Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction

Sergio PÉREZ- ELIZALDE, Jaime CUEVAS, Paulino PÉREZ- RODRÍGUEZ, and José CROSSA

One of the most widely used kernel functions in genomic-enabled prediction is the Gaussian kernel. Selection of the bandwidth parameter for kernel regression has generally been based on cross-validation. We propose a Bayesian method for estimating the bandwidth parameter h of a Gaussian kernel as the modal component of the joint posterior distribution of h and the form parameter φ. We present a theory for the Bayesian selection of h in a Transformed Gaussian Kernel (TGK) model and its application in two plant breeding datasets (maize and wheat) that were already predicted using the kernel averaging (KA) model in the context of Reproducing Kernel Hilbert Spaces (RKHS KA). We also compared the prediction accuracy of the proposed method with a model that also uses a Gaussian kernel and estimates the bandwidth parameter using a restricted maximum likelihood method (GK REML). Results for the wheat dataset show that the predictive ability of TGK was at least as good as the predictive ability of model RKHS KA, with TGK showing a significantly smaller Predictive Mean Squared Error (PMSE) than the other two approaches. The TGK model was statistically a better predictor than methods GK REML and RKHS KA in terms of mean PMSE and mean correlations in seven (out of 17) trait-environment combinations in the wheat dataset. Fewer differences were found between models for the maize data; the TGK model generally had similar or inferior prediction accuracy than GK REML and RKHS KA in various analyses. The superiority of GK REML over TGK based on mean PMSE was clear in seven maize traits.

Key Words: Non-parametric regression; Bandwidth selection; Genomic-enabled prediction.

Sergio Perez-Elizalde (E-mail: *sergiop@colpos.mx*) and Paulino Pérez-Rodríguez (E-mail: *perpdgo@gmail.com*) are Professors of Statistics, Department of Statistics, Colegio de Postgraduados, CP 56230 Montecillos, Edo. de México, Mexico. Jaime Cuevas is Professor of Statistics, Department of Science, Universidad de Quintana Roo, CP 77019 Chetumal, Quintana Roo, Mexico (E-mail: *cuevas.jaime@colpos.mx*). José Crossa (✉) is a Biometrician, Biometrics and Statistics Unit of the International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, DF, Mexico (E-mail: *j.crossa@cgiar.org*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 20, Number 4, Pages 512–532
DOI: 10.1007/s13253-015-0229-y

1. INTRODUCTION

The rapid development of sequencing technologies has made it possible to use dense molecular marker information for genomic selection (GS) (Meuwissen et al. 2001), which has been shown to improve prediction accuracy (de los Campos et al. 2010; Crossa et al. 2010; Heslot et al. 2012; Pérez-Rodríguez et al. 2012).

The standard regression model of phenotypes, y_i, ($i = 1, 2, \ldots, n$ individuals) on markers ($j = 1, 2, \ldots, p$ markers) is represented by $y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, or, in matrix notation, as

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where μ is a general mean, $\mathbf{1}$ is a vector of ones of order $n \times 1$, $\boldsymbol{X}$ is the $n \times p$ incidence matrix of standardized markers, $\boldsymbol{\beta}$ is the vector of unknown marker effects, and $\boldsymbol{\varepsilon}$ is the vector of model errors. Standard assumptions of normality, independency and homoscedasticity are such that $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right)$. Assuming $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}\right)$ and independent of $\boldsymbol{\varepsilon}$, then (1.1) is usually fitted with the Ridge Regression Best Linear Unbiased Predictor (RRBLUP) method. By letting $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta}$, (1.1) may be represented as

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\boldsymbol{u}$ is a vector of random genetic effects and independent of $\boldsymbol{\varepsilon}$. If $\boldsymbol{u} \sim N\left(\mathbf{0}, \boldsymbol{G}\sigma_u^2\right)$, and $\boldsymbol{G} = \frac{\boldsymbol{X}\boldsymbol{X}'}{p}$, then (1.2) indicates a genomic mixed model which may be estimated through the GBLUP method (VanRaden 2008). Although (1.1) and (1.2) are equivalent, estimation of (1.2) is computationally more convenient (de los Campos et al. 2012). Models (1.1) and (1.2) may be extended in order to account for the vector of fixed effects $\boldsymbol{\vartheta}$ of factors arranged in an incidence matrix $\boldsymbol{Z}$ by adding the term $\boldsymbol{Z}\boldsymbol{\vartheta}$ to the right-hand side of (1.1) or of (1.2). However, the data included in this study were already pre-corrected for fixed effects such as experimental design and location effects.

A great deal of research on developing Bayesian linear regression models for capturing the true genetic signal has been published recently; usually these Bayesian models differ in their prior distributions (de los Campos et al. 2012). In general, the parametric regression function has a rigid structure comprising a set of assumptions which may not be met in genomic selection problems. Furthermore, the sample size (n) is usually much smaller than the number of predictors (markers), p ($p \gg n$), a problem known as "the curse of dimensionality" (Bellman 1961). Departures from linearity can be addressed by semi-parametric approaches, such as Reproducing Kernel Hilbert Space (RKHS) regressions or different types of neural networks (Gianola et al. 2006, 2011; Gianola and van Kaam 2008; de los Campos et al. 2010; González-Camacho et al. 2012; Pérez-Rodríguez et al. 2012). Gianola et al. (2006, 2014) suggested using RKHS regression for semi-parametric, genomic-enabled prediction, and pointed out that non-parametric methods such as kernel regression are necessary to reduce the dimension of the parametric space, and may be able to capture complex cryptic interaction among markers. However, recovering non-additive interaction among markers is an open field of research and the most successful results have been obtained through kernel-based methods (Howard et al. 2014). Morota and Gianola

(2014) and Gianola et al. (2014) pointed out that most studies carried out so far suggest that whole-genome prediction coupled with combinations of kernels may capture non-additive variation.

The basic idea underlying the RKHS approach to GS (Kimeldorf and Wahba 1971; Gianola 2013) is to use the matrix of markers $\boldsymbol{X}$ to build a covariance structure among genetic values $\boldsymbol{u}$ in (1.2). Therefore, in this context, $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma_a^2 \boldsymbol{K}_h\right)$ is independent of $\boldsymbol{\varepsilon}$ (de los Campos et al. 2010), $\boldsymbol{K}_h$ is a symmetric positive semi-definite matrix of order $n \times n$, known as the reproducing kernel (RK) matrix, which depends on the markers, the bandwidth parameter $h > 0$, the additive genetic variance component $\sigma_a^2 > 0$, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of homoscedastic and independent normal errors. This general approach requires choosing an RK, for example, a Gaussian kernel function $K_h\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp\left(-h d_{ij}^2 / q_{0.05}\right)$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the marker vectors for the ith and jth individuals, respectively, and $q_{0.05}$ is the fifth percentile of the squared Euclidean distance $d_{ij}^2 = \sum_k \left(x_{ik} - x_{jk}\right)^2$ (González-Camacho et al. 2012).

Let $\boldsymbol{u} = \boldsymbol{K}_h \boldsymbol{f}$; then the (1.2) may be represented as

$$\boldsymbol{y} = \mu \boldsymbol{1} + \boldsymbol{K}_h \boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{1.3}$$

One of the methods used in GS to fit (1.2) and (1.3) is the standard ridge regression (Hoerl and Kennard 1970). Assume that shrinkage parameter γ is a known scalar and that the square of the norm of $\boldsymbol{u} \in H$, (where H is the collection of functions in RKHS of real-valued functions) is given by $\|\boldsymbol{u}\|_H^2 = \boldsymbol{f}' \boldsymbol{K}_h \boldsymbol{f}$ (Kimeldorf and Wahba 1971); then for each value of γ, the ridge estimator is $\hat{\boldsymbol{f}} = \left(\boldsymbol{K}_h + \gamma \boldsymbol{I}\right)^{-1} \left(\boldsymbol{y} - \bar{y}\boldsymbol{1}\right)$. Then it follows that $\hat{\boldsymbol{u}} = \boldsymbol{K}_h \left(\boldsymbol{K}_h + \gamma \boldsymbol{I}\right)^{-1} \left(\boldsymbol{y} - \bar{y}\boldsymbol{1}\right)$. A Bayesian version can be derived by assigning the prior $\boldsymbol{f} \sim N\left(\boldsymbol{0}, \sigma_a^2 \boldsymbol{K}_h^{-1}\right)$ or, equivalently, $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma_a^2 \boldsymbol{K}_h\right)$, where σ_a^2 and h are unknown hyperparameters and $\gamma = \sigma^2 / \sigma_a^2$, so that $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{u}}$ are the modes of the posterior distributions of $\boldsymbol{f}$ and $\boldsymbol{u}$, respectively.

