A2 in the Appendix). However, the difference between the predictive ability of $PM_{k,\theta}$ and $M_{k,\theta}$ models was larger for extreme values of θ, indicating that PM models are more robust than M models with respect to the choice of θ. Again, this occurs because $PM_{k,\theta}$ involves some form of KA (between the RK evaluated in the pedigree, **A**, and the one evaluated in marker genotypes, $\mathbf{K}_\theta$).

In all environments, KA had an estimated PMSE that was either close or lower than the one obtained with any specific value of the bandwidth parameter (Fig. 4 and Table A2 in the appendix). This was observed both in models with and without pedigree. These results suggest that KA can be an effective way of choosing the RK. Finally, PM_{KA} had higher predictive ability than PM_{BL}; this suggests a superiority of semi-parametric methods. However, PM_{BL} outperformed $PM_{k,\theta}$ for extreme values of the bandwidth parameter, illustrating, again, the importance of kernel selection. Moreover, the superiority of RKHS methods may not generalize to other traits or populations.

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

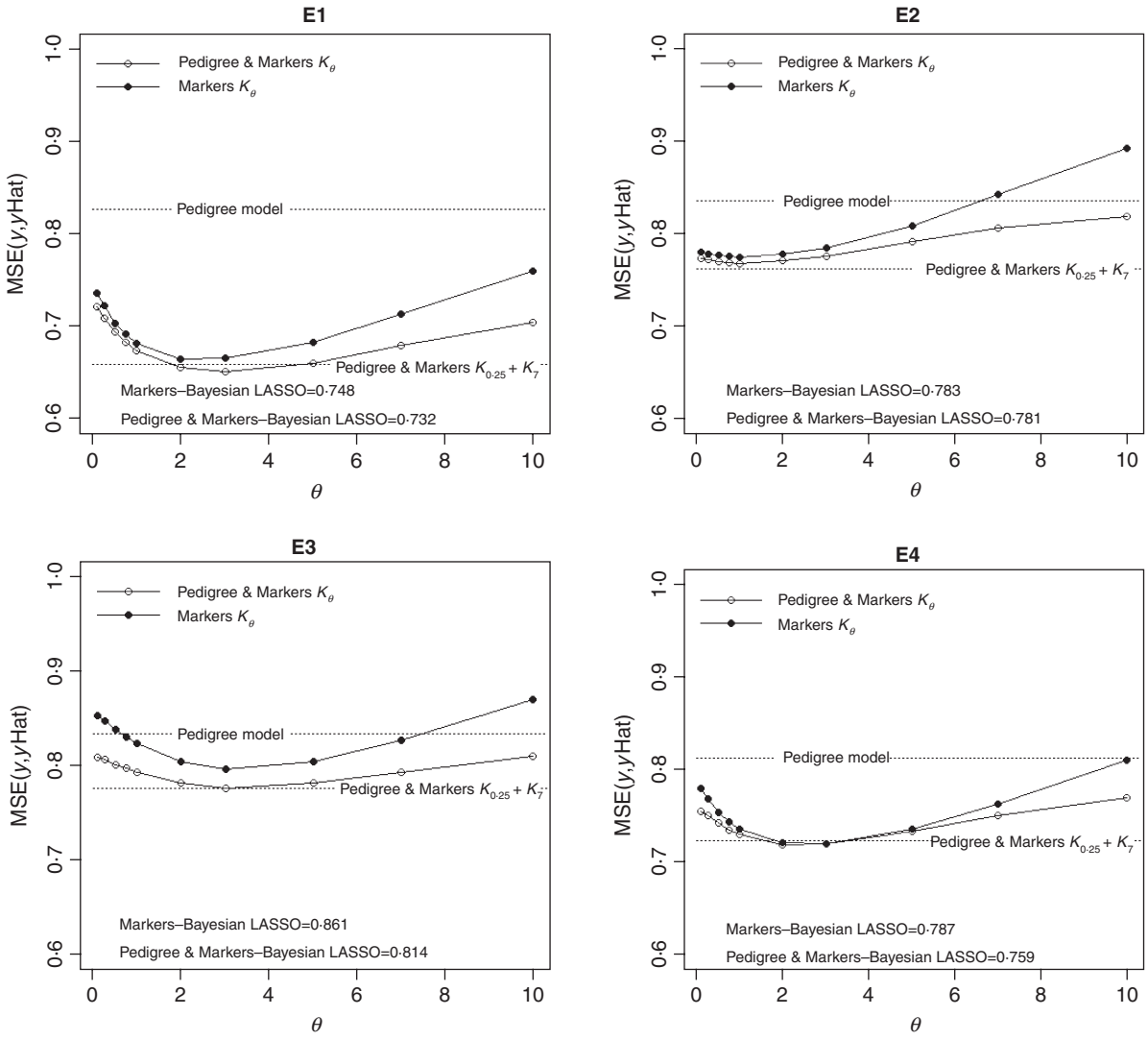

Fig. 4. Estimated MSE between CV predictions (*y*Hat) and observations (*y*) versus values of the bandwidth parameter, θ, by environment and model. K_θ is a marker-based RKHS regression with bandwidth parameter θ; Pedigree & Markers K_θ uses pedigree and markers, here, θ is the value of the bandwidth parameter for markers. Pedigree & Markers $K_{0\cdot25}+K_7$ uses pedigree and markers with KA. E1–E4: environments where the lines were evaluated.

Using data from US Jersey sires ($n=1446$) genotyped with the BovineSNP50 BeadChip (42 552 Single-nucleotide polymorphisms (SNPs)) de los Campos *et al*. (2010) compared the predictive ability of several RKHS models for predicted transmitting abilities of milk production, protein content and daughter pregnancy rate. Models evaluated in that study were: (a) BRR, i.e. $\mathbf{K}=\mathbf{XX}'$; (b) a Gaussian kernel evaluated over a grid of values of the bandwidth parameter, i.e. $\mathbf{K}_\theta$; (c) KA using the two most extreme kernels in the sequence $\{\mathbf{K}_\theta\}$; and (d) a model where $\mathbf{K}$ was a marker-based estimate of a kinship matrix, i.e. $\mathbf{K}=\mathbf{G}$. Results in that study are in agreement with findings reported here in that using KA gave predictive ability similar to that achieved with best performing kernel in the sequence $\{\mathbf{K}_\theta\}$. The comparison between KA, BRR and using $\mathbf{K}=\mathbf{G}$ yielded mixed results: for milk yield all models performed similarly; however, for protein content BRR and $\mathbf{G}$ outperformed KA and the opposite was observed for daughter fertility, illustrating that the optimal choice of kernel may be trait dependent.

