frontiers in Genetics

METHODS
published: 09 August 2016
doi: 10.3389/fgene.2016.00145

A Unified and Comprehensible View of Parametric and Kernel Methods for Genomic Prediction with Application to Rice

*Laval Jacquin *, Tuong-Vi Cao and Nourollah Ahmadi*

Centre de Coopération Internationale en Recherche Agronomique pour le Développement, BIOS, UMR AGAP, Montpellier, France

One objective of this study was to provide readers with a clear and unified understanding of parametric statistical and kernel methods, used for genomic prediction, and to compare some of these in the context of rice breeding for quantitative traits. Furthermore, another objective was to provide a simple and user-friendly R package, named KRMM, which allows users to perform RKHS regression with several kernels. After introducing the concept of regularized empirical risk minimization, the connections between well-known parametric and kernel methods such as Ridge regression [i.e., genomic best linear unbiased predictor (GBLUP)] and reproducing kernel Hilbert space (RKHS) regression were reviewed. Ridge regression was then reformulated so as to show and emphasize the advantage of the kernel "trick" concept, exploited by kernel methods in the context of epistatic genetic architectures, over parametric frameworks used by conventional methods. Some parametric and kernel methods; least absolute shrinkage and selection operator (LASSO), GBLUP, support vector machine regression (SVR) and RKHS regression were thereupon compared for their genomic predictive ability in the context of rice breeding using three real data sets. Among the compared methods, RKHS regression and SVR were often the most accurate methods for prediction followed by GBLUP and LASSO. An R function which allows users to perform RR-BLUP of marker effects, GBLUP and RKHS regression, with a Gaussian, Laplacian, polynomial or ANOVA kernel, in a reasonable computation time has been developed. Moreover, a modified version of this function, which allows users to tune kernels for RKHS regression, has also been developed and parallelized for HPC Linux clusters. The corresponding KRMM package and all scripts have been made publicly available.

Keywords: genomic prediction, parametric, semi-parametric, non-parametric, kernel "trick", epistasis

OPEN ACCESS

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Ashok Ragavendran, Massachusetts General Hospital, USA
Li Zhang, University of California, San Francisco, USA

****Correspondence:***
Laval Jacquin
laval.jacquin@cirad.fr

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 May 2016*
Accepted: *26 July 2016*
Published: *09 August 2016*

Citation:
Jacquin L, Cao T-V and Ahmadi N (2016) A Unified and Comprehensible View of Parametric and Kernel Methods for Genomic Prediction with Application to Rice. Front. Genet. 7:145. doi: 10.3389/fgene.2016.00145

1. INTRODUCTION

Since the seminal contribution of Meuwissen et al. (2001), genomic selection (GS) has become a popular strategy for genetic improvement of livestock species and plants. Moreover numerous methods from statistics and machine learning have been proposed for genomic prediction since, due to the high modeling complexity associated to the large amount of markers available. For instance, modeling the effects of thousands interacting genes (i.e., epistasis) associated to complex

quantitative traits is not trivial. There is an increasing number of studies supporting that epistasis may be the most prevalent form of genetic architecture for quantitative traits (Flint and Mackay, 2009; Moore and Williams, 2009; Huang et al., 2012). Hence genomic prediction methods which can account for epistatic genetic architectures have been proposed. For example, Gianola et al. (2006) and Gianola and van Kaam (2008) first proposed reproducing kernel Hilbert space (RKHS) regression for genomic prediction when dealing with epistatic genetic architectures. Later Howard et al. (2014) showed that RKHS and support vector machine regression (SVR), when dealing with an additive genetic architecture, could be almost as competitive as parametric methods such as best linear unbiased predictor (BLUP), least absolute shrinkage and selection operator (LASSO) or Bayesian linear regressions (Bayes A, Bayes B, Bayes C, Bayes Cπ, and Bayesian LASSO). These authors also showed that RKHS regression and SVR, with some other non parametric methods, clearly outperformed parametric methods for an epistatic genetic architecture.