In the non-parametric regression framework, the key idea is to develop more flexible models that will capture complex signals, thus improving the prediction accuracy of the genetic values. The RK function takes as input the markers used to create a covariance structure for the genetic values $\text{Cov}\left(u_i, u_j\right) = \sigma_a^2 K_h\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ $\left(\sigma_a^2 > 0\right)$, where u_i, and u_j are the genetic values of the ith and jth individuals. For example, the linear kernel given by $\boldsymbol{K} = \frac{XX'}{p}$ [see the definition of $\boldsymbol{G}$ in (1.2)] can be used to reduce the dimensionality of the genotypic data and, hence, the number of parameters to be estimated (de los Campos et al. 2010; VanRaden 2008). A comprehensive review of various kernel-based approaches for capturing genetic variation in the context of genomic-enabled prediction was recently published by Morota and Gianola (2014).

The RK function has two components: a distance measure between individuals based on markers and the bandwidth parameter h that controls the rate of decay of the covariance between genotypes. The RK function can either be selected from different RK options or constructed with Mercer's theorem, as described by Genton (2001). The most commonly used RK function in GS is the Gaussian kernel (Gianola et al. 2006). Other RK functions, such as the t (Tussel et al. 2014) and the exponential (Endelman 2011), have also been

used. However, none of them have proven to be consistently superior to the Gaussian RK function (Ober et al. 2011). Assuming that the Gaussian kernel has been selected, a remaining crucial step is to tune the bandwidth parameter, which is usually based on cross-validation or penalization (Härdle 1990).

Assume that the RKHS in (1.3) is non-parametric and involves a fixed operator $\boldsymbol{K}_h$ for a known h. Then, given h, we need to estimate $\boldsymbol{f}$ or, equivalently, $\boldsymbol{u}$ and the variance components. One way of achieving this goal is to use Bayesian inference within the theoretical framework of inverse problems (Aster et al. 2005), such as those proposed by Knapick et al. (2012) and Cavalier (2008), and apply them to genomic-enabled prediction (Cuevas et al. 2014). In an attempt to optimize prediction ability in GS, the bandwidth parameter h is selected by means of cross-validation in a grid of values, as shown in Heslot et al. (2012). However, cross-validation methods for selecting h are computationally very intense and sometimes not easy to apply in GS (Gianola and van Kaam 2008). Endelman (2011) used the restricted maximum likelihood (REML) to jointly fit (1.2) including h for a model with the Gaussian RK function; the proposed method (GK REML) is implemented in the R package rrBLUP. Note that when prior distributions of variance components are assumed to be flat, the REML estimators coincide with the mode of the marginal posterior density (Harville 1974; Blasco 2001). An efficient Bayesian method that minimizes the effect of selecting an inadequate value of h and thus improves GS prediction accuracy was proposed by de los Campos et al. (2010), who used an extension of (1.2) with $\boldsymbol{u} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_k$, such that each random effect, $\boldsymbol{u}_i, i = 1, \ldots, k$, is weighted by its variance component in a process termed Kernel Averaging (KA) or multi-kernel. Kernel averaging uses a prior distribution that assumes independence among random effects $p\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k | \sigma^2_{u_1}, \ldots, \sigma^2_{u_k}\right) = N\left(\boldsymbol{u}_1 | \boldsymbol{0}, \boldsymbol{K}_1 \sigma^2_{u_1}\right) \times \cdots \times N(\boldsymbol{u}_k | \boldsymbol{0}, \boldsymbol{K}_k \sigma^2_{u_k})$. The set of kernels $\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_k\}$ is defined based on a set of values of $h \in \{h_1, \ldots, h_k\}$. The set of values of h is selected by first defining a range of values that will avoid small and large values of h. Achieving these criteria is important because small values of h will produce a kernel matrix of ones, whereas large values of h will produce a kernel matrix whose off-diagonal elements will approach zero. These two extreme cases are avoided because they do not contribute information that can be used for prediction; for example, in the case of small values of h, the information provided is redundant with the intercept [which is already included in (1.3)] and in the case of large values of h, will lead to a random effect with a variance covariance structure similar to that of the error term in (1.3). de los Campos et al. (2010) also show that $\boldsymbol{u} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_k \sim N\left(\boldsymbol{0}, \sigma^2_u \bar{K}\right)$, where $\bar{\boldsymbol{K}} = \sum_l \boldsymbol{K}_l \sigma^2_{u_l} / \sigma^2_u$ and argue that inferring the weights leads to a kernel $\bar{\boldsymbol{K}}$ that is optimal. The strategy for specifying the values of h is also described in Pérez-Rodríguez and de los Campos (2014).

In this study, we propose a Bayesian method for selecting the bandwidth parameter h following a simple and logical idea put forward by Gianola and van Kaam (2008), that is, to assign a prior $p(h)$ and obtain a posterior point estimate of h. This is achieved by transforming (1.3) following the approach of Cuevas et al. (2014), which is similar to that proposed by de los Campos et al. (2010), but transforming both sides of (1.3) as in Cavalier (2008). This paper is organized as follows. In Sect. 2, we define the general framework (prior and posterior distributions) of the proposed Bayesian Transformed Gaussian Kernel (TGK) model; we also describe the likelihood, the prior, and the posterior of the bandwidth

parameter. In Sect. 3, we describe the two real datasets analyzed using the TGK model, as well as details of the prediction assessment using a random cross-validation scheme. Section 4 gives the prediction accuracy results for the two datasets; results are discussed in Sect. 5 and conclusions are provided in Sect. 6.

2. STATISTICAL METHODS

2.1. Bayesian Selection of the Bandwidth for a Transformed Gaussian Kernel (TGK)

For genomic-enabled prediction, Cuevas et al. (2014) showed the advantages of transforming both sides of the parametric linear regression model by an orthonormal matrix based on a singular value decomposition (SVD) of the matrix of regression variables, $\boldsymbol{X}$. Some advantages of this method are (i) it reduces computing time because of the reduction in dimensionality of the vector of regression parameters, (ii) the prior distributions for regression parameters have variances that mimic the decay of the singular values of $\boldsymbol{X}$, thus controlling over-fitting, and (iii) the posteriors of the transformed regression parameters are conditionally independent. Following Cuevas et al. (2014), a model based on $\boldsymbol{K}_h$ (not directly $\boldsymbol{X}$) is described below.

In (1.3), we replace the linear operator $\boldsymbol{K}_h$, with its eigenvalue decomposition $\boldsymbol{K}_h = \boldsymbol{U}_h \boldsymbol{S}_h \boldsymbol{U}_h'$, where $\boldsymbol{U}_h$ is a square orthogonal matrix and $\boldsymbol{S}_h$ is the diagonal matrix of eigenvalues, and where sub-indexes express dependency on the value of h; this eigenvalue decomposition is a common strategy also used in parametric methods for genome prediction (Zhou et al. 2013). When multiplying both sides of (1.3) by $\boldsymbol{U}_h'$, we have

$$\boldsymbol{U}_h'\boldsymbol{y} = \mu \boldsymbol{U}_h'\boldsymbol{1} + \boldsymbol{U}_h'\left(\boldsymbol{U}_h \boldsymbol{S}_h \boldsymbol{U}_h'\right)\boldsymbol{f} + \boldsymbol{U}_h'\boldsymbol{\varepsilon}.$$

Let $\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{y}) = \boldsymbol{U}_h'\boldsymbol{y}, \boldsymbol{t} = \boldsymbol{U}_h'\boldsymbol{1}, \boldsymbol{b} = \boldsymbol{U}_h'\boldsymbol{f}, \tilde{\boldsymbol{\varepsilon}} = \boldsymbol{U}_h'\boldsymbol{\varepsilon}$. Since $\boldsymbol{U}_h'\boldsymbol{U}_h = \boldsymbol{U}_h\boldsymbol{U}_h' = \boldsymbol{I}_n$, the model becomes

$$\boldsymbol{d} = \mu \boldsymbol{t} + \boldsymbol{S}_h \boldsymbol{b} + \tilde{\boldsymbol{\varepsilon}}. \tag{2.1}$$

By assuming that $\boldsymbol{\varepsilon} \sim N\left(\boldsymbol{0}, \boldsymbol{I}_n \sigma^2\right)$, it follows that $\tilde{\boldsymbol{\varepsilon}} \sim N\left(\boldsymbol{0}, \boldsymbol{I}_n \sigma^2\right)$. Then, the joint distribution of transformed data $\boldsymbol{d}$, given h, $\boldsymbol{b}$ and σ^2, is

$$p\left(\boldsymbol{d}|\mu, \boldsymbol{b}, \sigma^2, h\right) = \prod_{i=1}^{n} N(d_i|\mu t_i + s_i b_i, \sigma^2).$$

Note that (2.1) may be simply expressed as

$$d_i = \mu t_i + s_i b_i + \tilde{\varepsilon}_i, (i = 1, \ldots, n). \tag{2.2}$$

In the theory of inverse problems, this model is known as the sequential spatial model (Cavalier 2008). As will be seen later, this representation makes it easy to draw samples from the joint posterior distribution of the target parameters. Estimation of the target parameters

in the original model is done directly through the inverse transformation $\boldsymbol{f} = \boldsymbol{U}_h\boldsymbol{b}$ and $\boldsymbol{u} = \boldsymbol{U}_h\boldsymbol{S}_h\boldsymbol{b}$.