4. Concluding remarks

Incorporating molecular markers into models for prediction of genetic values poses important statistical and computational challenges. Ideally, models for dense molecular markers should be: (a) able to cope with the curse of dimensionality; (b) flexible enough to capture the complexity of quantitative traits and (c) amenable for computations. RKHS regressions can be used to address some of these challenges.

Coping with the curse of dimensionality and with complexity. In RKHS, the curse of dimensionality is controlled by defining a notion of smoothness of the

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

unknown function with respect to pairs of points in input space, $\text{Cov}[g(t_i), g(t_{i'})] \propto K(t_i, t_{i'})$. The choice of RK becomes a central element of model specification in RKHS regressions.

As a framework, RKHS is flexible enough to accommodate many non-parametric and some parametric methods, including some classical choices such as the infinitesimal model. The frontier between parametric and non-parametric methods becomes fuzzy; models are better thought as decision rules (i.e. maps from data to estimates) and best evaluated based on performance. Predictive ability appears as a natural choice for evaluating model performance from a breeding perspective.

From a non-parametric perspective, kernels are chosen based on their properties (e.g. predictive ability). To a certain extent, this choice can be made a task of the algorithm. KA offers a computationally convenient method for kernel selection, and results on this study, as well as those of de los Campos *et al.* (2010), suggests that KA is an effective strategy for kernel selection.

Computational considerations. RK Hilbert spaces methods offer enormous computational advantages relative to most of the parametric methods for regression on molecular markers. This occurs for two reasons: (a) the model can be represented in terms of n unknowns and (b) factorizations such as EV or Singular value decompositions can be used to arrive at highly efficient algorithms. Unfortunately, these benefits cannot be exploited in linear models, $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, with marker-specific prior precision variances of effects such as BayesA or Bayesian LASSO. This provides RKHS with a great computational advantage relative to those methods, especially when $p >> n$.