The SVR and kernel Ridge regression (abusively called RKHS regression in this paper with respect to previous studies (Konstantinov and Hayes, 2010; Howard et al., 2014) are popular methods known as kernel methods in the machine learning community (Cristianini and Shawe-Taylor, 2000), while they are commonly and respectively referred to as non-parametric and semi-parametric methods in statistics. Like kernel Ridge regression, SVR also performs regularization in a RKHS and this explains why kernel Ridge regression is somehow abusively called RKHS regression. Nevertheless, the term RKHS regression for kernel Ridge regression will be used in this paper so as to remain consistent with previous studies. For RKHS regression, a part of the model can be specified parametrically with fixed effects and this explains why it is also called semi-parametric regression.

There is an increasing number of studies, based on either real or simulated data, showing that kernel methods can be more appropriate than parametric methods for genomic prediction in many situations (Konstantinov and Hayes, 2010; Pérez-Rodríguez et al., 2012; Sun et al., 2012; Howard et al., 2014). In a recent review study, Morota and Gianola (2014) conjectured that RKHS regression is at least as good as linear additive models whether non-additive or additive effects are the main source of genetic variation. Their conjecture came from a series of comparison between parametric and kernel methods based on several real data sets, especially in plant breeding. The main difference between kernel and parametric methods rely in model assumptions and functional form specification. For example, SVR or RKHS regression can account for complex epistatic genetic architectures without explicitly modeling them, i.e., the model is data-driven and hence there is no pre-specified functional form relating covariates to the response (Howard et al., 2014). On the other hand classical linear regression, which is a parametric method, rely on a pre-specified functional relationship between covariates and the response.

One objective of this paper is to provide readers with a clear and unified understanding of conventional parametric and kernel methods, used for genomic prediction, and to compare some of these in the context of rice breeding for quantitative traits. Another objective is to provide an R package named `KRMM` which allows users to perform RKHS regression with several kernels. The first part of the paper reviews the concept of regularized empirical risk minimization as a classical formulation of learning problems for prediction. The second part reviews the equivalence between some well-known regularized linear models, such as Ridge regression and LASSO, and their Bayesian formulations. The main objective of this part is to highlight the equivalences between Ridge regression, Bayesian Ridge regression, random regression BLUP (RR-BLUP) and genomic BLUP (GBLUP), through the connections between regularized, Bayesian and mixed model regressions within the Ridge regression framework. These equivalences are important in order to understand the reformulation of Ridge regression in terms of kernel functions known as the dual formulation (Saunders et al., 1998).

In the third part we use the dual formulation of Ridge regression in order to explain and emphasize the RKHS regression methodology, in the context of epistatic genetic architectures, by the use of the so-called kernel "trick". To our best knowledge, and according to Jiang and Reif (2015), it has not been well clarified how RKHS regression can capture multiple orders of interaction between markers and we aim at providing a simple and clear explanation to this. Jiang and Reif (2015) gave an excellent explanation on how RKHS regression, based on Gaussian kernels, can capture epistatic effects. Nevertheless, our approach is different from these authors in the sense that it is directly motivated from the kernel "trick" perspective, and hence we did not restricted ourselves to Gaussian kernels. Moreover, we used a simpler kernel function (than the Gaussian kernel) in order to give a simple and clear explanation on how RKHS regression can capture multiple orders of interaction between markers.

In the fourth part we show that solutions to many parametric and machine learning problems have similar form, due to the so-called representer theorem (Kimeldorf and Wahba, 1971), and that these solutions differ only in the choice of the loss and kernel functions used for the regularized empirical risk. We show that many parametric methods can be framed as machine learning methods with simple kernels. In the last part we compare four methods which are LASSO, GBLUP, SVR, and RKHS regression for their genomic predictive ability in the context of rice breeding using three real data sets. Finally, we provide a simple and user-friendly R function, and a tuned and parallelized version of the latter, which allow users to perform RR-BLUP of marker effects, GBLUP and RKHS regression with a Gaussian, Laplacian, polynomial or ANOVA kernel. The corresponding `KRMM` package and all scripts have been made publicly available at https://sourceforge.net/u/ljacquin/profile/.

2. MATERIALS AND METHODS

2.1. Regularized Empirical Risk Minimization (RERM)

Here we review RERM as a classical formulation of learning problems for prediction. For simplicity reason, we consider

a motivating example to RERM problems only in the linear regression framework.