2.2. JOINT POSTERIOR DISTRIBUTION

If the prior of $\boldsymbol{f}$ is $N\left(\boldsymbol{0}, \sigma_a^2\boldsymbol{K}_h^{-1}\right)$, where σ_a^2 is the additive genetic variance, then it can be assumed that $\sigma_a^2 = \varphi\sigma^2$, with $\varphi > 0$ (note that $\varphi = 1/\gamma$), such that the prior distribution of $\boldsymbol{b} = \boldsymbol{U}_h'\boldsymbol{f}$ is $N\left(\boldsymbol{b}|\boldsymbol{0}, \varphi\sigma^2\boldsymbol{S}_h^{-1}\right)$. We also assume that the scale hyperparameter φ is an unknown quantity such that we should assign a prior distribution to φ. Therefore, the joint posterior distribution of $\boldsymbol{b}$, μ, σ^2, φ, and h in (2.1) is given by

$$p\left(\mu, \boldsymbol{b}, \sigma^2, h, \varphi|\boldsymbol{y}\right) \propto N\left(\boldsymbol{d}|\mu\boldsymbol{t} + \boldsymbol{S}_h\boldsymbol{b}, \sigma^2\boldsymbol{I}\right) p\left(\mu\right) N\left(\boldsymbol{b}|\boldsymbol{0}, \varphi\sigma^2{\boldsymbol{S}_h}^{-1}\right) \chi^{-2}\left(\sigma^2|\upsilon_\varepsilon, \tau_\varepsilon\right) p\left(h\right) p\left(\varphi\right), \tag{2.3}$$

where the prior for μ is uniform and the prior for σ^2 is the inverse scaled chi-squared with υ_ε degrees of freedom (*df*) and scale parameter τ_ε; the priors for φ and h are described below.

2.3. BANDWIDTH ESTIMATION

Selection of bandwidth h is a challenging problem. The ideal solution is to consider h as an unknown quantity and obtain its marginal distribution through Bayes' rule, which implies approximation of the joint posterior distribution of $\left(\mu, \boldsymbol{b}, \sigma^2, \varphi, h\right)$ via Markov Chain Monte Carlo (MCMC) or other computational tools. However, any simulation method requires recalculating the kernel at each iteration, which is computationally intensive.

Here, we propose fixing h and φ to the mode $(\tilde{h}, \tilde{\varphi})$ of their joint posterior distribution $p(h, \varphi|\boldsymbol{y})$, which is given by

$$\begin{aligned} p\left(h, \varphi|\boldsymbol{y}\right) &= \int p\left(\boldsymbol{\theta}, h, \varphi|\boldsymbol{y}\right) \mathrm{d}\boldsymbol{\theta} = \int \frac{L\left(\boldsymbol{\theta}, h, \varphi|\boldsymbol{y}\right) p\left(\boldsymbol{\theta}, h, \varphi\right)}{m\left(\boldsymbol{y}\right)} \mathrm{d}\boldsymbol{\theta} \\ &= \frac{p\left(h\right) p\left(\varphi\right) m\left(\boldsymbol{y}|h, \varphi\right)}{m\left(\boldsymbol{y}\right)} \int \frac{L\left(\boldsymbol{\theta}, h, \varphi|\boldsymbol{y}\right) p\left(\boldsymbol{\theta}|h, \varphi\right)}{m\left(\boldsymbol{y}|h, \varphi\right)} \mathrm{d}\boldsymbol{\theta} \propto p\left(h\right) p\left(\varphi\right) m\left(\boldsymbol{y}|h, \varphi\right), \end{aligned}$$

where $\boldsymbol{\theta} = \left(\mu, \boldsymbol{b}, \sigma^2\right)'$, $L\left(\boldsymbol{\theta}, h, \varphi|\boldsymbol{d}\right)$ is the likelihood function, the joint prior is of the form $p\left(\boldsymbol{\theta}, h, \varphi\right) = p\left(\boldsymbol{\theta}|h, \varphi\right) p\left(h\right) p\left(\varphi\right)$, and $m\left(\boldsymbol{y}|h, \varphi\right) = \int L\left(\boldsymbol{\theta}, \varphi, h|\boldsymbol{y}\right) p\left(\boldsymbol{\theta}|h, \varphi\right) \mathrm{d}\boldsymbol{\theta}$ is the predictive distribution given h and φ.

Since the posterior distribution $p\left(h, \varphi|\boldsymbol{y}\right)$ is proportional to marginal likelihood $m\left(\boldsymbol{y}|h, \varphi\right)$ times joint prior density $p\left(h\right) p\left(\varphi\right)$, the first step is to analytically derive $m\left(\boldsymbol{y}|h, \varphi\right)$; then, the posterior mode of $p\left(h, \varphi|\boldsymbol{y}\right)$ can be estimated. Note that the problem reduces to finding the maximum of a function of only two parameters (h, φ).

2.4. THE MARGINAL LIKELIHOOD OF THE BANDWIDTH

To obtain the marginal distribution of h from (1.3), we integrate out $\boldsymbol{\theta}$ from the likelihood using the prior density as a weighting function (Berger et al. 1999; Maruyama and George

2011) such that the marginal likelihood of $\boldsymbol{y}$ given (h, φ) is

$$m\left(\boldsymbol{y}|h,\varphi\right)=\int_{0}^{\infty}\int_{\mathbb{R}^{n}}\int_{-\infty}^{\infty}p(\boldsymbol{d}|\mu,\boldsymbol{b},\sigma^{2},h)p\left(\mu,\boldsymbol{b},\sigma^{2}\right)\mathrm{d}\mu\mathrm{d}\boldsymbol{b}\mathrm{d}\sigma^{2},$$

where $p\left(\mu,\boldsymbol{b},\sigma^{2}\right)$ is the joint prior distribution of $\left(\mu,\boldsymbol{b},\sigma^{2}\right)$, which are assumed to be independent, that is, $p\left(\mu,\boldsymbol{b},\sigma^{2}\right)=p\left(\mu\right)p\left(\boldsymbol{b}\right)p\left(\sigma^{2}\right)$. Assigning uniform, Gaussian, and inverted scaled Chi-square prior distributions to $\mu,\boldsymbol{b}$ and σ^{2}, respectively, it follows that the marginal likelihood is (see Appendix A):

$$m\left(\boldsymbol{y}|h,\varphi\right)\propto\prod_{i=1}^{n}(1+\varphi s_{i})^{-\frac{1}{2}}\left[\tau_{\varepsilon}+\sum_{i=1}^{n}\frac{\tilde{d}_{i}^{2}}{(1+\varphi s_{i})}\right]^{\frac{\upsilon_{\varepsilon}+n-1}{2}},\tag{2.4}$$

where $\tilde{d}_i$ is the ith element of the vector $\tilde{\boldsymbol{d}}=\boldsymbol{U}_{h}'\left(\boldsymbol{y}-\bar{y}\boldsymbol{1}\right)$ and s_i is the ith eigenvalue of the spectral decomposition of $\boldsymbol{K}_h$. Then, $m\left(\boldsymbol{y}|h,\varphi\right)$ depends implicitly on h. We propose finding the optimal value of h and φ by maximizing $p\left(h,\varphi|\boldsymbol{y}\right)$, which implies Bayesian estimators under 0–1 loss function for h and φ.

2.5. PRIOR DISTRIBUTIONS OF h AND φ

An appropriate prior for h may be a Gamma distribution $p\left(h\right)\sim Ga\left(h|\upsilon,\tau\right)$,where υ should be greater than 2 to avoid infinite prior variance; by fixing υ and the prior mean μ_h, the scale parameter τ is given by ν/μ_h. Based on the KA method of de los Campos et al. (2010), the idea is to find an interval of values of h that would tend, at one extreme, to have the $\boldsymbol{K}_h$ matrix as a diagonal matrix (local bandwidth) and, at the other extreme, to have the $\boldsymbol{K}_h$ matrix as a matrix of ones (global bandwidth), which would allow an appropriate interval to be computed. For φ we should assign a proper prior with support on the positive real numbers; then, an alternative is the uniform prior on $(0, B)$, where $B > 0$ is large enough to include values with high posterior density.

2.6. ESTIMATION OF THE POSTERIOR MODE OF h AND φ

The posterior distribution of (h, φ) is $p\left(h,\varphi|\boldsymbol{y}\right)\propto m(\boldsymbol{y}|h,\varphi)Ga\left(h|\upsilon,\tau\right)p\left(\varphi\right)$ and its mode may be numerically estimated by the values of (h, φ) within an area that produces the highest value of the posterior density. As demonstrated later, the evaluation of $p\left(h,\varphi|\boldsymbol{y}\right)$ does not require computation of repetitive matrix inverses or multiplications, such that the joint posterior mode $\left(\tilde{h},\tilde{\varphi}\right)$ can be determined by a numerical approximation.

2.7. GIBBS SAMPLER

It is difficult to directly get samples from the joint posterior distribution in (2.3), as it does not have a closed form. However, it is possible to obtain the closed forms for the conditional distributions of the parameters. This allows using Markov Chain Monte Carlo (MCMC) through the Gibbs Sampler (Gelfand and Smith 1990) algorithm, which samples sequentially

from the full conditional distribution until it reaches a stationary process, converging to the joint posterior distribution.