Contribution of marker genotypes to prediction of genetic values. Unlike pedigrees, molecular markers allow tracing Mendelian segregation; potentially, this should allow better predictions of genetic values. Results from this study confirm this expectation. Overall, PM models outperformed P models. Further, most RKHS regression yielded better predictions than those attained with the Bayesian LASSO. However, this did not occur for every RK, indicating that the choice of the kernel is one of the main challenges when applying kernel-based methods. As stated, our results as well as those of de los Campos *et al.* (2010) suggest that KA provides an effective way of choosing a kernel.

Future challenges. In the kernels used in this study all SNPs contributed equally to the RK. As the number of available markers increases, a high number is expected to be located in regions of the genome that are not associated with genetic variability of a quantitative trait. Ideally, the RK should weight each marker based on some measure of its contribution to genetic variance. In linear models such as the Bayesian LASSO, or methods Bayes A or Bayes B, the prior variances of marker effects, which are marker specific, act as weights assigned to each of the markers (e.g. de los Campos *et al.*, 2009*b*).

In RKHS models, one could think of kernels where the contribution of each marker to the kernel is weighted according to some measure of its contribution to genetic variance. For example, one could derive weighted estimates of kinship in which each marker obtains a differential contribution. Alternatively, with a Gaussian kernel, one could think of attaching a bandwidth parameter to each marker. For example, one could use $K(i, i') = \exp\{-\sum_{k=1}^{p} \theta_k d(x_{ik}, x_{i'k})\}$, where θ_k and $d(x_{ik}, x_{i'k})$ are a bandwidth parameter and a distance function associated with the kth marker.

An approach similar to that above-described was evaluated by Long *et al.* (2010) who used radial-basis functions evaluated on principal components (as opposed to individual markers) derived from marker genotypes. Results of that study indicate that the use of input-specific bandwidth parameters may improve predictive ability relative to a model based on a single bandwidth parameter. However, inferring these weights (or bandwidth parameters) poses several statistical challenges when $p >> n$. This occurs because the kernel must be re-computed every time the bandwidth parameters are updated. A natural alternative is to use two-step procedures, with a first step in which an approximation to the weights (or bandwidth parameters) is employed (e.g. with some form of single-marker regression) and a second step where genetic values are inferred. Irrespective of whether single or two-step approaches are used, the development and evaluation of algorithms for computing weighted kernels seem to constitute a central area of research for the application of RKHS to genomic models.

APPENDIX

1. Gibbs sampler

The Appendix describes a Gibbs sampler for a Bayesian RKHS regression. The parameterization is as in equation (5), extended to two random effects and with the inclusion of an intercept. Extension of the model to more than two random effects is straightforward. The derivation of the fully conditional distributions presented here uses standard results for Bayesian linear models (e.g. Gelman *et al.*, 2004; Sorensen & Gianola, 2002).

Let $\mathbf{K}_1 = \boldsymbol{\Lambda}_1 \boldsymbol{\Psi}_1 \boldsymbol{\Lambda}_1'$ and $\mathbf{K}_2 = \boldsymbol{\Lambda}_2 \boldsymbol{\Psi}_2 \boldsymbol{\Lambda}_2'$ be the EV decompositions of the two kernel matrices. Extending (5) to two random effects and by including an intercept, the data equation and likelihood function become $\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2 + \boldsymbol{\varepsilon}$ and $p(\mathbf{y}|\mu, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \sigma^2_\varepsilon) = \text{N}(\mathbf{y}|\mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2, \ \mathbf{I}\sigma^2_\varepsilon)$, respectively. The joint

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

prior is (upon assuming a flat prior for μ)

$$p(\mu, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \sigma^2_\varepsilon, \sigma^2_{g_1}, \sigma^2_{g_2}) \propto N(\boldsymbol{\delta}_1|\mathbf{0}, \boldsymbol{\Psi}_1\sigma^2_{g_1})N(\boldsymbol{\delta}_2|\mathbf{0}, \boldsymbol{\Psi}_2\sigma^2_{g_2}) \times \chi^{-2}(\sigma^2_\varepsilon|df_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_{g_1}|df_{g_1}, S_{g_1})\chi^{-2}(\sigma^2_{g_2}|df_{g_2}, S_{g_2}).$$

Above, $\chi^{-2}(.|\mathrm{df}_\cdot, S_\cdot)$ is a scaled inverse chi-square density with degree of freedom $\mathrm{df}_\cdot$ and scale-parameter $S_\cdot$, with the parameterization presented in Gelman *et al.* (2004).

