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

Gustavo de los Campos[1,2,3], Daniel Gianola[1], Guilherme J.M. Rosa[1], Kent A. Weigel[1], and José Crossa[2]

Corresponding author: Gustavo de los Campos, gdeloscampos@ms.soph.uab.edu

Profs to be sent to: Gustavo de los Campos, 1665 University Boulevard

Ryals Public Health Building 414, AL 35294, US.

Running headlines: Kernel Methods for Genomic Selection

1 University of Wisconsin-Madison, 1675 Observatory Dr. WI 53706, US.
2 International Maize and Wheat Improvement Center (CIMMYT), Ap. Postal 6-641, 06600, México DF, México.
3 Now at University of Alabama-Birmingham, 1665 University Boulevard, AL 35294, US.

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. Nowadays, dense molecular markers are available in the genome of humans plants and animals, and this information can be use to enhance 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 'automatic' kernel selection, and an assessment of the proposed methods using a collection of 599 wheat lines evaluated for grain yield in four mega-environments.

KEY WORDS: Complex Traits; Bayesian Methods; Genomics; RKHS.

1. Introduction

Prediction of genetic values is relevant in plant and animal breeding, or for assessing probability of disease in medicine. Standard genetic models view phenotypic outcomes $(y_i; i = 1,...,n)$ 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 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 $Cov(g_i, g_{i'})$, and this provides a means for smoothing phenotypic records.

Dense molecular marker panels (MM) are now available in humans and in many plant an 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 MM 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 and 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.

In parametric regression models for MM (e.g., Meuwissen, Hayes and Goddard, 2001), g_i is a parametric regression on marker covariates, x_{ij}. The linear model takes the form:

$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, where β_j is the regression of y_i on x_{ij}. Often, $p>>n$ and some shrinkage estimation method such as ridge regression (Hoerl and Kennard, 1970a, 1970b) or LASSO (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 and Casella (2008) or methods BayesA or BayesB of Meuwissen, Hayes and Goddard (2001) are the most commonly used. 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 modeling interactions is only feasible to a limited degree.

Reproducing Kernel Hilbert spaces regressions have been proposed for semi-parametric regression on marker genotypes, e.g., Gianola, Fernando and Stella (2006); Gianola and van Kaam (2008). In RKHS, markers are used to build a covariance structure among genetic values; for example, $Cov(g_i, g_{i'}) \propto K(\mathbf{x}_i, \mathbf{x}_{i'})$ where $\mathbf{x}_i$, $\mathbf{x}_{i'}$ are vectors of marker genotypes and $K(.,.)$, the Reproducing Kernel (RK), is some positive definite function. This approach has several attractive features: (a) the methodology can be used with almost any type of information set (e.g., covariates, strings, images, graphs). This is particularly important because techniques for characterizing genomes change rapidly; (b) some parametric methods for GS appear as special cases; and (c) computations are performed in a *n*-dimensional space. This gives to RKHS methods 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. Genetic models (e.g., additive infinitesimal) can be used to choose the kernel and we discuss the connection between

RKHS regressions and some of the standard models of quantitative genetics. We also discuss the case where the reproducing kernel is chosen based on its properties (e.g., predictive ability) and how the problem of kernel choice be formulated as an estimation problem. Section 3 presents an application to an extensive plant breeding data where some of the methods discussed in Section 2 are evaluated. Concluding remarks are provided in Section 4.

2. Reproducing Kernel Hilbert Spaces Regression

This section provides a brief overview of kernel methods (*i*), introduce a parameterization of RKHS regressions that yields highly efficient computational implementations (*ii*), and discusses the problem of kernel choice (*iii*); a central problem in RKHS regressions.

(i) Overview of reproducing kernel Hilbert spaces regressions

Reproducing kernel Hilbert spaces 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. In all these applications the learning task is the following (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, *f*. Inferring f requires defining a collection of functions (hereinafter denoted as $f \in H$, standing for all functions in RKHS of real-valued functions *H*) from which an element, $\hat{f}$,will be chosen and a criterion (e.g., a penalized residual sum of squares or a posterior density) for comparing functions in *H*. In RKHS estimates are obtained by solving the following optimization problem:

$$\hat{f} = \arg\min_{f \in H} \left\{ l(f, \mathbf{y}) + \lambda \|f\|_H^2 \right\}, \qquad [1]$$

where $l(f, \mathbf{y})$ 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 $\|f\|_H^2$ is the square of the norm of f on H; a measure of model complexity.

Each RKHS is uniquely associated to a positive definite (PD) function[4] and this is known as the Moore-Aronszajn theorem (Aronszajn, 1950). Therefore, choosing $K(t_i, t_j)$ amounts to select H. Using this duality, Kimeldorf and Wahba (1971) showed that the finite dimensional solution of [1] admits a linear representation $f(t_i) = f_i = \sum_j K(t_i, t_j)\alpha_j$, or in matrix notation, $\hat{\mathbf{f}} = \mathbf{K}\hat{\boldsymbol{\alpha}} = [\hat{f}_1, ..., \hat{f}_n]'$. Further, in this finite dimensional setting, $\lambda \|f\|_H^2 = \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$, where $\mathbf{K} = \{K(t_i, t_j)\}$. Using this in [1] and setting $l(f, \mathbf{y})$ to be a residual sum of squares, one obtains: $\hat{\mathbf{f}} = \mathbf{K}\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)'$ is the solution of :

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} \right\}, \qquad [2]$$

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 $\hat{\mathbf{f}} = \mathbf{K}\hat{\boldsymbol{\alpha}} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} = \mathbf{P}(\lambda, K)\mathbf{y}$, 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 in the objective function and on the solution only through $\mathbf{K}$. This allows using RKHS for regression with any class of information sets (vectors,

[4] That is, any function, $K(t_i, t_j)$, satisfying $\sum_i \sum_j \alpha_i \alpha_j K(t_i, t_j) > 0$ for all sequences, $\{\alpha_i\}$, with $\alpha_i \neq 0$ for some *i*.

graphs, images, etc.) where a PD function can be evaluated; the choice of kernel becomes the key element of model specification.

From a Bayesian perspective, $\hat{\boldsymbol{\alpha}}$ can be viewed as a posterior mode in the following model: $\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$; $p\left(\boldsymbol{\varepsilon}, \boldsymbol{\alpha} | \sigma_\varepsilon^2, \sigma_g^2\right) = N\left(\boldsymbol{\varepsilon} | \mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right) N\left(\boldsymbol{\alpha} | \mathbf{0}, \mathbf{K}^{-1}\sigma_g^2\right)$. The relationship between RKHS regressions and Gaussian processes was first noted by Kimeldorf and Wahba (1970) and has re-visited by many authors, e.g., Harville, 1983; Speed, 1991. The Bayesian view of RKHS spaces gives a connection to Gaussian processes. Following de los Campos *et al.* (2009a), one can change variables in the above model, with $\mathbf{g} = \mathbf{K}\boldsymbol{\alpha}$, yielding:

$$\begin{cases} \mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon} \\ p\left(\boldsymbol{\varepsilon}, \mathbf{g} | \sigma_\varepsilon^2, \sigma_g^2\right) = N\left(\boldsymbol{\varepsilon} | \mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right) N\left(\mathbf{g} | \mathbf{0}, \mathbf{K}\sigma_g^2\right) \end{cases} \qquad [3]$$

Thus, from a Bayesian perspective, the evaluations of functions, can be viewed as Gaussian processes satisfying $Cov\left(g_i, g_j\right) \propto K\left(t_i, t_j\right)$. The fully-Bayesian RKHS regression assumes unknown variance parameters, and the model becomes:

$$\begin{cases} \mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon} \\ p\left(\boldsymbol{\varepsilon}, \mathbf{g}, \sigma_\varepsilon^2, \sigma_g^2\right) = N\left(\boldsymbol{\varepsilon} | \mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right) N\left(\mathbf{g} | \mathbf{0}, \mathbf{K}\sigma_g^2\right) p\left(\sigma_\varepsilon^2, \sigma_g^2\right) \end{cases} \qquad [4]$$

where $p\left(\sigma_\varepsilon^2, \sigma_g^2\right)$ is a (proper) prior density assigned to variance parameters.