2.1.1. Classical Formulation of RERM Problems

Many statistical and machine learning problems for prediction are often formulated as follows:

$$\hat{f}(.) = \underset{f \in \mathcal{H}}{argmin}\{ \underbrace{\mathbb{E}[||Y - f(X)||_2^2]}_{\text{Empirical risk term } (T_1)} + \underbrace{\lambda |f|_{\mathcal{H}}}_{\substack{\text{Regularization term } (T_2), \\ \text{i.e., "penalty"}}} \} \quad (1)$$

where $(Y, X) = (Y_i, X_i)_{1 \leq i \leq n}$ are n independent and identically distributed (i.i.d.) data samples, according to the joint distribution of (Y, X), and f is a functional relating Y and X. $\mathcal{H}$ corresponds to a Hilbert space and we can take $\mathcal{H} = \mathbb{R}^p$ for example in the finite dimensional case, which is the Euclidean space, if f is a linear functional. In term T_2, $||.||_{\mathcal{H}}$ is a mathematical norm defined over $\mathcal{H}$. For $a \in \mathcal{H} = \mathbb{R}^p$, we can define $||a||_{\mathcal{H}} = ||a||_q = (\sum_{i=1}^{p} |a_i|^q)^{\frac{1}{q}}$ which is the L^q norm for example. In Expression (1) $\hat{f}(.)$ corresponds to a functional (i.e., "model") minimizing simultaneously T_1 and T_2 over $\mathcal{H}$. Note that the uniqueness of $\hat{f}(.)$ depends on the norm used in T_2 and the sizes of n and p. Term T_2 is called the regularization (or penalization) term which has a tuning parameter λ controlling the "size" of f (i.e., model complexity). Term T_1 is called the empirical risk and corresponds, for some loss function, to the expected (i.e., $\mathbb{E}[.]$) data prediction error which can be estimated using the empirical mean by the weak law of large numbers (Cornuéjols and Miclet, 2011). A common choice for the loss function is the squared L^2 norm (i.e., $||.||_q^2$ with $q = 2$), even though other choices such as the L^1 norm, or the ε-insensitive loss like in the case of SVR (Smola and Schlkopf, 1998), are possible. Finally, finding the solution to Expression (1) is known as a RERM problem.

2.1.2. A Motivating Example for RERM Problems

Here we review the motivation behind RERM problems within the classical linear regression framework for the sake of simplicity. Assume that we have a functional relationship $Y = f^*(X) + \varepsilon^*$, where $Y = [Y_1, .., Y_i, .., Y_n]$ is a vector of n measured phenotypic responses, $X = (X_i)_{1 \leq i \leq n}$ is an n x p marker genotype matrix with $X_i = [X_i^{(1)}, X_i^{(2)}, .., X_i^{(j)}, .., X_i^{(p)}] \in \mathbb{R}^p$ (i.e., genotypes at p markers for individual i) and $\varepsilon^* = [\varepsilon_1^*, \varepsilon_2^*, .., \varepsilon_i^*, .., \varepsilon_n^*]'$ is the error vector of n i.i.d elements with $\mathbb{E}[\varepsilon_i^*] = 0$ and $\mathbb{V}ar[\varepsilon_i^*] = \sigma_{\varepsilon^*}^2 > 0$, where $\sigma_{\varepsilon^*}^2$ is unknown. $f^*(.)$ can be interpreted as the "true" deterministic model, or the data generating process (DGP), generating the true genetic values of individuals. Note that we do not assume gaussianity for ε^* here. Our aim is to identify a model with linear regression that best approximates $f^*(.)$. Consider the following linear model with full rank X ($\Rightarrow p \leq n$):

$$Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + ... + \beta_j X_i^{(j)} + ... + \beta_p X_i^{(p)} + \varepsilon_i = f_p(X_i) + \varepsilon_i \text{ where } f_p(X_i) = \sum_{j=1}^{p} \beta_j X_i^{(j)} \quad (2)$$