The joint posterior density is proportional to the product of the likelihood and the prior; thus

$$\begin{aligned} p(\mu, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \sigma^2_\varepsilon, \sigma^2_{g_1}, \sigma^2_{g_2}|\mathbf{y}) &\propto N(\mathbf{y}|\mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2, \mathbf{I}\sigma^2_\varepsilon) \\ &\times N(\boldsymbol{\delta}_1|\mathbf{0}, \boldsymbol{\Psi}_1\sigma^2_{g_1})N(\boldsymbol{\delta}_2|\mathbf{0}, \boldsymbol{\Psi}_2\sigma^2_{g_2}) \\ &\times \chi^{-2}(\sigma^2_\varepsilon|\mathrm{df}_\varepsilon, S_\varepsilon)\chi^{-2}(\sigma^2_{g_1}|\mathrm{df}_{g_1}, S_{g_1})\chi^{-2}(\sigma^2_{g_2}|\mathrm{df}_{g_2}, S_{g_2}). \end{aligned}$$

The Gibbs sampler draws samples of the unknowns from their fully conditional distributions, with the conjugate priors chosen, all fully conditionals are known, as described next.

Intercept. Parameter μ enters only in the likelihood; therefore,

$$\begin{aligned} p(\mu|\mathrm{ELSE}) &\propto N\left(\mathbf{y}|\mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2, \mathbf{I}\sigma^2_\varepsilon\right) \\ &\propto N\left(\mathbf{y}^\mu|\mathbf{1}\mu, \mathbf{I}\sigma^2_\varepsilon\right), \end{aligned}$$

where $\mathbf{y}^\mu = \mathbf{y} - \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 - \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2$, and ELSE denotes all other unknowns except for μ. The fully conditional distribution is then normal with mean $n^{-1}\sum_i y_i^\mu$ and variance $n^{-1}\sigma^2_\varepsilon$.

Regression coefficients. The fully conditional distribution of $\boldsymbol{\delta}_1$ is

$$\begin{aligned} p(\boldsymbol{\delta}_1|\mathrm{ELSE}) &\propto N(\mathbf{y}|\mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2, \mathbf{I}\sigma^2_\varepsilon) \\ &\quad \times N(\boldsymbol{\delta}_1|\mathbf{0}, \boldsymbol{\Psi}_1\sigma^2_{g_1}) \\ &\propto N(\mathbf{y}^{\delta_1}|\boldsymbol{\Lambda}_1\boldsymbol{\delta}_1, \mathbf{I}\sigma^2_\varepsilon)N(\boldsymbol{\delta}_1|\mathbf{0}, \boldsymbol{\Psi}_1\sigma^2_{g_1}), \end{aligned}$$

where $\mathbf{y}^{\delta_1} = \mathbf{y} - \mathbf{1}\mu - \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2$. This is known to be a multivariate normal distribution with mean (covariance matrix) equal to the solution (inverse of the matrix of coefficients) of the following system of equations: $[\boldsymbol{\Lambda}'_1\boldsymbol{\Lambda}_1\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_1^{-1}\sigma_{g_1}^{-2}]\hat{\boldsymbol{\delta}}_1 = \boldsymbol{\Lambda}'_1\mathbf{y}^{\delta_1}\sigma_\varepsilon^{-2}$. Using $\boldsymbol{\Lambda}'_1\boldsymbol{\Lambda}_1 = \mathbf{I}$, the system becomes $[\mathbf{I}\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_1^{-1}\sigma_{g_1}^{-2}]\hat{\boldsymbol{\delta}}_1 = \boldsymbol{\Lambda}'_1\mathbf{y}^{\delta_1}\sigma_\varepsilon^{-2}$. Since $\boldsymbol{\Psi}$ is diagonal, so is the matrix of coefficients of the above system, implying that the elements of $\boldsymbol{\delta}_1$ are conditionally independent. Moreover, $p(\delta_{1j}|\mathrm{ELSE})$ is normal, centred at $[1 + \sigma^2_\varepsilon\sigma_{g_1}^{-2}\Psi_{1j}^{-1}]^{-1}y_{\cdot j}^{\delta_1}$ and with variance $\sigma^2_\varepsilon[1 + \sigma^2_\varepsilon\sigma_{g_1}^{-2}\Psi_{1j}^{-1}]^{-1}$, where $y_{\cdot j}^{\delta_1} = \boldsymbol{\lambda}'_{1j}\mathbf{y}^{\delta_1}$. Here, $\boldsymbol{\lambda}_{1j}$ is the jth column (eigenvector) of $\boldsymbol{\Lambda}_1$.