(ii) Representation using orthogonal random variables

Representing model [4] with orthogonal random variables simplifies computations greatly and provides additional insights on the nature of the RKHS regressions. To this end, we make use of the eigenvalue decomposition (e.g., Golub and Van Loan, 1996) of the kernel matrix $\mathbf{K} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}$ is a matrix of eigenvectors satisfying $\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \mathbf{I}$ and $\boldsymbol{\Psi} = Diag\left\{\Psi_j\right\}$, $\Psi_1 \geq \Psi_2 \geq ... \geq \Psi_n > 0$, is a diagonal matrix whose non-zero entries are the

eigenvalues (EV) of **K**. Using these [4] becomes:

$$\begin{cases} \mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \\ p\left(\boldsymbol{\varepsilon}, \boldsymbol{\delta}, \sigma_\varepsilon^2, \sigma_g^2\right) \propto N\left(\boldsymbol{\varepsilon} \middle| \mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right) N\left(\boldsymbol{\delta} \middle| \mathbf{0}, \boldsymbol{\Psi}\sigma_g^2\right) p\left(\sigma_\varepsilon^2, \sigma_g^2\right) \end{cases} \qquad [5]$$

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 $Cov(\boldsymbol{\Lambda}\boldsymbol{\delta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'\sigma_g^2 = \mathbf{K}\sigma_g^2$. Therefore, [4] and [5] are two parameterizations of the same probability model. However, [5] is much more computationally convenient.

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, $p\left(\boldsymbol{\delta} \middle| \mathbf{y}, \mu, \sigma_\varepsilon^2, \sigma_g^2\right)$, is usually the most computationally demanding step. From standard results of Bayesian linear models, one can show that $p\left(\boldsymbol{\delta} \middle| ELSE\right) = N\left(\hat{\boldsymbol{\delta}}, \sigma_\varepsilon^2 \mathbf{C}^{-1}\right)$, where: $\mathbf{C} = \left[\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \sigma_\varepsilon^2\sigma_g^{-2}\boldsymbol{\Psi}^{-1}\right] = Diag\left\{1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_j^{-1}\right\}$, and $\hat{\boldsymbol{\delta}} = \mathbf{C}^{-1}\boldsymbol{\Lambda}'\mathbf{y}$. This simplification occurs because $\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \mathbf{I}$. The fully conditional distribution of $\boldsymbol{\delta}$ is multivariate normal, and the (co)variance matrix, $\sigma_\varepsilon^2\mathbf{C}^{-1}$, is diagonal, therefore: $p\left(\boldsymbol{\delta} \middle| ELSE\right) = \prod_{j=1}^{n} p\left(\delta_j \middle| ELSE\right)$. Moreover, $p\left(\delta_j \middle| ELSE\right)$ is normal, centered at $\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_j^{-1}\right]^{-1} y_{\cdot j}$ and with variance $\sigma_\varepsilon^2\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_j^{-1}\right]^{-1}$. Here, $y_{\cdot j} = \boldsymbol{\lambda}_i'\mathbf{y}$, where $\boldsymbol{\lambda}_i'$ is the transpose of the i^{th} column (eigenvector) of $\boldsymbol{\Lambda}$. Note that model unknowns are not required for computing $y_{\cdot j}$, implying that these quantities remain constant across iterations of a sampler. The only quantities that need to be updated are $\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_j^{-1}\right]$ and $\sigma_\varepsilon^2\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_j^{-1}\right]$. 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 $p\left(\boldsymbol{\delta} \middle| ELSE\right)$ will need to be updated at each iteration of the

sampler; however, the matrix of coefficient remains diagonal and this simplifies computations greatly (see Appendix).

In [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 EVs are usually sorted such that $\Psi_1 \geq \Psi_2 \geq ... \geq \Psi_n > 0$. The prior precision variance of regression coefficients is proportional to the EVs, that is, $Var(\delta_j) \propto \Psi_j$. Therefore, the extent of shrinkage increases as *j* does. For most RKs, the decay of the EVs will be such that for the first EVs $\left[1 + \sigma_\varepsilon^2 \sigma_g^{-2} \Psi_j^{-1}\right]$ 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 f that are not penalized.

(iii) Choosing the Reproducing Kernel

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 to predict future observations. 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, Gianola and Rosa, 2009a). 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 infinitesimal model and certain mode of gene action (additive, dominance or diverse forms of epistasis, e.g., Cockerham, 1954; Kempthorne, 1954).

One way of incorporating marker information into models for prediction of genetic values is to use [4] with **K** being a marker-based estimate of a kinship matrix (usually denoted as **G**) and several estimates have been suggested and used in applications (c.f., Ritland, 1996; Lynch and Ritland, 1999; Eding and Meuwissen, 2001; VanRaden, 2007; Hayes and Goddard,

2008). As with the pedigree-based infinitesimal model, here the (co)variance structure is defined so as to represent the type of patterns expected under a particular mode of gene action (e.g., infinitesimal additive model). Therefore, using **K=G** is another way of generating a model-driven RK.