We carried out convergence and diagnostic tests on different datasets. The Gelman–Rubin convergence tests (Gelman and Rubin 1992) for the model were satisfactory. The Raftery–Lewis test (Raftery and Lewis 1992) suggested a small burn-in period and that the number of iterations should be between 10,000 and 20,000 for the datasets used. With the aim of decreasing the potential impact of MCMC errors on prediction accuracy, we performed a total of 60,000 iterations with a burn-in of 10,000 and a thinning of 10, so that 5,000 samples were used for inference. The Gelman–Rubin convergence tests and the Raftery–Lewis tests were performed using the Convergence Diagnosis and Output Analysis (CODA) R package for MCMC (Plummer et al. 2006). The Effective Sample Size (ESS) varied between 1,400 and 2,000 for all parameters.

2.8. Data and Software Repository

An R script (R Core Team 2015) to implement the Gibbs Sampler described above and a brief document in Help and Programs.zip file can be downloaded from the following link http://hdl.handle.net/11529/10234 together with the file Data.zip containing the datasets used in this study. Supplemental Tables S1 and S2 can also be found at this link.

3. EXPERIMENTAL DATA

The two data sources used in this study were previously used by several authors (Crossa et al. 2010; Pérez-Rodríguez et al. 2012; González-Camacho et al. 2012) and most recently by Cuevas et al. (2014) for fitting several Bayesian inverse regression methods.

3.1. Wheat DataSet

This dataset included 306 elite wheat lines from CIMMYT's Global Wheat Program that were used by Pérez-Rodríguez et al. (2012). These lines were genotyped with 1,717 diversity array technology (DArT) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). These DArT corresponds to scores for presence (1) and absence (0) for dominant marker (see Wenzl et al. 2004, for further details). Two traits were analyzed: days to heading (DTH) measured in ten different environments and grain yield (GY) measured in seven different environments.

3.2. Maize DataSet

The maize data represented 21 trait-environment combinations for 300 tropical inbred lines genotyped with 55,000 SNPs each (González-Camacho et al. 2012). A first group of traits included female flowering (FFL) or days to silking, male flowering (MFL) or days to anthesis, and the anthesis-silking interval (ASI). Each trait was evaluated under severe drought stress (SS) and in well-watered (WW) environments. This dataset was also used by Crossa et al. (2010) for assessing prediction performance.

In the second group of traits, grain yield (GY) was obtained under severe drought (SS) and in well-watered (WW) environments. Further, GY of the 300 maize lines was also measured in a large number of relatively high yielding (GY-HI) and low yielding environments (GY-LO). In the third group of traits, the 300 maize lines were also evaluated in nine international environments for gray leaf spot (GLS), a disease caused by the fungus *Cercospora zeae-maydis*. Finally, in the fourth group, the same 300 lines were evaluated in another set of trials for northern corn leaf blight (NCLB), a disease caused by the fungus *Exserohilum turcicum*.

3.3. Assessing Prediction Accuracy Using Cross-Validation

Once the value of h was estimated, its predictive ability was evaluated using the TGK model, where conditional distributions were computed according to (2.3). Model predictions for each of the maize and wheat datasets were done in each of 50 random partitions, with 90 % of individuals in the training set and 10 % of individuals in the testing set to be predicted. We used the same 50 random partitions as those used by Pérez-Rodríguez et al. (2012), González-Camacho et al. (2012), and Cuevas et al. (2014). We computed both Pearson's correlation coefficient between the posterior predictive mean and the corresponding observed values and the Predictive Mean Squared Error (PMSE) as measures of prediction ability.

The GK REML method included in the rrBLUP package (Endelman 2011) was fitted and the mean Pearson's correlation between predicted and observed values and the PMSE was also calculated for the same 50 random partitions. Therefore, Pearson's correlation between predicted and observed values and the PMSE for the same 50 random partitions were computed for the TGK model, GK REML, and RKHS KA, as reported by Pérez-Rodríguez et al. (2012) for the wheat dataset, and by González-Camacho et al. (2012) for the maize dataset.

3.4. Assessing the Significant Differences Among the Models (Methods)

We separated the traits in the wheat datasets into two subsets, those comprising the DTH traits and those including the YLD traits. The trait-environment combinations of the maize dataset included flowering time (FFL, FML, ASI) and grain yield (GY) in different stress environments (WW, SS, HI and LOW) and those including the disease traits GLS and NCBL measured in different environments. The significant differences among the methods and models were assessed based on the two criteria, correlation between the predicted and observed values and the PMSE. We computed two-tailed hypothesis testing of the three models, TGK, RKHS KA, GK REML (using a Type I error rate of 5 %) to examine significant differences between models based on correlation and PMSE criteria.

3.5. Full Data Analyses

For each dataset (trait-environment combination), we computed the corresponding parameters using the full data. For the TGK model, we used three different priors with the objective of examining the influence of each of them on the predictive ability of the TGK

model. We fitted the TGK model using, as priors for h, Gamma distributions with mean 2 (TGK2) and 1 (TGK1), and with a flat prior (TGKF). The inferences for TKG and RKHS KA were based on 30,000 MCMC samples that were obtained after discarding 5,000 samples (burn-in).

For the TGK model, a Gamma prior distribution for h with mean 2 was used. For GK REML, the default values given in the rrBLUP R package for the grid were used. For RKHS KA, the values used were taken directly from Table 2 in Pérez-Rodríguez et al. (2012) for the wheat datasets and from Table 1 in González-Camacho et al. (2012) for the maize datasets (see footnote in Supplemental Tables S1 and S2 of this study).

4. RESULTS

4.1. Full Data Analysis

Supplemental Tables S1 and S2 (see http://hdl.handle.net/11529/10234) contain the results for TGK1, TGK2, and TGKF. In general, it can be seen that R, defined as the ratio [genomic variance]/[genomic + error variance], for methods TGK1, TGK2, and TGKF are very similar; therefore, the prior should have very little impact on the bandwidth parameter h. Estimates of the bandwidth parameter depend on the data and traits. For the wheat data (DTH and GY), the estimates of h varied around 1 and 2 (Supplemental Table S1), whereas for the maize dataset, the estimates of h from the full data varied from 0.27 to 3.9 (Supplemental Table S2).

4.2. Wheat DataSet

The results of the two-tailed hypothesis tests all conducted based on a Type I error rate of 5 % are shown in Table 1 for the mean PMSE and the mean correlation across 50 random partitions for a total of 17 trait-environment combinations. For the PMSE criterion, TGK was significantly more accurate than GK REML in seven trait-environments combinations (DTH2, DTH8-11, YLD1, and YLD3) and significantly more accurate than RKHS KA in 14 cases. Also, GK REML gave significantly higher prediction accuracy than TGK in four cases (DTH4, DTH12, YLD4, and YLD7). RKHS KA was significantly the best in only one case (YLD2).

Based on correlations, TGK was significantly more accurate than RKHS KA and GK REML in seven trait-environment combinations (the same as those detected as significant based on the PMSE criterion, except for YLD1 and DTH2) (Table 1). In only one case (DTH4), was GK REML significantly more accurate, based on correlation, than TGK.

4.3. Maize DataSet

The mean PMSE and correlations across 50 random partitions and the significant differences between them based on two-tailed hypothesis tests (Type I error rate = 5 %) are given in Table 2 for a total of 21 trait-environment combinations. Fewer significant differences between the methods were found for the maize data compared to the wheat data; for nine

Table 1. Mean Predictive Mean Squared Error (PMSE) and correlation from 50 random cross-validation partitions (and their standard deviation, SD) of TGK, RKHA KA, and GK REML for various trait-environment combinations of the wheat dataset.