By symmetry, the fully conditional distribution of $\boldsymbol{\delta}_2$ is also multivariate normal and the associated system of equations is $[\mathbf{I}\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_2^{-1}\sigma_{g_2}^{-2}]\hat{\boldsymbol{\delta}}_2 = \boldsymbol{\Lambda}'_2\mathbf{y}^{\delta_2}\sigma_\varepsilon^{-2}$, where $\mathbf{y}^{\delta_2} = \mathbf{y} - \mathbf{1}\mu - \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1$.

Variance parameters. The fully conditional distribution of the residual variance is

$$\begin{aligned} p\left(\sigma^2_\varepsilon|\mathbf{y}\right) &\propto N\left(\mathbf{y}|\ \mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2\ , \mathbf{I}\sigma^2_\varepsilon\right)\chi^{-2}\left(\sigma^2_\varepsilon|\mathrm{df}_\varepsilon, S_\varepsilon\right) \\ &\propto N\left(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon\right)\chi^{-2}\left(\sigma^2_\varepsilon|\mathrm{df}_\varepsilon, S_\varepsilon\right), \end{aligned}$$

where $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{1}\mu - \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 - \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2$. The above is a scaled inverse chi-square distribution with $\mathrm{df} = n + \mathrm{df}_\varepsilon$ and scale parameter $S = (\sum_i \varepsilon_i^2 + \mathrm{df}_\varepsilon S_\varepsilon)/(n + \mathrm{df}_\varepsilon)$.

The fully conditional distribution of $\sigma^2_{g_1}$ is $p(\sigma^2_{g_1}|\mathrm{ELSE}) \propto N(\boldsymbol{\delta}_1|\mathbf{0}, \boldsymbol{\Psi}_1\sigma^2_{g_1})\chi^{-2}(\sigma^2_{g_1}|\mathrm{df}_{g_1}, S_{g_1})$, which is a scaled inverse chi-square distribution with $\mathrm{df} = n + \mathrm{df}_{g_1}$ and scale parameter $S = (\sum_i \Psi_{1j}^{-1}\delta_{1j}^2 + \mathrm{df}_{g_1}S_{g_1})/(n + \mathrm{df}_g)$. Here, Ψ_{1j} is the jth EV of $\mathbf{K}_1$. Similarly, the fully conditional distribution of $\sigma^2_{g_2}$ is scaled inverse chi-square with $\mathrm{df} = n + \mathrm{df}_{g_2}$ and scale parameter $S = (\sum_j \Psi_{2j}^{-1}\delta_{2j}^2 + \mathrm{df}_{g_2}S_{g_2})/(n + \mathrm{df}_{g_2})$.