A Bayesian ridge regression (BRR) provides another way of choosing **K** so as to represent the patterns generated by a parametric model. This model is defined by $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ and $p\left(\boldsymbol{\varepsilon},\boldsymbol{\beta},\sigma_\varepsilon^2,\sigma_\beta^2\right)=N\left(\boldsymbol{\varepsilon}|\mathbf{0},\mathbf{I}\sigma_\varepsilon^2\right)N\left(\boldsymbol{\beta}|\mathbf{0},\mathbf{I}\sigma_\beta^2\right)p\left(\sigma_\varepsilon^2,\sigma_g^2\right)$. To see how a BRR constitutes a special case of [5] one can make use of the singular value decomposition (SVD, e.g., Golub and Van Loan, 1996) of $\mathbf{X}=\mathbf{UDV}'$. Here, $\mathbf{U}$ (n×n) and $\mathbf{V}$ (p×n) are matrices whose columns are orthogonal, and $\mathbf{D}=Diag\{\xi_i\}$is a diagonal matrix whose non-null entries are the singular values of $\mathbf{X}$. Using this in the data equation, we get $\mathbf{y}=\mathbf{UDV}'\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{U}\boldsymbol{\delta}+\boldsymbol{\varepsilon}$, 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 $Cov(\boldsymbol{\delta})=\mathbf{DV}'\mathbf{VD}'\sigma_\beta^2=\mathbf{DD}'\sigma_\beta^2$, thus, $\boldsymbol{\delta}\sim N\left[\mathbf{0},Diag\left\{\xi_j^2\right\}\sigma_\beta^2\right]$. Therefore a BRR can be equivalently represented using [5] with $\boldsymbol{\Lambda}=\mathbf{U}$ and $\boldsymbol{\Psi}=Diag\left\{\xi_j^2\right\}$. Once an estimate of $\boldsymbol{\delta}$ has been obtained, estimates of marker effects can be retrieved using $\boldsymbol{\beta}=\mathbf{VD}^{-1}\boldsymbol{\delta}$. Note that using $\boldsymbol{\Lambda}=\mathbf{U}$ and $\boldsymbol{\Psi}=Diag\left\{\xi_j^2\right\}$ in [5] implies $\mathbf{K}=\mathbf{UDD}'\mathbf{U}'=\mathbf{UDV}'\mathbf{VD}'\mathbf{U}'=\mathbf{XX}'$ in [4].

Habier Fernando and Dekkers (2009) argues that as the number of markers increases, $\mathbf{XX}'$ approaches the numerator relationship matrix, **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 as fixed, whereas in BRR, the opposite is true (see, Gianola et al., 2009, for further discussion of this).

In the examples given above the RK was defined in such a manner that it represents a parametric model. An advantage of using parametric models is that estimates can be interpreted in terms of the theory used to derive **K**. For example, if **K=A** then σ_g^2 is interpretable as an additive genetic variance and $\sigma_g^2\left(\sigma_g^2+\sigma_\varepsilon^2\right)^{-1}$ 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 and Cristianini, 2004) views RK's 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 $K\left(t_i,t_{i'}|\theta\right)=\exp\{-\theta d\left(t_i,t_{i'}\right)\}$. Here, $d\left(t_i,t_{i'}\right)$ is some distance function and θ controls how fast the covariance function drops as points get further apart as measured by $d\left(t_i,t_{i'}\right)$. The bandwidth parameter may be chosen by cross-validation (CV) or with Bayesian methods (e.g., Mallick, Ghosh and Ghosh, 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 may be computationally easier to evaluate model performance over a grid of values of θ—this is illustrated in Section 3.

The decay of the EVs controls, to a certain extent, shrinkage of estimates of **δ** and, with this, the trade-offs between goodness of fit and model complexity. Transformations of EVs (indexed with unknown parameters) can also be used to generate a family of kernels. One such example is the diffusion kernel $\mathbf{K}_\alpha=\mathbf{\Lambda}Diag\left\{\exp\left(\alpha\Psi_j\right)\right\}\mathbf{\Lambda}'$ (e.g., Kondor and Lafferty, 2002). Here, $\alpha>0$ is used to control the decay of EVs.

A third way of generating families of kernels is to use closure properties of PD functions (Shawe-Taylor and Cristianini, 2004). For example, linear combinations of PD functions, $\tilde{K}(t_i,t_j)=\sigma^2_{g1}K_1(t_i,t_j)+\sigma^2_{g2}K_2(t_i,t_j)$, with $\sigma^2_{g.}\geq 0$, are PD as well. From a Bayesian perspective, σ^2_{g1} and σ^2_{g2} 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, $p(\mathbf{g}_1,\mathbf{g}_2|\sigma^2_{g1},\sigma^2_{g2})=N(\mathbf{g}_1|\mathbf{0},\mathbf{K}_1\sigma^2_{g1})N(\mathbf{g}_2|\mathbf{0},\mathbf{K}_2\sigma^2_{g2})$. It follows that $\mathbf{g}\sim N(\mathbf{0},\mathbf{K}_1\sigma^2_{\alpha 1}+\mathbf{K}_2\sigma^2_{\alpha 2})$, or equivalently $\mathbf{g}\sim N(\mathbf{0},\tilde{\mathbf{K}}\tilde{\sigma}^2_{\alpha})$, where $\tilde{\sigma}^2_{\alpha}=(\sigma^2_{\alpha 1}+\sigma^2_{\alpha 2})$ and $\tilde{\mathbf{K}}=\mathbf{K}_1\sigma^2_{\alpha 1}\tilde{\sigma}^{-2}_{\alpha}+\mathbf{K}_2\sigma^2_{\alpha 2}\tilde{\sigma}^{-2}_{\alpha}$. Therefore, fitting a RKHS with two random effects is equivalent to using $\tilde{\mathbf{K}}$ in [4]. We refer to this approach as automatic 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 $K_1(t_i,t_j)$ and $K_2(t_i,t_j)$ are PD, so is $K(t_i,t_j)=K_1(t_i,t_j)K_2(t_i,t_j)$, in matrix notation this is usually denoted as $\mathbf{K=K_1\#K_1}$. From a genetic perspective, this formulation can be used to accommodate non-additive infinitesimal effects (e.g., Cockerham, 1954; Kempthorne, 1954). For example, under certain conditions, **K=A#A** gives the expected degree of resemblance between relatives under an infinitesimal model for additive×additive interactions.

3. Application to Plant Breeding Data

Some of the methods discussed in the previous section were evaluated using a dataset consisting of a collection of historical wheat lines from the Global Wheat Breeding Program of CIMMYT (International Maize and Wheat Improvement Center). In plant breeding programs, 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 programs 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 and Yu, 2007; Heffner, Sorrels and Jannink 2009). Thus, assessing the ability of a model to predict future outcomes is central in breeding programs.

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 and 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 1,447 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 (ICIS), as described in McLaren *et al*. (2005).

A sequence of models was fitted to the entire dataset and in cross-validation. 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 $(i = 1,...,599)$ is the phenotype of the i^{th} line; μ is an

effect common to all lines; g_i is the genetic value of the i^{th} 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 $\varepsilon_i \overset{IID}{\sim} N\left(0,\sigma_\varepsilon^2\right)$, where σ_ε^2 is the residual variance. The conditional distribution of the data was $p\left(\mathbf{y}|\mu,\mathbf{g},\sigma_\varepsilon^2\right)=\prod_{i=1}^{n} N\left(y_i|\mu+g_i,\sigma_\varepsilon^2\right)$ where, $\mathbf{g}=\left(g_1,...,g_n\right)'$. Models differed on how g_i was described.