Trait-environment	TGK	SD	RKHS KA	SD	GK REML	SD
Mean PMSE						
DTH1	**10.8477***	3.0137	11.0235	3.0458	**10.8640**	2.9814
DTH2	**10.1821**	11.9974	10.1949	11.6981	10.2789	11.8519
DTH3	**6.2090**	2.1940	6.3071	2.2361	**6.2365**	2.1981
DTH4	21.3998	6.2440	**21.1418**	6.1762	**21.1175**	6.0726
DTH5	7.9089	2.2520	7.9491	2.2058	7.9247	2.2572
DTH8	**12.8310**	3.0244	13.1249	3.1391	13.2197	3.1712
DTH9	**20.3664**	5.9080	20.6926	5.9438	20.5104	5.9239
DTH10	**6.5580**	2.0023	6.6379	2.0517	6.6918	1.9853
DTH11	**5.9693**	1.8404	6.0122	1.8057	6.0849	1.8363
DTH12	13.8573	7.1541	**13.2890**	6.9796	**13.3424**	6.9308
YLD1	**0.0668**	0.0158	0.0673	0.0159	0.0678	0.0159
YLD2	0.0607	0.0232	**0.0594**	0.0227	0.0612	0.0232
YLD3	**0.0488**	0.0127	0.0501	0.0132	0.0496	0.0132
YLD4	0.1990	0.0497	0.2029	0.0505	**0.1977**	0.0496
YLD5	**0.3455**	0.1077	0.3498	0.1059	**0.3468**	0.1121
YLD6	**0.1160**	0.0344	0.1182	0.0347	**0.1163**	0.0351
YLD7	0.3746	0.1076	0.3755	0.1090	**0.3721**	0.1079
Mean correlation						
DTH1	**0.6654**	0.0905	0.6585	0.0920	**0.6637**	0.0903
DTH2	**0.6804**	0.1340	0.6767	0.1299	0.6749	0.1336
DTH3	**0.6999**	0.1000	0.6947	0.1004	**0.6984**	0.0999
DTH4	0.1330	0.1978	0.1547	0.1829	**0.1735**	0.1864
DTH5	0.6901	0.0819	0.6884	0.0819	0.6890	0.0825
DTH8	**0.4882**	0.1268	0.4708	0.1339	0.4657	0.1331
DTH9	**0.6204**	0.1083	0.6159	0.1112	0.6161	0.1072
DTH10	**0.6111**	0.1249	0.6058	0.1219	0.6030	0.1258
DTH11	**0.5909**	0.1396	0.5888	0.1339	0.5832	0.1397
DTH12	**0.5058**	0.1730	0.4989	0.1736	0.4974	0.1702
YLD1	0.5197	0.1186	0.5162	0.1154	0.5132	0.1185
YLD2	0.5066	0.1498	**0.5237**	0.1431	0.5013	0.1488
YLD3	**0.4134**	0.1872	0.3807	0.1970	0.3928	0.1960
YLD4	**0.5475**	0.1216	0.5368	0.1248	**0.5510**	0.1237
YLD5	0.6432	0.1356	0.6410	0.1337	0.6400	0.1387
YLD6	**0.7379**	0.0801	0.7339	0.0777	**0.7368**	0.0821
YLD7	**0.5369**	0.1295	0.5335	0.1286	**0.5408**	0.1335

* Methods significantly (at the 0.05 probability level) best or tied with the best method are in boldface and underlined. Methods significantly worst or tied with the worst method are in plain typeface. Methods significantly worse than the best but significantly better than the worst (i.e., second place methods) are underlined (but not in boldface).

trait-environment combinations, no significant differences among any of the three methods were detected.

Based on PMSE, TGK was the most accurate method together with either RKHS KA or GK REML in four cases and was significantly the solely best predictive method only once (GLS7). GK REML was the best method in three cases (GY-LOW, GLS6, and NCBL2); it tied for the highest position with RKHS KA four times, and with TGK three times. GK REML was a significantly better predictor than TGK in seven cases.

Table 2. Mean Predictive Mean Squared Error (PMSE) and correlation and from 50 random cross-validation partitions (and their standard deviation, SD) of TGK, RKHA KA, and GK REML for various trait-environment combinations of the maize dataset.

Trait-environment	TGK	SD	RKHS KA	SD	GK REML	SD
Mean PMSE						
FFL-WW	0.2570*	0.1257	**0.2190**	0.0909	**0.2233**	0.0942
FFL-SS	**0.3229**	0.0870	**0.3295**	0.0870	0.3234	0.0895
MFL-WW	0.2553	0.1218	**0.2211**	0.0895	**0.2254**	0.0917
MFL-SS	0.3071	0.0933	0.3113	0.0904	0.3056	0.0950
ASI-WW	0.6459	0.2465	0.6499	0.2508	0.6453	0.2464
ASI-SS	0.6460	0.2245	0.6490	0.2281	0.6479	0.2230
GY-SS	0.8880	0.2424	0.8879	0.2436	0.8834	0.2439
GY-WW	**0.6841**	0.1936	0.6925	0.2027	**0.6785**	0.1908
GY-HI	0.5739	0.1485	**0.5710**	0.1477	**0.5670**	0.1474
GY-LOW	0.8536	0.1950	0.8551	0.1940	**0.8483**	0.1965
GLS1	0.9243	0.2193	0.9184	0.2137	0.9212	0.2176
GLS2	0.8116	0.1768	0.8145	0.1718	0.8122	0.1792
GLS3	**0.6299**	0.1597	0.6349	0.1533	**0.6260**	0.1621
GLS4	0.7110	0.1695	0.7125	0.1685	0.7129	0.1701
GLS5	0.8174	0.1834	0.8167	0.1799	0.8173	0.1823
GLS6	0.9674	0.1422	0.9707	0.1418	**0.9654**	0.1452
GLS7	**0.7282**	0.1751	0.7287	0.1716	0.7333	0.1813
GLS8	**0.6243**	0.1611	0.6301	0.1547	**0.6203**	0.1636
GLS9	0.7110	0.1698	0.7121	0.1674	0.7128	0.1700
NCBL1	0.5498	0.1386	**0.5194**	0.1293	**0.5145**	0.1296
NCBL2	0.7050	0.2155	0.7083	0.2172	**0.7046**	0.2139
Mean correlation						
FFL-WW	0.8140	0.2071	**0.8364**	0.1853	0.8235	0.2009
FFL-SS	0.7622	0.1727	**0.7628**	0.1714	0.7604	0.1740
MFL-WW	0.8210	0.1889	**0.8409**	0.1686	0.8287	0.1849
MFL-SS	**0.7810**	0.1532	**0.7817**	0.1523	0.7775	0.1575
ASI-WW	0.5863	0.1576	0.5855	0.1588	0.5854	0.1598
ASI-SS	**0.6211**	0.1162	**0.6212**	0.1153	0.6186	0.1163
GY-SS	0.3240	0.1870	0.3297	0.1881	0.3285	0.1872
GY-WW	0.5520	0.1342	0.5484	0.1351	**0.5558**	0.1352
GY-HI	0.6530	0.0909	**0.6632**	0.0844	**0.6596**	0.0872
GY-LOW	0.4004	0.1295	0.4021	0.1310	0.4048	0.1298
GLS1	0.2615	0.2111	0.2591	0.2045	0.2522	0.1990
GLS2	0.4393	0.1472	0.4386	0.1443	0.4380	0.1469
GLS3	0.5827	0.1335	0.5786	0.1335	**0.5864**	0.1340
GLS4	**0.5423**	0.1370	**0.5439**	0.1367	0.5398	0.1372
GLS5	0.3336	0.1868	0.3318	0.1856	0.3371	0.1837
GLS6	**0.2670**	0.1437	0.2625	0.1465	**0.2696**	0.1449
GLS7	**0.5014**	0.1414	**0.5022**	0.1430	0.4969	0.1473
GLS8	0.5889	0.1335	0.5842	0.1331	**0.5924**	0.1341
GLS9	**0.5420**	0.1371	**0.5438**	0.1374	0.5398	0.1375
NCBL1	0.6793	0.0882	**0.7085**	0.0813	**0.7077**	0.0815
NCBL2	**0.4970**	0.1631	0.4909	0.1683	0.4969	0.1627

*Methods significantly (at the 0.05 probability level) best or tied with the best method are in boldface and underlined. Methods significantly worst or tied with the worst method are in plain typeface. Methods significantly worse than the best but significantly better than the worst (i.e., second place methods) are underlined (but not in boldface).

Based on the correlation criterion, TGK shared the highest ranking position with RKHS KA five times and was significantly the best predictive method for GLS6 and NCBL2 (Table 2). RKHS KA was the best predictive method 10 times, sharing the top position with other models 6 times, and was the best model overall on four occasions. GK REML showed a pattern similar to that of TGK, that is, it was the most significant predictive method six times; it shared the top ranking with other methods for three trait-environment combinations, and in three cases it was the most significant predictive method overall (GY-WW, GLS3, and GLS8).

5. DISCUSSION

The marginal likelihood is important in selecting h and φ can be easily computed after eliminating the nuisance parameters through integration with the joint prior of these parameters (Berger et al. 1999). In this study, the assumption that $p(\mu)$ has a uniform distribution and that $\sigma_a^2 = \varphi\sigma^2$ allows developing an expression such as the one shown in (2.3). However, expression (2.3) still has a nuisance parameter (φ) that is difficult to eliminate by analytical integration. An alternative solution may be Monte Carlo integration, but this could produce an unstable solution.

5.1. SIMILARITIES AND DIFFERENCES BETWEEN REML AND THE PROPOSED BAYESIAN ESTIMATION OF h

It should be pointed out that although we solved the problem of finding estimates of h and φ by maximizing a function of the data, our approach is not the same as Endelman's (2011) strategy. While Endelman (2011) uses the iterative REML approach, our approach simply follows the probability rules of finding the posterior density of h and φ.