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

2. Tables and Figures

Table A1. *Posterior mean (SD) of residual variance by model and environment*

	Models using Pedigree or Markers				Models using Pedigree and Markers			
	E1	E2	E3	E4	E1	E2	E3	E4
Pedigree model	0·562 (0·057)	0·580 (0·056)	0·493 (0·058)	0·519 (0·055)			NA	
$\mathbf{K}_{0\cdot10}$	0·520 (0·049)	0·561 (0·049)	0·646 (0·056)	0·579 (0·052)	0·410 (0·049)	0·485 (0·052)	0·459 (0·056)	0·451 (0·051)
$\mathbf{K}_{0\cdot25}$	0·484 (0·048)	0·545 (0·051)	0·618 (0·057)	0·548 (0·053)	0·386 (0·049)	0·469 (0·051)	0·446 (0·055)	0·437 (0·051)
$\mathbf{K}_{0\cdot50}$	0·432 (0·048)	0·524 (0·051)	0·565 (0·061)	0·502 (0·053)	0·347 (0·048)	0·458 (0·051)	0·428 (0·055)	0·414 (0·051)
$\mathbf{K}_{0\cdot75}$	0·401 (0·048)	0·507 (0·051)	0·520 (0·062)	0·467 (0·052)	0·318 (0·047)	0·442 (0·052)	0·408 (0·055)	0·397 (0·050)
$\mathbf{K}_{1\cdot00}$	0·373 (0·047)	0·490 (0·052)	0·486 (0·062)	0·440 (0·052)	0·294 (0·048)	0·431 (0·053)	0·392 (0·056)	0·379 (0·050)
$\mathbf{K}_{2\cdot00}$	0·313 (0·044)	0·436 (0·053)	0·379 (0·060)	0·373 (0·050)	0·232 (0·043)	0·392 (0·053)	0·327 (0·056)	0·330 (0·048)
$\mathbf{K}_{3\cdot00}$	0·277 (0·043)	0·399 (0·054)	0·320 (0·056)	0·333 (0·047)	0·199 (0·042)	0·364 (0·056)	0·284 (0·053)	0·300 (0·047)
$\mathbf{K}_{5\cdot00}$	0·238 (0·041)	0·347 (0·056)	0·262 (0·051)	0·286 (0·048)	0·155 (0·039)	0·335 (0·060)	0·246 (0·054)	0·269 (0·050)
$\mathbf{K}_{7\cdot00}$	0·214 (0·042)	0·323 (0·060)	0·232 (0·052)	0·255 (0·050)	0·136 (0·037)	0·332 (0·067)	0·238 (0·059)	0·255 (0·053)
$\mathbf{K}_{10\cdot00}$	0·203 (0·044)	0·309 (0·070)	0·218 (0·057)	0·226 (0·055)	0·121 (0·037)	0·333 (0·075)	0·240 (0·064)	0·261 (0·059)
$\mathbf{K}_{0\cdot25}+\mathbf{K}_{7\cdot00}$	0·244 (0·044)	0·402 (0·059)	0·276 (0·060)	0·314 (0·055)	0·152 (0·040)	0·337 (0·058)	0·243 (0·056)	0·276 (0·052)
Bayesian LASSO	0·532 (0·045)	0·555 (0·047)	0·644 (0·050)	0·582 (0·048)	0·370 (0·044)	0·446 (0·047)	0·427 (0·049)	0·419 (0·045)

E1–E4 are the four environments where wheat lines were evaluated; $\mathbf{K}_\theta$ are (Bayesian) RKHS models using a Gaussian kernel evaluated at marker-genotypes with bandwidth parameter θ; $\mathbf{K}_{0\cdot25}+\mathbf{K}_7$ is a model that includes two Gaussian kernels differing only in the value of θ.

Table A2. *MSE between realized phenotypes and CV predictions, by model and environment*

	Models using Pedigree or Markers				Models using Pedigree and Markers			
	E1	E2	E3	E4	E1	E2	E3	E4
Pedigree model	0·826	0·835	0·834	0·812			NA	
$\mathbf{K}_{0\cdot10}$	0·736	0·779	0·853	0·780	0·721	0·773	0·808	0·755
$\mathbf{K}_{0\cdot25}$	0·722	0·778	0·847	0·768	0·708	0·772	0·806	0·750
$\mathbf{K}_{0\cdot50}$	0·703	0·776	0·838	0·754	0·694	0·769	0·801	0·742
$\mathbf{K}_{0\cdot75}$	0·691	0·775	0·830	0·744	0·682	0·769	0·797	0·734
$\mathbf{K}_{1\cdot00}$	0·681	0·775	0·823	0·735	0·674	0·768	0·793	0·730
$\mathbf{K}_{2\cdot00}$	0·664	0·778	0·804	0·721	0·655	0·771	0·781	0·719
$\mathbf{K}_{3\cdot00}$	0·665	0·785	0·796	0·719	0·651	0·775	0·776	0·720
$\mathbf{K}_{5\cdot00}$	0·683	0·809	0·803	0·736	0·660	0·792	0·781	0·733
$\mathbf{K}_{7\cdot00}$	0·713	0·842	0·827	0·763	0·679	0·806	0·792	0·750
$\mathbf{K}_{10\cdot00}$	0·759	0·892	0·870	0·811	0·704	0·818	0·809	0·770
$\mathbf{K}_{0\cdot25}+\mathbf{K}_{7\cdot00}$	0·679	0·768	0·801	0·729	0·658	0·762	0·775	0·723
Bayesian LASSO	0·748	0·783	0·861	0·787	0·732	0·781	0·814	0·759

E1–E4 are the four environments where wheat lines were evaluated; $\mathbf{K}_\theta$ are (Bayesian) RKHS models using a Gaussian kernel evaluated at marker genotypes with bandwidth parameter θ; $\mathbf{K}_{0\cdot25}+\mathbf{K}_{7\cdot00}$ is a model that includes two Gaussian kernels differing only in the value of θ.