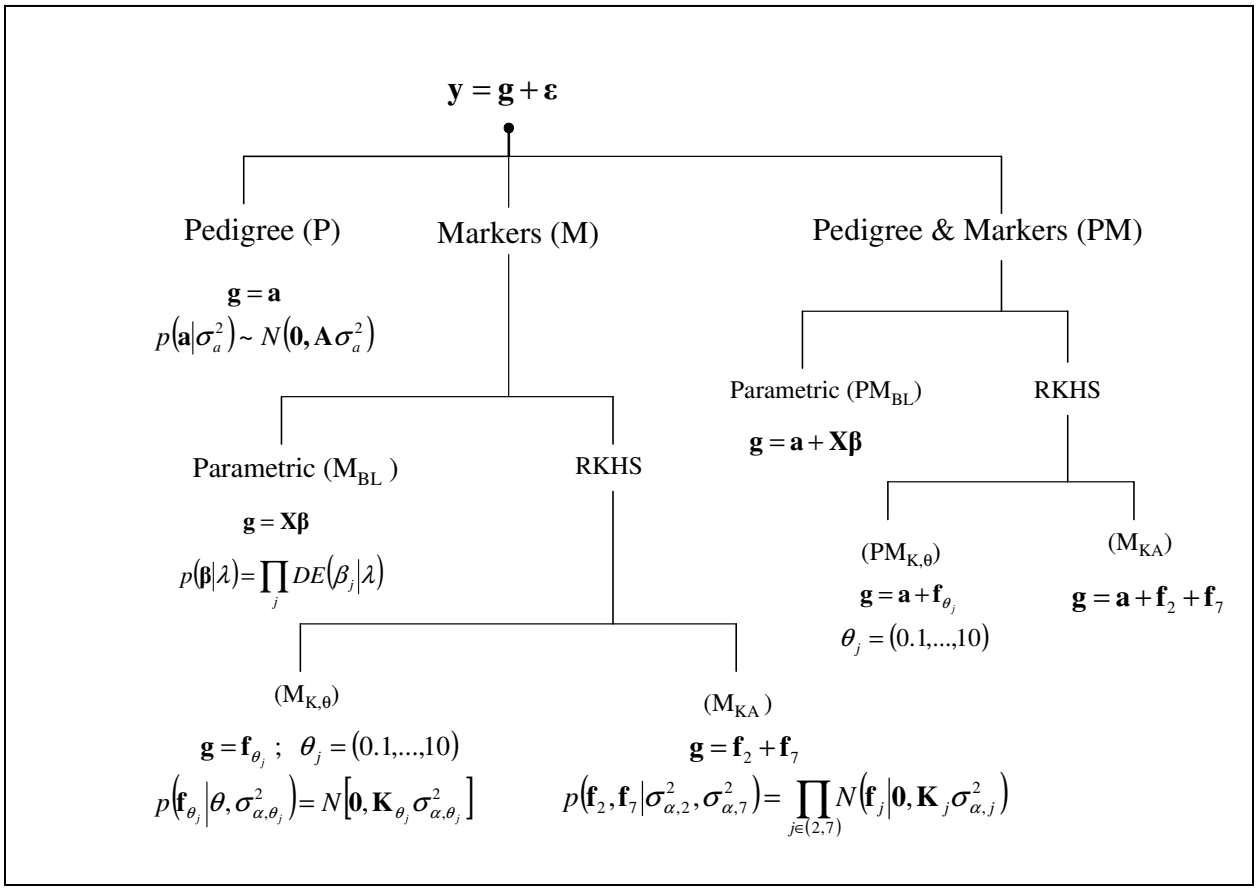

Figure 1. Alternative models for prediction of genetic values. Phenotypic records (**y**) were always the sum of a genetic signal (**g**) and a vector of Gaussian residuals (**ε**). Models differed on how **g** was represented, as described in the figure. BL=Bayesian LASSO; RKHS=Reproducing Kernel Hilbert Spaces regression; λ=LASSO regularization parameter; θ=RKHS bandwidth parameter; $\sigma^2_{\cdot}$ =variance parameter; KA=Kernel Averaging; N(.,.) normal density, DE(.) double-exponential density.

In a standard infinitesimal additive model (P, standing for pedigree-model), genetic values are $\mathbf{g}=\mathbf{a}$ with $\mathrm{p}\left(\mathbf{a}\middle|\sigma_a^2\right)=N\left(\mathbf{0},\mathbf{A}\sigma_a^2\right)$, where σ_a^2 is the additive 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 **K=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 $\mathbf{X}=\{x_{ij}\}$ is an incidence matrix with marker genotypes codes, and $\boldsymbol{\beta}=\left(\beta_1,...,\beta_p\right)'$, the vector of regression coefficients, was inferred using the BL of Park and Casella (2008). This model is denoted as $\mathrm{M_{BL}}$. Following de los Campos *et al*. (2009b), the prior of on the regularization parameter, λ, that controls the prior distribution of marker effects was $Beta\left(\frac{\lambda}{150}\middle|\alpha_3=1.2,\alpha_4=1.2\right)$,which is flat over a fairly wide range.

In marker-based RKHS regressions ($\mathrm{M_K}$) $\mathbf{g}=\mathbf{f}_\theta$, where $\mathbf{f}_\theta=\left(f_{\theta,1},...,f_{\theta,n}\right)'$ was assigned a Gaussian prior with null mean and (co)variance matrix $Cov(\mathbf{f}_\theta)\propto\mathbf{K}_\theta=\left\{\exp\left(-\theta k^{-1}d_{ii'}\right)\right\}$. Here, θ is a bandwidth parameter, $d_{ii'}=\left\|\mathbf{x}_i-\mathbf{x}_{i'}\right\|^2$ is the square Euclidean distance between marker codes $\mathbf{x}_i=\left(x_{i1},...,x_{ip}\right)'$ and $\mathbf{x}_{i'}=\left(x_{i'1},...,x_{i'p}\right)'$, and $k=\max_{(i,i')}\left\{\left\|\mathbf{x}_i-\mathbf{x}_{i'}\right\|^2\right\}$. Models were fitted over a grid of values of $\theta\in\{.1,\ .25,.5,\ .75,1,2,3,5,7,10\}$, and are denoted as $M_{K,\theta}$. A model where $\mathbf{g}$ was the sum of two components: $\mathbf{g}=\mathbf{f}_{0.25}+\mathbf{f}_7$, with $p\left(\mathbf{f}_{0.25},\mathbf{f}_7\middle|\sigma^2_{\alpha,0.25},\sigma^2_{\alpha,7}\right)=N\left(\mathbf{f}_{0.25}\middle|\mathbf{0},\mathbf{K}_{0.25}\sigma^2_{\alpha,0.25}\right)N\left(\mathbf{f}_7\middle|\mathbf{0},\mathbf{K}_7\sigma^2_{\alpha,7}\right)$ was fitted as well. This model is referred to as M_{KA}, standing for marker-based model with "kernel-averaging".

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}; $\mathbf{g} = \mathbf{a} + \mathbf{f}_\theta$, $\theta = \{.1,\ .25, .5,\ .75, 1, 2, 3, 5, 7, 10\}$, denoted as $PM_{K,\theta}$; and, $\mathbf{g} = \mathbf{a} + \mathbf{f}_{0.25} + \mathbf{f}_7$, denoted as PM_{KA}.

In all models variance parameters were treated as unknown and assigned identical independent scaled inverse chi-square prior distributions with small degrees of freedom and scale parameters, $p(\sigma^2_\cdot) = \chi^{-2}(\sigma^2_\cdot | df = 3, S = 1)$. Samples from posterior distributions for each of the models were obtained with a Gibbs sampler (see de los Campos *et al.,* 2009b, 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 2,000 samples as burn-in. The distribution of prediction errors was estimated using a 10-fold CV (e.g., Hastie, Tibshirani and Friedman, 2009), with random assignment of lines to folds.