Blasco (2001) explained that given the variance components (σ^2, σ_a^2) and h, the BLUP of the vector of random effects $\boldsymbol{b}$ is equal to its conditional posterior mean (mode or median) because the conditional posterior is normal if the prior for μ is flat or normal itself. So, BLUP of $\boldsymbol{b}$ is the posterior mean of $p(\boldsymbol{b}|\sigma^2, \sigma_a^2, h)$ if $\boldsymbol{b}$ is included in the model as a Gaussian random effect with mean $\mathbf{0}$ and variance $\sigma_a^2\boldsymbol{S}^{-1}, \sigma_a^2 = \varphi\sigma^2$ However, σ^2 and σ_a^2 are unknown, and a common way to estimate them, for a given h, is through REML which uses the marginal posterior modes as estimators. However, for both REML and Bayesian approaches, the problem of estimating h remains. It is important to recognize that h is not a variance component, though it is the scale parameter in the Gaussian kernel; that is, h is not the variance of a Gaussian random effect on the linear predictor. Therefore, it cannot be affirmed that the posterior mode of h is equal to the REML estimate, even though the corresponding estimates will be similar as both correspond to the same quantity in the same model.

It is worth mentioning that Endelman (2011) followed the data transformation proposed by Kang et al. (2008), such that $\boldsymbol{V}'\boldsymbol{y} = (u_1, u_2, \ldots, u_n)'$, where $\boldsymbol{V}$ and $\lambda_i + \varphi^{-1}$ are the $n \times n$ matrix of eigenvectors and the ith eigenvalue of the spectral decomposition of $\boldsymbol{SHS}$, respectively, with $\boldsymbol{S} = \boldsymbol{I} - \mathbf{1}\mathbf{1}'/n$ and $\boldsymbol{H} = \boldsymbol{K}_h + \varphi^{-1}\boldsymbol{I}$. Then, the restricted likelihood

(RL) to be optimized for computing the REML estimate of h is given by

$$\mathrm{RL} = \frac{1}{2}\left[(n-1)\log(2\pi) - \log n - \sum_{i=1}^{n}\log\left(\lambda_i + \varphi^{-1}\right) - Q/\left(\varphi\sigma^2\right)\right]$$

with $Q = \sum_{i=1}^{n}\left(\frac{u_i^2}{\lambda_i+\varphi^{-1}}\right)$. Note that as expression (2.4), RL also implicitly depends on h. However, the main difference between RL and (2.4) is that RL is a function of σ^2, while $m(\boldsymbol{y}|h,\varphi)$ is not. In Endelman's method, σ^2 is assumed to be equal to $\hat{\sigma}^2 = \hat{\sigma}_a^2/\hat{\varphi}$, where $\hat{\sigma}_a^2$ and $\hat{\varphi}$ are REML estimations, and this value is plugged into RL. On the other hand, in our Bayesian proposal, the function to be maximized is the marginal posterior density of h, $p(h|\boldsymbol{y},\varphi) \propto m(\boldsymbol{y}|h,\varphi)\,p(h)$, which by definition does not depend on σ^2.

5.2. The Prior Distribution of h

The prior distribution of h may dominate the likelihood and its correct selection depends on the researcher's knowledge of h for a particular trait. This previous knowledge of the values of h for certain traits can be used a priori to avoid using large values of h that may cause over-fitting, as suggested by Härdle (1990). It is important to note that large values of h generate a $\boldsymbol{K}_h$ close to the identity matrix, with the obvious consequence of generating confusion in the signal/noise ratio. Furthermore, low values of h may produce $\boldsymbol{K}_h$ matrices with all off-diagonal entries close to 1, which causes serious problems when fitting the model.

The correct selection of $p(h)$ influences the signal/noise ratio, and a flat prior is suggested when nothing is known about h. With prior information, we suggest using a Gamma distribution as a prior for h, because this distribution is flexible enough to represent the prior knowledge of h. The parameters of the Gamma distribution may be selected by exploring matrix $\boldsymbol{K}_h$ with different values of h and previous knowledge of the range of h values.

We used a Gaussian kernel of the form $K_{ij}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right) = \exp\left(-hd_{ij}^2/q_{0.05}\right)$, where d_{ij}^2 is the squared Euclidean distance (González-Camacho et al. 2012), and $q_{0.05}$ is the 0.05 quantile of this squared distance. The value $q_{0.05}$ is used to scale the values of the kernel such that the values of h close to zero will originate a matrix $\boldsymbol{K}_h$ near to one and large values of h will tend to produce an identity matrix for $\boldsymbol{K}_h$. For example, for markers of the wheat data, Fig. 1 depicts the histograms of the upper off-diagonal elements of the Gaussian kernel for three values of h. Figure 1a shows a high frequency of high values of $\boldsymbol{K}_h$ for $h = 0.2$, indicating a global bandwidth; Fig. 1c shows a high frequency of low values of $\boldsymbol{K}_h$ for $h = 6$, and Fig. 1b, with $h = 1$, shows intermediate off-diagonal values of $\boldsymbol{K}_h$. It seems reasonable to think that there is a high probability that a Gamma distribution with mean 1 or 2 covers the range of values of h that generates intermediate values for the off-diagonal elements of $\boldsymbol{K}_h$.

For the application of TGK in the maize and wheat datasets, we considered a Gamma distribution with $\nu = 3$, and a scale parameter $\tau = 1.5$. The R program that we provide in the Supplemental Material allows a researcher to choose a flat prior or proper Gamma distribution for h with scale parameters that the researcher considers to be appropriate.

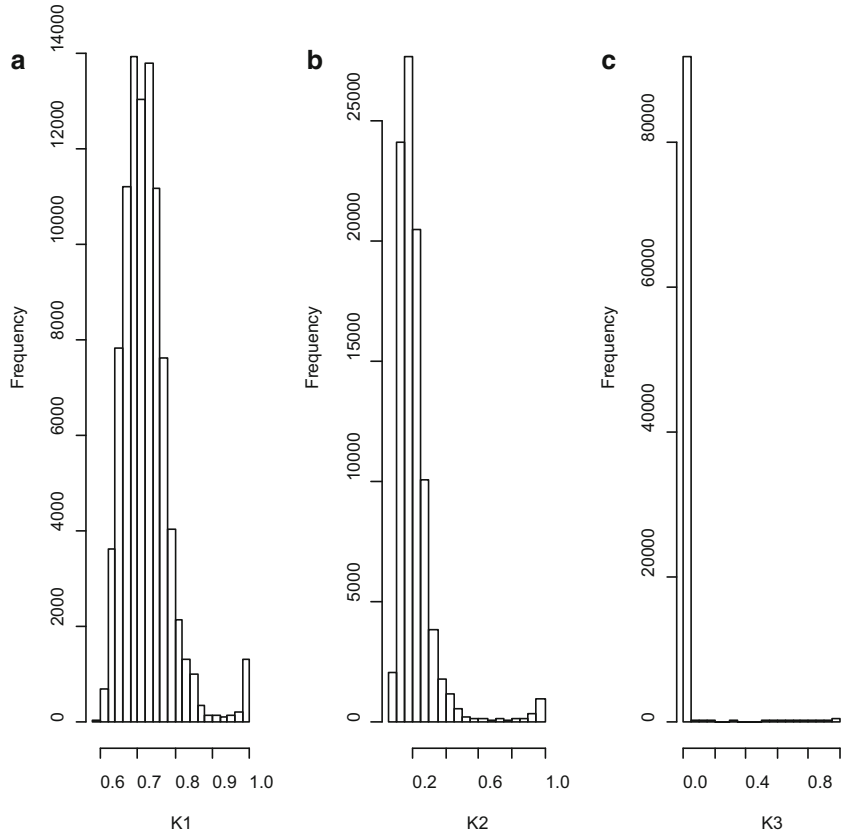

Figure 1. **a–c** Histograms of the off-diagonal entries for three kernels in the wheat dataset: **a K1** with $h = 0.2$; **b K2** with $h = 1$, and **c K3** with $h = 6$.

5.3. PREDICTION ACCURACY OF THE WHEAT DATASET

The wheat data may have genetic complexities that favor semi-parametric regression over parametric regression, as clearly shown by Pérez-Rodríguez et al. (2012); some of these complexities may be due to cryptic gene × gene epistatic effects that might be only captured by a non-additive model. In general, TGK had better prediction accuracy more often than RKHS KA and GK REML for DTH and YLD wheat datasets. Supplemental Table S1 shows the estimated values of h and φ for the full data using TGK2 with a Gamma prior with mean 2, and RKHS KA and GK REML with a Gaussian kernel of the form $K = \exp\left(-\frac{d_{ij}^2}{h^2}\right)$. Also, Supplemental Table S1 shows that in almost all cases, the residual variance from GK REML is smaller than those from the other methods and therefore produces larger R. On the other hand, RHKS KA tended to have a smaller R than the other methods, whereas TGK had intermediate values of R that were closer to those estimated by GK REML than those estimated by RHKS KA. This is related to the generally better prediction accuracy of TGK2 for the wheat dataset. An exception is trait DTH4 with a very low signal/noise ratio; the posterior of h is dominated by the prior distribution and thus the posterior mode is close to the prior mean (close to 2).