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

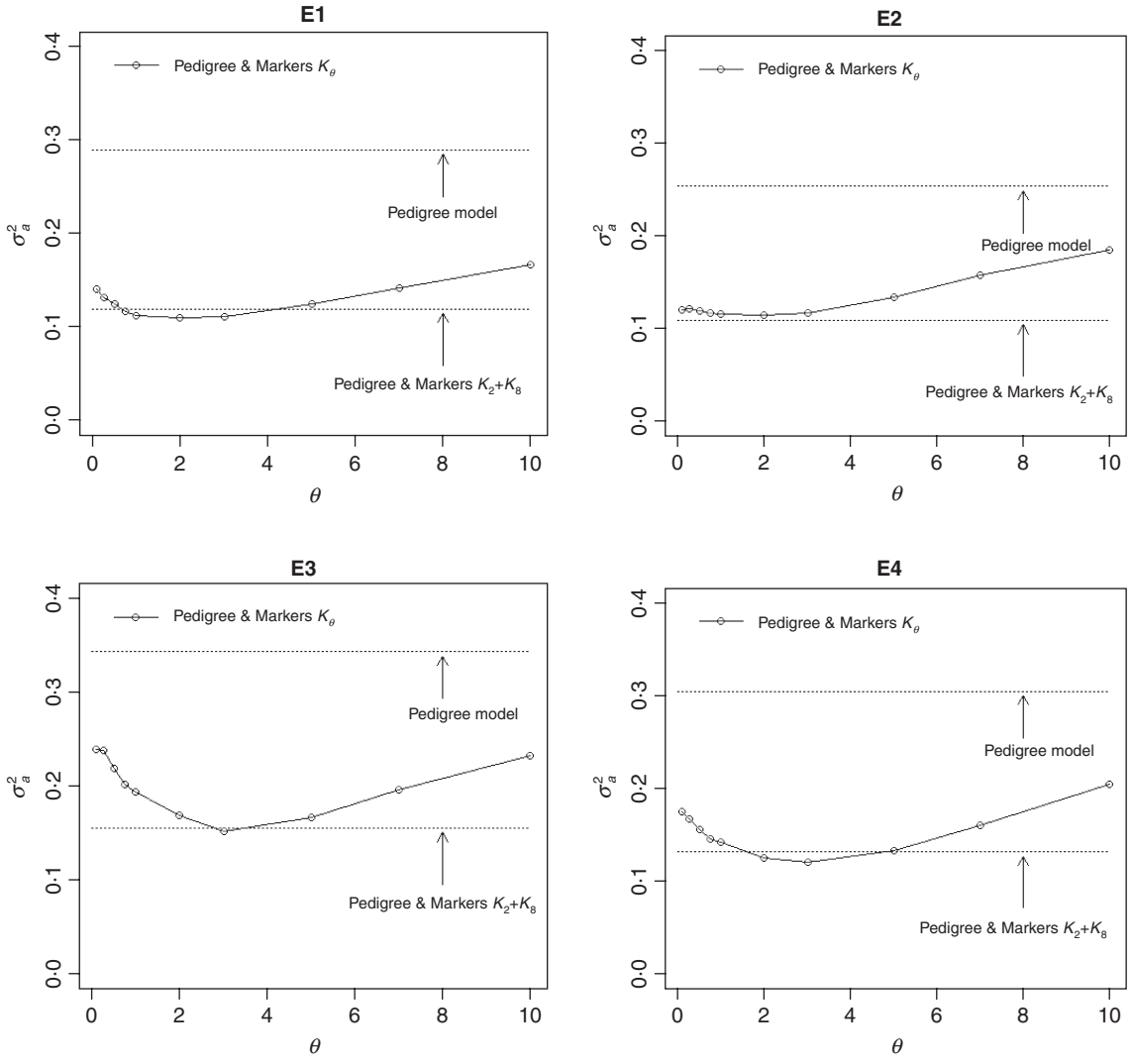

Fig. A1. Posterior mean of the variance of the regression on the pedigree, σ_a^2, versus values of the bandwidth parameter, θ, by environment and model. Pedigree & Markers K_θ uses pedigree and markers, here, θ is the value of the bandwidth parameter for markers. Pedigree & Markers $K_{0\cdot25}+K_7$ uses pedigree and markers with KA. E1–E4: environments where the lines were evaluated.