(ii) Results

Figure 2 shows the posterior means of the residual variance in $M_{K,\theta}$ and $PM_{K,\theta}$ versus values of the bandwidth parameter θ obtained when models were fitted to the entire data. Each panel in Figure 2 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. Overall, models M and PM fitted the data better than P, and PM_{KA} gave almost always better fit than $M_{K,\theta}$ and $PM_{K,\theta}$. 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 smaller for $PM_{K,\theta}$ than for $M_{K,\theta}$. This occurs, because in $PM_{K,\theta}$, 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 $M_{K,0.1}$ and $PM_{K,0.5}$, 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 Figure 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 $PM_{K,\theta}$ 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 conditional expectation function fitted would depend more strongly on the regression on the pedigree (large values of σ_a^2).

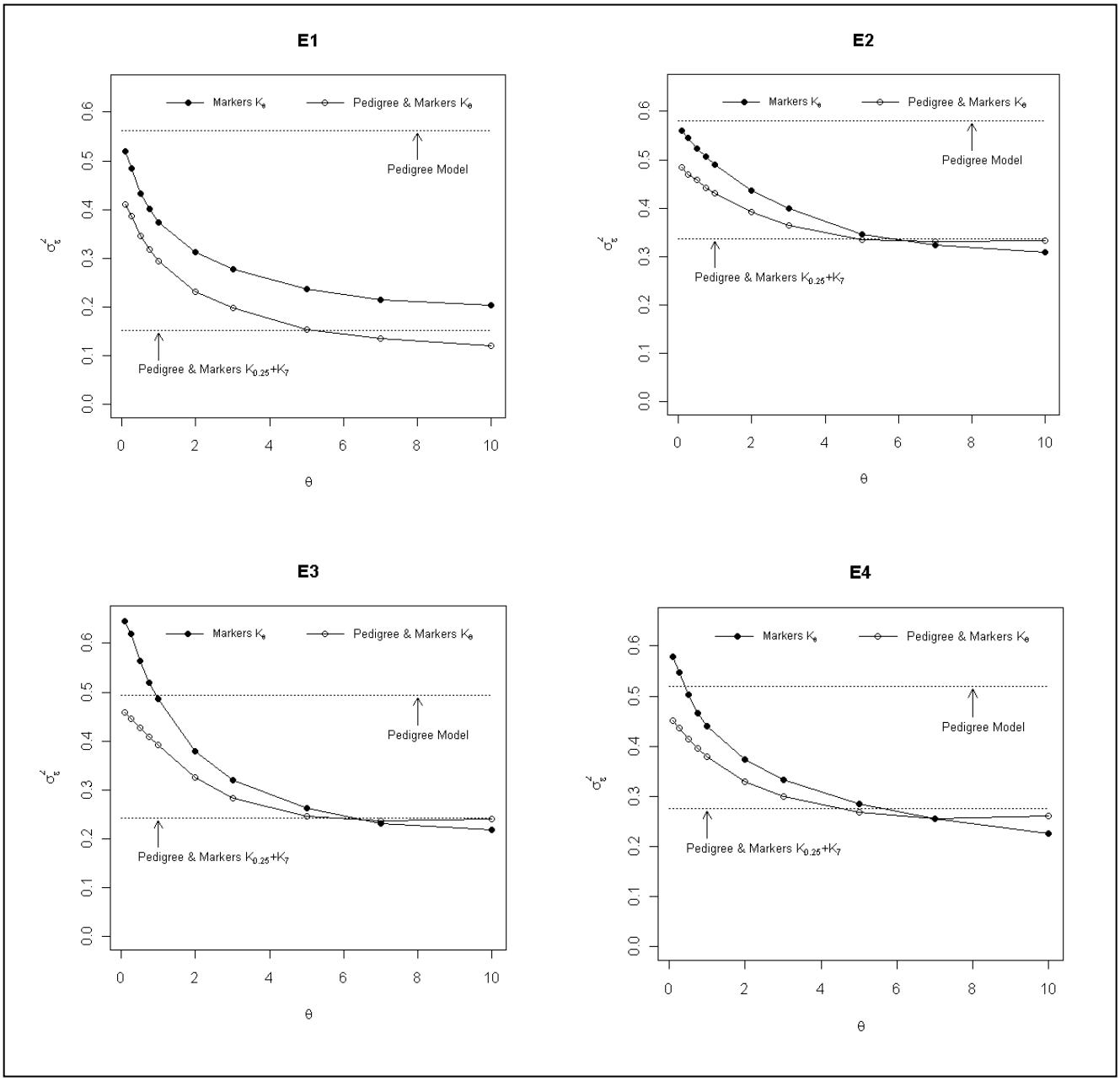

Figure 2. Estimated posterior mean of the residual variance versus values of the bandwidth parameter, θ, by environment and model. K_θ is a marker-based reproducing kernel Hilbert spaces regression (RKHS) with bandwidth parameter θ ; Pedigree & Markers K_θ uses pedigree and markers, here, θ is the value of the bandwidth parameter for markers. Pedigree & Markers $K_{0.25} + K_7$ uses pedigree and markers with kernel averaging. E1-E4: environments where the lines were evaluated.

Plots in Figure 3 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 give 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, as in PM_{KA}. These results are in agreement with some studies (e.g., Corrada Bravo *et al.* 2009; de los Campos *et al.* 2009b) that provided evidence of a gain in predictive ability by jointly considering markers and pedigree information. However, not all $PM_{K,\theta}$ performed better than the P-models highlighting the importance of kernel choice.

As shown in Figure 3, 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 (Figure 3 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 kernel averaging (between the RK evaluated in the pedigree, **A**, and the one evaluated in marker

genotypes, $\mathbf{K}_\theta$). In all environments PM_{KA} gave a predictive MSE as small as that of the best of the $PM_{K,\theta}$ models, suggesting 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.