In matrix notation we can write the model defined by Equation (2) as $Y = f_p(X) + \varepsilon = X\beta + \varepsilon$ where $\beta = [\beta_1, \beta_2, .., \beta_j, .., \beta_p]'$. By ordinary least squares (OLS), the estimated model for Equation (2) is given by $\hat{f}_p(X) = X\hat{\beta}_{OLS}$, where $\hat{\beta}_{OLS}$ is the unique minimizer of $||Y - X\beta||_2^2 = ||\varepsilon||_2^2$ (which is strictly convex and quadratic) and is given by $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. For this estimated model we have the following property which holds (see Supplementary Material, lemma 4):

$$\underbrace{\mathbb{E}[\, ||\hat{f}_p(X) - f^*(X)||_2^2\,]}_{\substack{\text{Risk of the model, i.e., distance} \\ \text{between estimated model} \\ \text{and true model } (R_1)}} = \underbrace{\mathbb{E}[\, ||Y - \hat{f}_p(X)||_2^2\,]}_{\text{Empirical risk term } (R_2)} + \underbrace{2\sigma_{\varepsilon^*}^2 p}_{\substack{\text{Term with dependence} \\ \text{on number of} \\ \text{parameters } (R_3)}} \underbrace{-\sigma_{\varepsilon^*}^2 n}_{(R_4)} \quad (3)$$

For a fixed sample size n, from Equation (3) we clearly see that $R_2 \to 0$ and $R_3 \to +\infty$ when $p \to +\infty$. This situation is common in genomic prediction where p is often much bigger than n, i.e., $p >> n$. Precisely, we have $R_2 = 0$ (i.e., $Y = \hat{f}_p(X)$) when $p \geq n$. This is due to the fact that $\hat{f}_p(X)$ is the orthogonal projection of Y on the subspace of $\mathbb{R}^n$, generated by columns of X, which becomes $\mathbb{R}^n$ when $p \geq n$ (see Supplementary Material, lemma 5). This phenomena is a case of what is known as overfitting since the estimated model reproduces the data, which contains the error term, and is not describing the underlying relation defined by $f^*(.)$. Note that R_4 is unaffected by p for a fixed n. Hence, if we want to decrease the distance between the estimated model and the true model (i.e., R_1), we need to minimize simultaneously R_2 and R_3 and this motivates the RERM formulation seen in Equation (1). Note that minimizing R_3 (i.e., decreasing p), with R_2 simultaneously, will penalize model complexity, and size, and this explains why a regularization term is also called a penalization term.

2.2. Equivalence of Regularized and Bayesian Formulations of Ridge and LASSO Regressions

In what follows, we assume that matrix X and vector Y are centered. Two popular examples of regularized linear regressions are Ridge regression (Hoerl and Kennard, 1970) and LASSO (Tibshirani, 1996). These (estimated) models and their regularized estimates are given by:

$$\hat{f}(X)_{Ridge} = X\hat{\beta}_{Ridge}$$
$$\text{where } \hat{\beta}_{Ridge} = \underset{\beta \in \mathbb{R}^p}{argmin}\{\, ||Y - X\beta||_{2,\mathbb{R}^n}^2 + \lambda ||\beta||_{2,\mathbb{R}^p}^2 \,\} \quad (4)$$
$$= (X'X + \lambda I_p)^{-1}X'Y \quad (5)$$

$$\hat{f}(X)_{LASSO} = X\hat{\beta}_{LASSO}$$
$$\text{where } \hat{\beta}_{LASSO} = \underset{\beta \in \mathbb{R}^p}{argmin}\{\, ||Y - X\beta||_{2,\mathbb{R}^n}^2 + \lambda ||\beta||_{1,\mathbb{R}^p} \,\} \quad (6)$$