The authors thank Vivi Arief from the School of Land Crop and Food Sciences of the University of Queensland, Australia, for assembling the historical wheat phenotypic and molecular marker data and for computing the additive relationships between the wheat lines. We acknowledge valuable comments from Grace Wahba, David B. Allsion, Martin Schlather, Emilio Porcu and two anonymous reviewers. Financial support by the Wisconsin Agriculture Experiment Station; grant DMS-NSF DMS-044371 is acknowledged.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**, 337–404.

Bernardo, R. & Yu, J. (2007). Prospects for genome-wide selection for quantitative traits in maize. *Crop Science* **47**, 1082–1090.

Calus, M. P. L. & Veerkamp, R. F. (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* **124**, 362–388.

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

Corrada Bravo, H., Lee, K. E., Klein, B. E. K., Klein, R., Iyengar, S. K. & Wahba, G. (2009). Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences, USA* **106**, 8128–8133.

Cressie, N. (1993). *Statistics for Spatial Data*. New York Wiley.

de los Campos, G., Gianola, D. & Rosa, G. J. M. (2009*a*). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* **87**, 1883–1887.

https://doi.org/10.1017/S0016672310000285 Published online by Cambridge University Press

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. & Cotes, J. M. (2009*b*). Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics* **182**, 375–385.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., Vazquez, A. I. & Allison, D. B. (2010). Semi-Parametric Marker-enabled Prediction of Genetic Values using Reproducing Kernel Hilbert Spaces methods. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, in press.

Eding, J. H. & Meuwissen, T. H. E. (2001). Marker based estimates of between and within population kinships for the conservation of genetic diversity. *Journal of Animal Breeding and Genetics* **118**, 141–159.

Fernando, R. L. & Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21**, 467–477.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis*. London, UK: Chapman and Hall.

Gianola, D. & de los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genetics Research* **90**, 525–540.

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

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

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–363.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations* 3rd ed. Baltimor and London: The Johns Hopkins University Press.

Habier, D., Fernando, R. L. & Dekkers, J. C. M. (2007). The impact of genetic relationships information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.

Harville, D. A. (1983). Discussion on a section on interpolation and estimation. In David, H. A. & David, H. T. (ed.). *Statistics an Appraisal*, pp. 281–286. Ames, IA: The Iowa State University Press.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning* (*Data Mining, Inference, and Prediction*) 2nd edition. New York, NY: Springer.

Hayes, B. J. & Goddard, M. E. (2008). Prediction of breeding values using marker-derived relationship matrices. *Journal of Animal Science* **86**, 2089–2092.

Heffner, E. L., Sorrells, M. E. & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

Hoerl, A. E. & Kennard, R. W. (1970*a*). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–67.

Hoerl, A. E. & Kennard, R. W. (1970*b*). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 69–82.

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

Kimeldorf, G. S. & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic process and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

Kimeldorf, G. S. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematic Analysis and Applications* **33**, 82–95.

Kondor, R. I. & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete inputs. *Proceedings of 19th International Conference on Machine Learning* (*ICML*-2002).

Long, N., Gianola, D., Rosa, G., Weigel, K., Kranis, A. & González-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research*, in press.

Lynch, M. & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766.

Mallick, B., Ghosh, D. & Ghosh, M. (2005). Bayesian kernel-based classification of microarray data. *Journal of the Royal Statistical Society, Series B* **2**, 219–234.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

McLaren, C. G., Bruskiewich, R., Portugal, A. M. & Cosico, A. B. (2005). The international Rice information system. A platform for meta-analysis of rice crop data. *Plant Physiology* **139**, 637–642.

Park, T. & Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* **103**, 681–686.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research* **67**, 175–186.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.

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

Speed, T. (1991). [That BLUP is a good thing: the estimation of random effects]: Comment. *Statistical Science* **6**, 42–44.

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

Van Raden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bulletin* **37**, 33–36.

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

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

Wright, S. (1921). Systems of mating. I. The biometric relations between parents and offspring. *Genetics* **6**, 111–123.