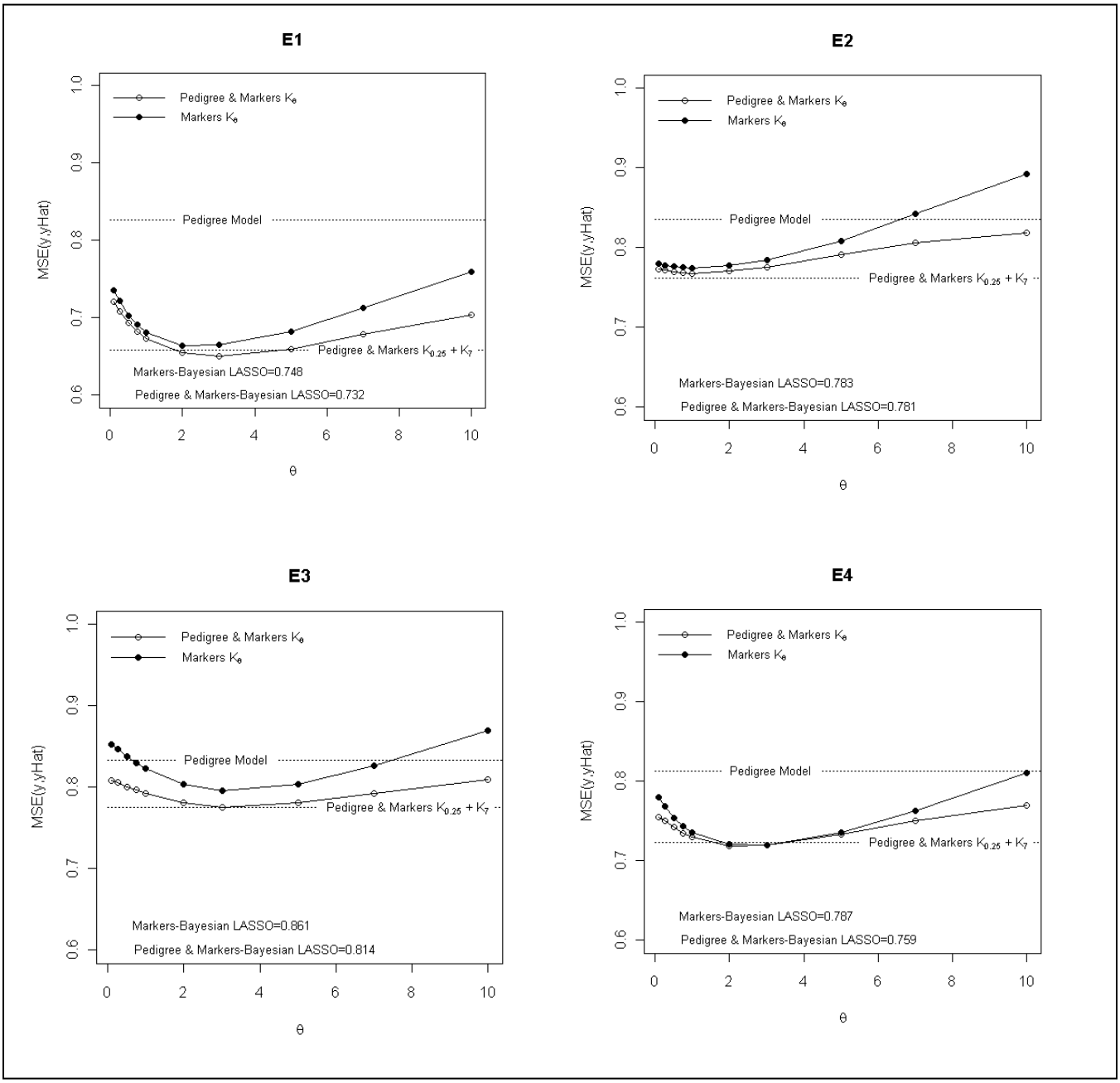

Figure 3. Estimated mean-squared error (MSE) between cross-validation predictions (yHat) and observations (y) versus values of the bandwidth parameter, θ, by environment and model. K_θ is a marker-based reproducing kernel Hilbert spaces regression (RKHS) with bandwidth parameter θ ; Pedigree & Markers 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. E1-E4: environments where the lines were evaluated.

4. Discussion

Incorporating molecular markers into model for prediction of genetic values poses important statistical and computational challenges. Ideally, models for MM 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 unknown function with respect to pairs of points in input space, $Cov[g(t_i), g(t_j)] \propto K(t_i, t_j)$ and 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. Kernel averaging offers a computationally convenient method for automatic kernel selection. In the applications presented in this article, $K_1(t_i, t_j)$ and $K_2(t_i, t_j)$ where either two PD functions evaluated in the same input set (e.g., M_{KA}) or two RKs, each evaluated in a different input set (e.g., PM_{θ}), or a combination of both (e.g., PM_{KA}).

Computational considerations. Gaussian processes offer enormous computational advantages relative to most of the parametric methods for regression on MM. This occurs for two reasons: (a) the model can be represented in terms of *n* unknowns, and (b) factorizations such as EVD or SVD can be used to arrive at highly efficient algorithms. Unfortunately, these benefits cannot be exploited in linear models, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with marker-specific prior precision variances of effects such as BayesA or Bayesian LASSO. This gives to Gaussian process 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 that 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.

Some 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. However, such contribution is model dependent, and the development of algorithms for choosing these weights is not trivial, especially for large *p*.

ACKNOWLEDGMENTS

The authors would like to 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, Martin Schlather and Emilio Porcu. 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.

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 Science* **106**, 8128-8133.

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

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

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

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.

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 & Hall.

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. (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., 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. The Johns Hopkins University Press, Baltimor and London.

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. (Eds). *Statistics an Appraisal,* pp 281-286. Ames, Iowa: 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. (1970a). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55-67.

Hoerl, A.E. & Kennard, R.W. (1970b). 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*, Series 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.

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, United Kingdom: Cambridge University Press.

Sorensen, D. & Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, New York, NY: 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*, Series B **58**, 267-288.

Van Raden, P.M. (2007) Genomic measures of relationship and inbreeding. *Interbull Bull* **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.

APPENDIX

1. Gibbs Sampler

The Appendix describes a Gibbs sampler for a Bayesian RKHS regression. The parameterization is as in [5], extended to two random effects and with the inclusion of an intercept. The derivation uses standard results for Bayesian linear models (e.g., Gelman et al., 2004; Sorensen and Gianola, 2002).

Let $\mathbf{K}_1 = \mathbf{\Lambda}_1\mathbf{\Psi}_1\mathbf{\Lambda}_1'$ and $\mathbf{K}_2 = \mathbf{\Lambda}_2\mathbf{\Psi}_2\mathbf{\Lambda}_2'$ be the eigenvalue decompositions of the two kernel matrices. Extending [5] to two random effects and by including an intercept, the data equation and likelihood function become $\mathbf{y} = \mathbf{1}\mu + \mathbf{\Lambda}_1\boldsymbol{\delta}_1 + \mathbf{\Lambda}_2\boldsymbol{\delta}_2 + \boldsymbol{\varepsilon}$ and $p\left(\mathbf{y}\middle|\mu,\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\sigma^2_\varepsilon\right) = N\left(\mathbf{y}\middle|\mathbf{1}\mu + \mathbf{\Lambda}_1\boldsymbol{\delta}_1 + \mathbf{\Lambda}_2\boldsymbol{\delta}_2\,,\mathbf{I}\sigma^2_\varepsilon\right)$, respectively. And the joint prior is (upon assuming a flat prior for μ):

$$p\left(\mu,\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\sigma^2_\varepsilon,\sigma^2_{\alpha 1},\sigma^2_{\alpha 2}\right) \propto N\left(\boldsymbol{\delta}_1\middle|\mathbf{0},\mathbf{\Psi}_1\sigma^2_{\alpha 1}\right)N\left(\boldsymbol{\delta}_2\middle|\mathbf{0},\mathbf{\Psi}_2\sigma^2_{\alpha 2}\right)$$
$$\times\chi^{-2}\left(\sigma^2_\varepsilon\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma^2_{\alpha 1}\middle|df_{\alpha 1},S_{\alpha 1}\right)\chi^{-2}\left(\sigma^2_{\alpha 2}\middle|df_{\alpha 2},S_{\alpha 2}\right).$$