Note that $\hat{\beta}_{LASSO}$ does not admit a closed form like $\hat{\beta}_{Ridge}$ in the general case. Indeed the L^1 norm in the LASSO penalty makes the objective function non-differentiable when $\beta_j = 0$ for any β_j. However, a closed form for $\hat{\beta}_{LASSO}$ is available via the *soft-thresholding operator* and the OLS estimate when X is orthonormal (i.e., $X'X = XX' = I$), which is however rarely the case with SNP markers. Nevertheless, there are many possible algorithms to compute LASSO solutions such as the least angle regression selection (LARS) (Efron et al., 2004), proximal gradient descent based iterative soft-thresholding (ISTA) (Gordon and Tibshirani, 2012), cyclic coordinate descent (Friedman et al., 2010) and etc. However, these algorithms are beyond the scope of this article and are not the focus here. The objective functions in Problem (4) and (6) are also particular type of functions called relaxed Lagrangians, which are unconstrained formulations of constrained optimization problems. Searching for the saddle points of these Lagrangians is equivalent to searching for the solutions to the constrained formulations of Problem (4) and (6). Specifically, the solutions to Problem (4) and (6) are obtained when the ellipses, defined by the contour lines of the empirical risk term, touch the different constrained regions for β imposed by the L^2 and L^1 norms respectively. Hence, the L^1 norm generally induces sparsity in the LASSO solution, compared to the L^2 norm which induces a shrinkage of the β_j in the Ridge solution, when λ increases (Friedman et al., 2001).

Another way to tackle Problem (4) and (6) is in a probabilistic manner via a Bayesian treatment. Moreover the Bayesian treatment allows one to see the direct equivalence between Ridge regression, Bayesian Ridge regression, RR-BLUP and GBLUP. The equivalence between Ridge regression, RR-BLUP and GBLUP is a direct consequence of the proof of equivalence between Ridge regression and Bayesian Ridge regression (Lindley and Smith, 1972; Bishop and Tipping, 2003; De los Campos et al., 2013a). The proof found in De los Campos et al. (2013a) is reported below.

Proof of equivalence between Ridge regression and Bayesian Ridge regression:

$$\hat{\beta}_{Ridge} = \underset{\beta\in\mathbb{R}^p}{argmin}\left\{ \sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p} X_i^{(j)}\beta_j\right]^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \right\} \tag{7}$$

$$\left(\text{take } \lambda = \frac{\sigma_\varepsilon^2}{\sigma_\beta^2} \text{ and } \times -\frac{1}{2}\right)$$

$$= \underset{\beta\in\mathbb{R}^p}{argmax}\left\{ -\frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p} X_i^{(j)}\beta_j\right]^2 - \frac{1}{2}\frac{\sigma_\varepsilon^2}{\sigma_\beta^2}\sum_{j=1}^{p}\beta_j^2 \right\} \tag{8}$$

(divide by σ_ε^2 and apply monotonic transformation e^x)

$$= \underset{\beta\in\mathbb{R}^p}{argmax}\left\{ \underbrace{\prod_{i=1}^{n}\mathcal{N}\left(Y_i|\sum_{j=1}^{p}X_i^{(j)}\beta_j, \sigma_\varepsilon^2\right)}_{i.e.,\ Y|\beta,\sigma_\varepsilon^2\sim\mathcal{N}_n(X\beta, I_n\sigma_\varepsilon^2)} \times \underbrace{\prod_{j=1}^{p}\mathcal{N}(\beta_j|0,\sigma_\beta^2)}_{i.e.,\ \beta|\sigma_\beta^2\sim\mathcal{N}_p(0,I_p\sigma_\beta^2)} \right\} \tag{9}$$

$$= \underset{\beta\in\mathbb{R}^p}{argmax}\left\{f(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)\right\} = \text{mode}\left\{f(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)\right\}$$

$$= \hat{\beta}_{Bayesian\ Ridge} \tag{10}$$

where $f(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)$ is the density of the posterior distribution for β (i.e., marker effects) in (10). Due to the proportionality between the posterior density, and the product of gaussian densities for the likelihood and the prior distribution for β, $f(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)$ is also the density of a gaussian distribution by conjugacy. Thus, by symmetry of the gaussian distribution for $\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2$, we have $\text{mode}\left\{f(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)\right\} = \mathbb{E}(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2) = \mathbb{C}ov(\beta,Y)\mathbb{V}ar(Y)^{-1}Y = I_p\sigma_\beta^2X'[XX'\sigma_\beta^2 + \sigma_\varepsilon^2I_n]^{-1}Y = X'[XX' + \lambda I_n]^{-1}Y$ under the assumption that $\mathbb{C}ov(\beta,\varepsilon) = 0$, where $\mathbb{E}(\beta|Y,\sigma_\varepsilon^2,\sigma_\beta^2)$ can be identified to be the BLUP (Robinson, 1991; Schaeffer, 2010) of β and corresponds to the solution of the RR-BLUP model: $\hat{\beta}_{RR-BLUP} = X'[XX' + \lambda I_n]^{-1}Y$. Hence we have $\hat{\beta}_{Ridge} = \hat{\beta}_{Bayesian\ Ridge} = \hat{\beta}_{RR-BLUP}$.