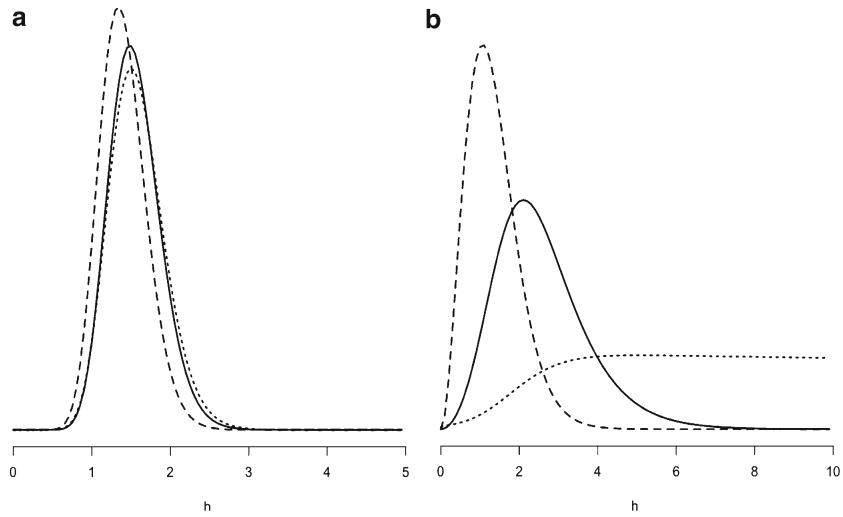

Figure 2. Graphs **a** and **b** depict the behavior of the posterior of h for wheat data DTH2 and DTH4, respectively. In each graph, three priors of h are used: proper flat prior distribution (*continuous line*), Gamma distribution with $\nu = 3, \tau = 1.5$ (*dashed line*) and Gamma distribution with $\nu = 3, \tau = 3$ (*dotted line*).

Depicted in Fig. 2a, b is the effect on the posterior of three different prior specifications for h given that $\varphi = \hat{\varphi}$, which is approximately equal to the REML estimate. Any other proper prior on $\mathbb{R}^+$ could certainly be used, but we used a Gamma prior in this study due mainly to practical considerations and mathematical convenience. The values of h in TGK2 (Supplemental Table S1) are intermediate, with moderate cross terms of $\boldsymbol{K}_h$ similar to those observed in Fig. 2b; it can thus be assumed that the selection of the prior is adequate. Figure 2a, b depict the behavior of the posterior distribution for two trait-environment combinations of the wheat dataset (DTH2 and DTH4). For trait-environment combination DTH2, Fig. 2a shows the posterior distribution of h with the Gamma prior distribution with mean 2 (continuous line) and the Gamma prior distribution with mean 1 (dotted line). These results indicate that for the DTH2 wheat dataset, the posterior is not sensitive to the prior selection. Note in Supplemental Table S1 that the R^2 for TGK2 of this trait-environment combination indicates a relatively high signal/noise rate.

On the other hand, Fig. 2b for trait-environment combination DTH4 shows the posterior distribution of h with the Gamma prior distribution with mean 2 (continuous line) and the Gamma prior distribution with mean 1 (dotted line). These results indicate that the posterior distribution of h is very sensitive to the prior distribution selection because the likelihood function does not give enough information about h and the remaining parameters of the kernel regression model (see R in Supplemental Table S1). In this case, the selection of an appropriate prior distribution of h will depend strongly on the prior knowledge about the value of h for this trait.

5.4. PREDICTION ACCURACY OF THE MAIZE DATASET

For the maize datasets, the prediction accuracy of TGK was statistically similar or lower than the prediction accuracy of GK REML; for seven trait-environment combinations, GK

REML was statistically better than TGK. Supplemental Table S2 shows the information after fitting TGK2, RKHS KA and GKREML with the full data. The pattern of results is similar to the patterns in the wheat dataset, with the residual variance lower in GK REML and higher in RKHS KA. The three models (TGK2, RKHS KA and GK REML) had very similar prediction accuracies, except in some specific trait-environment combinations. For example, flowering traits FFL-WW and MFL-WW, with the highest values of h (3.8914 and 3.5055), showed poor prediction performance under the TGK model, probably due to the bimodal distribution of the phenotypic data. Flowering traits such as FFL and MLF do not have a very complex genetic architecture and this may be another reason why the TGK2 model did not perform as well as the other models. Also, low values of h (0.2 and 0.4) produced a more global kernel and thus lowered the prediction ability of model TGK. Another case is NCBL1, where the values of h are low (0.2). However, despite the fact that the correlations of TGK2 were not all higher than those of the other two models, the PMSE are generally lower for methods GK REML and RKHS than for model TGK.

6. CONCLUSIONS

The proposed method for estimating h is based on a Bayesian formulation in which the selected value of h is a posterior point estimate of h, namely, the posterior mode $\tilde{h}$ of the marginal distribution $p(h|\boldsymbol{y}, \varphi)$. Then, conditional on $\tilde{h}$, the TGK model is fitted and used for prediction. The selection of h produced good prediction accuracy of unobserved individuals. When the proposed selection of h is combined with model TGK, the computational limitations derived from the high dimensionality are removed, as well as the poor mixing of the MCMC typical of most Bayesian genomic selection procedures.

The Gamma distribution used as a prior for estimating h penalizes the high values of the bandwidth parameter. To elicit the Gamma prior for h, we fixed the form parameter (φ) and either the mean or the mode may be selected such that K_h does not tend towards the identity matrix or, at the other extreme, to a K_h matrix with all its entries equal to one. In the application for wheat and maize data, a mean of 2 was used, which represents moderate values of the cross terms in $\boldsymbol{K}_h$.

When the proposed model (TGK) was applied to the wheat data, it indicated superiority in prediction accuracy over the RKHS KA and GK REML models. In the maize data, no superiority was achieved in some trait-environment combinations, due to the fact that some traits were bimodal and had different genetic architecture. Significant differences between models were less clear for the maize data; model GK REML had statistically lower PMSE than model TGK in seven cases, and, in general, both GK REML and RKHS KA had lower PMSE than model TGK.

ACKNOWLEDGEMENTS

We thank the CIMMYT field assistants and technicians as well as the national program researchers who collected the data used in this study. We also thank the anonymous reviewers and the Guest Editors of JABES for the time and effort they invested in correcting and improving the quality of the manuscript.

[Received December 2014. Accepted September 2015. Published Online October 2015.]

APPENDIX

Assume that uniform distribution is assigned as the prior of μ, that is, $p(\mu) \propto 1$; Berger et al. (1999) justify the use of non-informative prior distributions to obtain the predictive distributions in the context of model selection.

Let us first integrate the likelihood with respect to μ, such that the integrated likelihood for $(h, \boldsymbol{b}, \sigma^2, \varphi)$ is given by

$$\begin{aligned} L_{(h,\boldsymbol{b},\sigma^2,\varphi)}\left(h, \boldsymbol{b}, \sigma^2, \varphi|\boldsymbol{y}\right) &= \int_{-\infty}^{\infty} p\left(\boldsymbol{d}|\mu, \boldsymbol{b}, \sigma^2, h, \varphi\right) p\left(\mu, \boldsymbol{b}, \sigma^2|\varphi\right) \mathrm{d}\mu \\ &\propto p\left(\boldsymbol{b}|\sigma^2, \varphi\right) p\left(\sigma^2\right) \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}} \\ &\quad \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{d} - \mu\boldsymbol{t} - \boldsymbol{S_h b})'\,(d - \mu\boldsymbol{t} - \boldsymbol{S}_h\boldsymbol{b})\right\} \mathrm{d}\mu \end{aligned} \tag{6.1}$$

Define $\tilde{\boldsymbol{d}} = \boldsymbol{U}_h'\,(\boldsymbol{y} - \mathbf{1}\bar{y}) = \boldsymbol{d} - \boldsymbol{t}\bar{y}$, where $\bar{y}$ is the arithmetic mean of the elements of $\boldsymbol{y}$, such that

$$(\boldsymbol{d} - \mu\boldsymbol{t} - \boldsymbol{S}_h\boldsymbol{b})'(\boldsymbol{d} - \mu\boldsymbol{t} - \boldsymbol{S}_h\boldsymbol{b}) = n(\bar{y} - \mu)^2 + (\tilde{\boldsymbol{d}} - \mu\boldsymbol{t} - \boldsymbol{S}_h\boldsymbol{b})'(\tilde{\boldsymbol{d}} - \mu\boldsymbol{t} - \boldsymbol{S}_h\boldsymbol{b}),$$

Then, completing the normal density for μ in the integrand of (6.1), we have

$$L_{(h,\boldsymbol{b},\sigma^2,\varphi)}\left(h, \boldsymbol{b}, \sigma^2, \varphi|\boldsymbol{y}\right) \propto p\left(\boldsymbol{b}|\sigma^2, \varphi\right) p\left(\sigma^2\right) \left(\sigma^2\right)^{-(n-1)/2} N\left(\tilde{\boldsymbol{d}}|\boldsymbol{S}_h\boldsymbol{b}, \sigma^2\boldsymbol{I}\right) \tag{6.2}$$