Above, $\chi^{-2}\left(.\middle|df_.,S_.\right)$ is a scaled inverse chi-square density with degree of freedom $df_.$ and scale-parameter $S_.$, 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:

$$p\left(\mu,\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\sigma^2_\varepsilon,\sigma^2_{\alpha 1},\sigma^2_{\alpha 2}\middle|\mathbf{y}\right) \propto N\left(\mathbf{y}\middle|\mathbf{1}\mu + \mathbf{\Lambda}_1\boldsymbol{\delta}_1 + \mathbf{\Lambda}_2\boldsymbol{\delta}_2\,,\mathbf{I}\sigma^2_\varepsilon\right)$$
$$\times N\left(\boldsymbol{\delta}_1\middle|\mathbf{0},\mathbf{\Psi}_1\sigma^2_{\alpha 1}\right)N\left(\boldsymbol{\delta}_2\middle|\mathbf{0},\mathbf{\Psi}_2\sigma^2_{\alpha 2}\right)$$
$$\times\chi^{-2}\left(\sigma^2_\varepsilon\middle|df_\varepsilon,S_\varepsilon\right)\chi^{-2}\left(\sigma^2_{\alpha 1}\middle|df_{\alpha 1},S_{\alpha 1}\right)\chi^{-2}\left(\sigma^2_{\alpha 2}\middle|df_{\alpha 2},S_{\alpha 2}\right).$$

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,

$$p(\mu|ELSE) \propto N(\mathbf{y}|\mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2, \mathbf{I}\sigma_\varepsilon^2) \propto N(\mathbf{y}^\mu|\mathbf{1}\mu, \mathbf{I}\sigma_\varepsilon^2),$$

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 μ. The fully conditional distribution is then normal with mean $n^{-1}\sum_i y_i^\mu$ and variance $n^{-1}\sigma_\varepsilon^2$.

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

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

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: $\left[\boldsymbol{\Lambda}_1'\boldsymbol{\Lambda}_1\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_1^{-1}\sigma_{\alpha 1}^{-2}\right]\hat{\boldsymbol{\delta}}_1 = \boldsymbol{\Lambda}_1'\mathbf{y}^{\delta_1}\sigma_\varepsilon^{-2}$. Using $\boldsymbol{\Lambda}_1'\boldsymbol{\Lambda}_1 = \mathbf{I}$, the system becomes: $\left[\mathbf{I}\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_1^{-1}\sigma_{\alpha 1}^{-2}\right]\hat{\boldsymbol{\delta}}_1 = \boldsymbol{\Lambda}_1'\mathbf{y}^{\delta_1}\sigma_\varepsilon^{-2}$. 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, $p(\delta_j|ELSE)$ is normal, centered at $\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_{1j}^{-1}\right]^{-1} y_{.j}^{\delta_1}$ and with variance $\sigma_\varepsilon^2\left[1 + \sigma_\varepsilon^2\sigma_g^{-2}\Psi_{1j}^{-1}\right]^{-1}$ where $y_{.j}^{\delta_1} = \boldsymbol{\lambda}_{1j}'\mathbf{y}^{\delta_1}$. Here, $\boldsymbol{\lambda}_{1j}$ is the j^{th} 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: $\left[\mathbf{I}\sigma_\varepsilon^{-2} + \boldsymbol{\Psi}_2^{-1}\sigma_{\alpha 2}^{-2}\right]\hat{\boldsymbol{\delta}}_2 = \boldsymbol{\Lambda}_2'\mathbf{y}^{\delta_2}\sigma_\varepsilon^{-2}$, 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:

$$p\left(\sigma_\varepsilon^2 \middle| \mathbf{y}\right) \propto N\left(\mathbf{y} \middle| \mathbf{1}\mu + \boldsymbol{\Lambda}_1\boldsymbol{\delta}_1 + \boldsymbol{\Lambda}_2\boldsymbol{\delta}_2\ , \mathbf{I}\sigma_\varepsilon^2\right)\chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right)$$
$$\propto N\left(\boldsymbol{\varepsilon} \middle| \mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right)\chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right),$$

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 $df\ = n + df_\varepsilon$ and scale parameter $S = \dfrac{\sum_i \varepsilon_i^2 + df_\varepsilon S_\varepsilon}{n + df_\varepsilon}$.

The fully conditional distribution of $\sigma_{\alpha 1}^2$ is:

$p\left(\sigma_{\alpha 1}^2 \middle| ELSE\right) \propto N\left(\boldsymbol{\delta}_1 \middle| \mathbf{0}, \Psi_1\sigma_{\alpha 1}^2\right)\chi^{-2}\left(\sigma_{\alpha 1}^2 \middle| df_{\alpha 1}, S_{\alpha 1}\right)$, which is a scaled inverse chi-square distribution with $df\ = n + df_{\alpha 1}$ and scale parameter $S = \dfrac{\sum_i \Psi_{1i}^{-1}\delta_{1i}^2 + df_{\alpha 1}S_{\alpha 1}}{n + df_{\alpha 1}}$. Here, Ψ_{1i} is the i^{th} eigen-value of $\mathbf{K}_1$. Similarly, the fully conditional distribution of $\sigma_{\alpha 2}^2$ is scaled inverse chi-square with $df\ = n + df_{\alpha 2}$ and scale parameter $S = \dfrac{\sum_i \Psi_{2i}^{-1}\delta_{2i}^2 + df_{\alpha 2}S_{\alpha 2}}{n + df_{\alpha 2}}$.

2. Tables and Figures

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

	Marker-based Models				**Pedigree & Markers Models**			
	E1	E2	E3	E4	E1	E2	E3	E4
Pedigree Model	.562 (.057)	.580 (.056)	0.493 (.058)	.519 (.055)	NA			
$\mathbf{K}_{0.10}$	.520 (.049)	.561 (.049)	.646 (.056)	.579 (.052)	.410 (.049)	.485 (.052)	.459 (.056)	.451 (.051)
$\mathbf{K}_{0.25}$	.484 (.048)	.545 (.051)	.618 (.057)	.548 (.053)	.386 (.049)	.469 (.051)	.446 (.055)	.437 (.051)
$\mathbf{K}_{0.50}$	.432 (.048)	.524 (.051)	.565 (.061)	.502 (.053)	.347 (.048)	.458 (.051)	.428 (.055)	.414 (.051)
$\mathbf{K}_{0.75}$	.401 (.048)	.507 (.051)	.520 (.062)	.467 (.052)	.318 (.047)	.442 (.052)	.408 (.055)	.397 (.050)
$\mathbf{K}_{1.00}$	.373 (.047)	.490 (.052)	.486 (.062)	.440 (.052)	.294 (.048)	.431 (.053)	.392 (.056)	.379 (.050)
$\mathbf{K}_{2.00}$	.313 (.044)	.436 (.053)	.379 (.060)	.373 (.050)	.232 (.043)	.392 (.053)	.327 (.056)	.330 (.048)
$\mathbf{K}_{3.00}$	.277 (.043)	.399 (.054)	.320 (.056)	.333 (.047)	.199 (.042)	.364 (.056)	.284 (.053)	.300 (.047)
$\mathbf{K}_{5.00}$	.238 (.041)	.347 (.056)	.262 (.051)	.286 (.048)	.155 (.039)	.335 (.060)	.246 (.054)	.269 (.050)
$\mathbf{K}_{7.00}$	.214 (.042)	.323 (.060)	.232 (.052)	.255 (.050)	.136 (.037)	.332 (.067)	.238 (.059)	.255 (.053)
$\mathbf{K}_{10.00}$	.203 (.044)	.309 (.070)	.218 (.057)	.226 (.055)	.121 (.037)	.333 (.075)	.240 (.064)	.261 (.059)
$\mathbf{K}_{0.25}+\mathbf{K}_{7.00}$	.244 (.044)	.402 (.059)	.276 (.060)	.314 (.055)	.152 (.040)	.337 (.058)	.243 (.056)	.276 (.052)
Bayesian LASSO	.532 (.045)	.555 (.047)	.644 (.050)	.582 (.048)	.370 (.044)	.446 (.047)	.427 (.049)	.419 (.045)