To integrate expression (6.2) with respect to $\boldsymbol{b}$, recall that $\boldsymbol{b} = \boldsymbol{U}_h'\boldsymbol{f}$ and $\boldsymbol{K}_h = \boldsymbol{U}_h\boldsymbol{S}_h\boldsymbol{U}_h'$. Then if we assign the normal distribution $N\left(\mathbf{0}, \sigma^2\varphi\boldsymbol{K}_h^{-1}\right)$ as the prior of $\boldsymbol{f}$, the prior of $\boldsymbol{b}$

is $N\left(\mathbf{0}, \varphi\sigma^2 \boldsymbol{S}_h^{-\mathbf{1}}\right)$. It follows that the marginal likelihood of (h, φ, σ^2) is

$$
\begin{aligned}
L_{(h,\sigma^2,\varphi)}\left(h, \sigma^2, \varphi|\boldsymbol{y}\right) &= \int_{\mathbb{R}^n} L_{(h,\boldsymbol{b},\sigma^2,\varphi)}\left(h, \boldsymbol{b}, \sigma^2, \varphi|\boldsymbol{y}\right) N(\boldsymbol{b}|\mathbf{0}, \varphi\sigma^2 \boldsymbol{S}_h^{-1})\mathrm{d}\boldsymbol{b} \\
&\propto \left(\sigma^2\right)^{-(n-1)/2} p\left(\sigma^2\right) \int_{\mathbb{R}^n} N\left(\tilde{\boldsymbol{d}}|\boldsymbol{S}_h\boldsymbol{b}, \sigma^2\boldsymbol{I}\right) N(\boldsymbol{b}|\mathbf{0}, \varphi\sigma^2 S_h^{-1})\mathrm{d}\boldsymbol{b} \\
&\propto (\sigma^2)^{1/2} |\boldsymbol{S}_h|^{1/2} \left|\boldsymbol{S}_h^2 + \boldsymbol{S}_h/\varphi\right|^{-1/2} p\left(\sigma^2\right) N(\tilde{\boldsymbol{d}}|\boldsymbol{S}_h\tilde{\boldsymbol{b}}_h, \sigma^2\boldsymbol{I})
\end{aligned}
$$

where $\tilde{\boldsymbol{b}}_h = \left(\boldsymbol{S}_h^2 + \boldsymbol{S}_h/\varphi\right)^{-1} \boldsymbol{S}_h'\tilde{\boldsymbol{d}}$ is the conditional posterior mean of $\mathbf{b}$.

Finally, if we assign the prior $\chi^{-2}(\sigma^2|\upsilon_\varepsilon, \tau_\varepsilon)$ to σ^2, then

$$
\begin{aligned}
m(\boldsymbol{y}|h, \varphi) &\propto \int_0^\infty L_{(h,\sigma^2,\varphi)}\left(h, \boldsymbol{b}, \sigma^2, \varphi|\boldsymbol{d}\right) \chi^{-2}\left(\sigma^2|\upsilon_\varepsilon, \tau_\varepsilon\right) \mathrm{d}\sigma^2 \\
&\propto |\boldsymbol{S}_h|^{1/2} \left|\boldsymbol{S}_h^2 + \boldsymbol{S}_h/\varphi\right|^{-1/2} \int_0^\infty (\sigma^2)^{1/2} N(\tilde{\boldsymbol{d}}|\boldsymbol{S}_h\tilde{\boldsymbol{b}}_h, \sigma^2\boldsymbol{I}) \chi^{-2}(\sigma^2|\upsilon_\varepsilon, \tau_\varepsilon)\mathrm{d}\sigma^2
\end{aligned}
$$

Then, the marginal likelihood is:

$$
m\left(\boldsymbol{y}|h, \varphi\right) \propto |\boldsymbol{S}_h|^{1/2} \left|\boldsymbol{S}_h^2 + \boldsymbol{S}_h/\varphi\right|^{-1/2} \left[\tau_\varepsilon + \left(\tilde{\boldsymbol{d}} - \boldsymbol{S}_h\tilde{\boldsymbol{b}}_h\right)' \left(\tilde{\boldsymbol{d}} - \boldsymbol{S}_h\tilde{\boldsymbol{b}}_h\right)\right]^{-(\upsilon_\varepsilon+n-1)/2}
$$

or

$$
m\left(\boldsymbol{y}|h, \varphi\right) \propto \prod_{i=1}^n (1 + \varphi s_i)^{-\frac{1}{2}} \left[\tau_\varepsilon + \sum_{i=1}^n \frac{\tilde{d}_i^2}{(1 + \varphi s_i)}\right]^{-\frac{\upsilon_\varepsilon+n-1}{2}}
$$

REFERENCES

Aster, R., Borchers, B., Thurber, C. (2005). Parameter estimation and inverse problems. Elsevier Academic Press. New York.

Bellman, R. E. (1961). Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton, NJ.

Berger, J. O., Liseo, B., Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Sci.* **14**(1): 1-28.

Blasco, A. 2001. The Bayesian controversy in animal breeding. *Journal of Animal Breeding* 79:2023-2046.

Cavalier, L. (2008). Non-parametric statistical inverse problems. *Inverse Problems* 24, doi:10.1088/24/3/0034004.

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., Araus, J., Makumbi, D., Singh, R., Dreisigacker, S., Yan, J., Arief, V., Bänziger, M., Braun, H. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. DOI:10.1534/genetics10.118521.

Cuevas, J., Pérez-Elizalde, S., Soberanis, V., Pérez-Rodríguez, P., Gianola, D., Crossa, J. (2014). Bayesian genomic-enabled prediction as an inverse problem. *G3/Genes/Genome/Genetics* 4, 1991-2001. doi:10.1534/g3.114.013094.

de los Campos, G., Gianola, G., Rosa, G. J. M., Weigel, K. A., Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert space methods. *Genet. Res.* **92**(4): 295-308.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., Calus, M. P. L. (2012). Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* doi:10.1534/genetics.112.143313.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **4**(3):250-255.

Gelfand, A., Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85(410): 398-409.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457-511.

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *J. Mach. Learn.* **2**: 299-312.

Gianola, D., Fernando, R., Stella, A. (2006). Genomic-assisted prediction of genetic value with a semi-parametric procedure. *Genetics* **173**(3): 1761-1776.

Gianola, D., van Kaam, J. B. C. H. M. (2008). Reproducing Kernel Hilbert Space Regression Methods for Genomic-Assisted Prediction of Quantitative Traits. *Genetics* **178**(4): 2289-2303.

Gianola, D., Okut, H., Weigel, K. A., Rosa, G. J. M. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics* doi:10.1186/1471-2156-12-87.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics* 113.151753; Early online May 1, 2013, doi:10.1534/genetics.113.151753.

Gianola, D., Morota, G., Crossa, J. (2014). Genome-enabled prediction of complex traits with kernel methods: What have we learned? Proceedings, 10th World Congress Applied to Livestock Production. August 17-22, Vancouver, BC, Canada.

González-Camacho, J. M., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Cairns, J. E, Mahuku, G., Babu, R., Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function. *Theor. Appl. Genet.* **125**:759-771.

Harville, D. (1974). Bayesian inference for various variance components using only error contrasts. *Biometrika* **61**:383-385.

Härdle, B. W. (1990). Applied non parametric regression. Cambridge, U.K.: Cambridge University Press.

Heslot, N., Yang, H-P., Sorrells, E. M., Jannink, J. L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Sci.* **52**: 146-160.

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

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. *G3: Genes|Genetics* **4**, 1027-1046.

Kang, H.M., Zaitle, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**:1709-1723.

Kimeldorf, G., Wahba, G. (1971). Some results on Tchebycheffian Spline Functions. *Journal Mathematical Analysis and Applications* **33**(1): 82-95.

Knapick, B.T, van der Vaart, A.W., van Zanten, J.H. (2012). Bayesian inverse problems with Gaussian priors. *Annals of Statistics* **39**(5): 2626-2657. doi:10.1214/11-AOS920.

Maruyama, Y., George, E. I. (2011). Fully Bayes factors with generalized g-priors. *Annals of Statistics* 39(5): 2740-2765.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819-1829.

Morota, G., Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics* **5**: 1-13. doi:10.3389/fgene.2014.00363.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., Simianer, H. (2011). Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics* **188**: 695-708. doi:10.1534/genetics.111.128694.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manes, Y., Dreisigacker, S. (2012). Comparison between linear and non-parametric models for genome-enabled prediction in wheat. *G3/Genes/Genome/Genetics* **2**:1595-1605.

Pérez-Rodríguez, P., de los Campos, G. (2014). Genome Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198**: 483-495.

Plummer, M., Best, N., Cowles, K., Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* 6, 7-11.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raftery, A. E., Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science* **7**:493-497.

Tussel, L., Pérez-Rodríguez, P., Forni, S., Gianola, D. (2014). Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. *J. Anim. Breed. Genet.* 131:105-115.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91: 4414-4423.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, et al. (2004) Diversity arrays technology (DArT) for whole genome profiling of barley. Proc Natl Acad Sci (USA) **101**: 9915–9920. doi:10.1073/pnas.0401076101.

Zhou, X., Carbonetto, P., Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *Plos Genetics* **9**, e1003264.