E1-E4 are the four environments where wheat lines were evaluated; $\mathbf{K}_{\theta}$ are (Bayesian) reproducing kernel Hilbert spaces regression models using a Gaussian kernel evaluated at marker-genotypes with bandwidth parameter θ; $\mathbf{K}_{0.25}+\mathbf{K}_{7}$ is a model that includes two Gaussian kernels differing only in the value of θ.

Table A2. Mean-squared error between realized phenotypes and cross-validation predictions, by model and environment.

	Marker-based Models				Pedigree & Markers Models			
	E1	E2	E3	E4	E1	E2	E3	E4
Pedigree Model	0.826	0.835	0.834	0.812	NA			
$\mathbf{K}_{0.10}$	0.736	0.779	0.853	0.780	0.721	0.773	0.808	0.755
$\mathbf{K}_{0.25}$	0.722	0.778	0.847	0.768	0.708	0.772	0.806	0.750
$\mathbf{K}_{0.50}$	0.703	0.776	0.838	0.754	0.694	0.769	0.801	0.742
$\mathbf{K}_{0.75}$	0.691	0.775	0.830	0.744	0.682	0.769	0.797	0.734
$\mathbf{K}_{1.00}$	0.681	0.775	0.823	0.735	0.674	0.768	0.793	0.730
$\mathbf{K}_{2.00}$	0.664	0.778	0.804	0.721	0.655	0.771	0.781	0.719
$\mathbf{K}_{3.00}$	0.665	0.785	0.796	0.719	0.651	0.775	0.776	0.720
$\mathbf{K}_{5.00}$	0.683	0.809	0.803	0.736	0.660	0.792	0.781	0.733
$\mathbf{K}_{7.00}$	0.713	0.842	0.827	0.763	0.679	0.806	0.792	0.750
$\mathbf{K}_{10.00}$	0.759	0.892	0.870	0.811	0.704	0.818	0.809	0.770
$\mathbf{K}_{0.25} + \mathbf{K}_{7.00}$	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) reproducing kernel Hilbert spaces regression models using a Gaussian kernel evaluated at marker-genotypes with bandwidth parameter θ; $\mathbf{K}_{0.25} + \mathbf{K}_{7.00}$ is a model that includes two Gaussian kernels differing only in the value of θ.

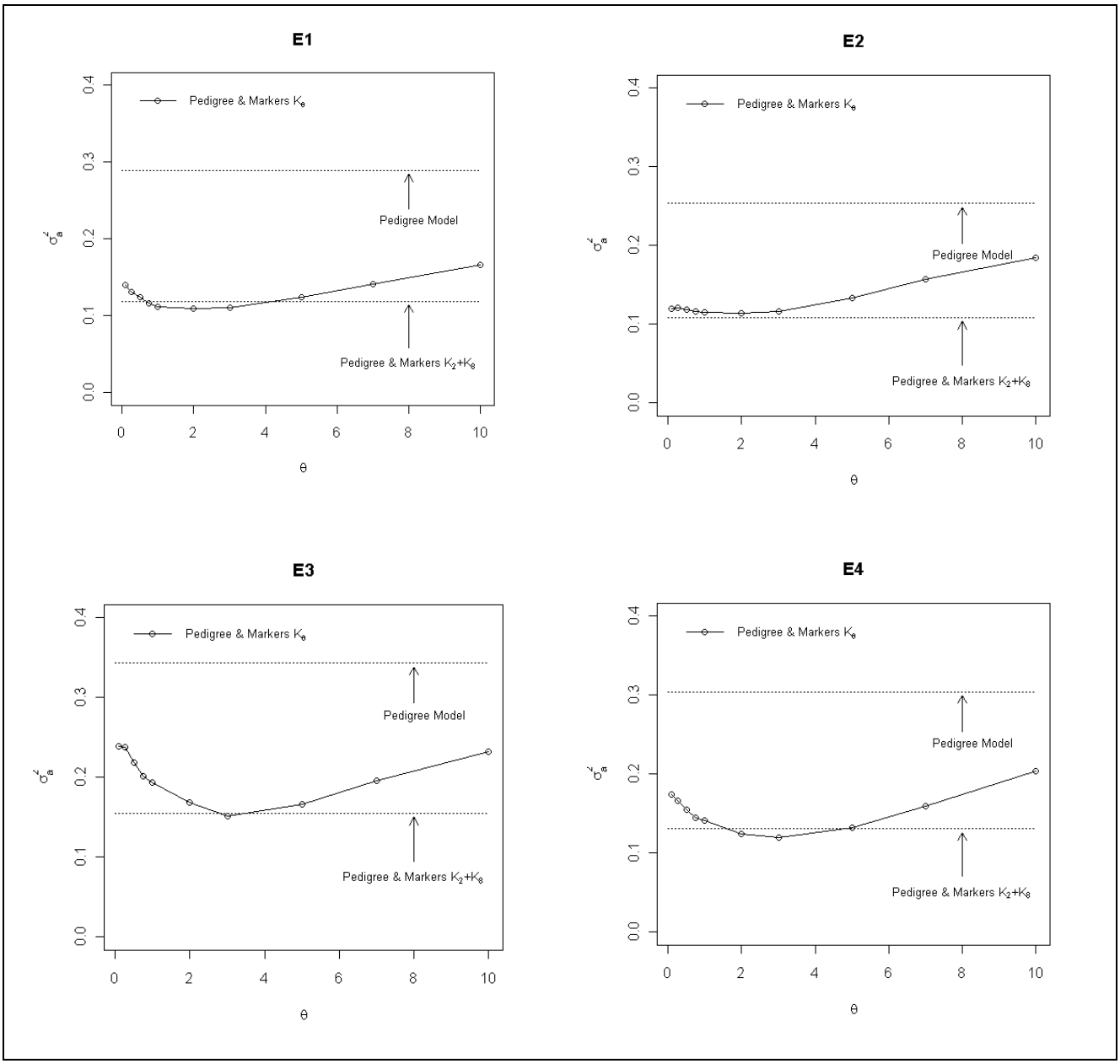

Figure A1. Posterior mean of the variance of the regression on the pedigree, σ_a^2, versus values of the bandwidth parameter, θ, by environment and model. K_θ is a marker-based reproducing kernel Hilbert spaces regression (RKHS) with bandwidth parameter θ ; Pedigree & Markers K_θ uses pedigree and markers, 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. E1-E4: environments where the lines were evaluated.