We recall that the RR-BLUP model (Ruppert et al., 2003) corresponds to the following mixed model $Y = X\beta + \varepsilon$, where $\beta \sim \mathcal{N}_p(0, I_p\sigma_\beta^2)$, $\varepsilon \sim \mathcal{N}_n(0, I_n\sigma_\varepsilon^2)$ and $\mathbb{C}ov(\beta,\varepsilon) = 0$. If $U = X\beta$, this model can be rewritten as $Y = U + \varepsilon$, where $U \sim \mathcal{N}_n(0, \sigma_U^2\mathbb{U})$ with genomic covariance matrix $\mathbb{U} = XX'$ and $\sigma_U^2 = \sigma_\beta^2$. The GBLUP of U for this model is given by $\hat{U} = \mathbb{C}ov(U,Y)\mathbb{V}ar(Y)^{-1}Y = XX'[XX' + \lambda I_n]^{-1}Y$. So it is clear that predictions obtained with RR-BLUP and GBLUP are mathematically equivalent.

The equivalence between LASSO and Bayesian LASSO, i.e., $\hat{\beta}_{LASSO} = \hat{\beta}_{Bayesian\ LASSO}$, can be shown using the same type of arguments as for the proof of equivalence between Ridge regression and Bayesian Ridge regression. For example, the proof of equivalence between LASSO and Bayesian LASSO can also be found in De los Campos et al. (2013a). In the case of Bayesian LASSO the prior density for β corresponds to the product of p i.i.d Laplace densities for the marker effects (Tibshirani, 1996; De los Campos et al., 2013a). Thus, the prior distributions for β, in Bayesian Ridge regression and Bayesian LASSO, give another insight on the shrinked and sparse solutions for Ridge and LASSO respectively.

2.3. Dual Formulation of Ridge Regression in Terms of Kernel Functions

We recall that the classical formulation of Ridge regression is given by $\hat{f}(X)_{Ridge} = X\hat{\beta}_{Ridge}$ where $\hat{\beta}_{Ridge} = (X'X + \lambda I_p)^{-1}X'Y$. This formulation is also known as the primal formulation of Ridge regression. However, one can notice that the Ridge solution can be written as $\hat{\beta}_{Ridge} = X'\hat{\alpha}^{Ridge}$ where $\hat{\alpha}^{Ridge} = \frac{1}{\lambda}[Y - X\hat{\beta}_{Ridge}]$. Therefore, by substituting $X'\hat{\alpha}^{Ridge}$ for $\hat{\beta}_{Ridge}$ in the expression for $\hat{\alpha}^{Ridge}$, we also have $\hat{\alpha}^{Ridge} = (XX' + \lambda I_n)^{-1}Y$. Hence Ridge regression can be reformulated as follows:

$$\hat{f}(X)_{Ridge} = XX'\hat{\alpha}^{Ridge} \quad \text{where} \quad \hat{\alpha}^{Ridge} = (XX' + \lambda I_n)^{-1}Y \in \mathbb{R}^n \tag{11}$$

Expression (11) is called the dual formulation of Ridge regression (Saunders et al., 1998), where the components of the vector $\hat{\alpha}^{Ridge} = (\hat{\alpha}_1^{Ridge}, \hat{\alpha}_2^{Ridge}, ..., \hat{\alpha}_n^{Ridge})$ are called dual variables. It is clear that Expression (11) is identical to the GBLUP expression seen in the previous section. Hence, the classical formulation of Ridge regression and GBLUP are primal and dual formulations, respectively, of the same solution to a RERM problem. Note that Expression (11) requires the inversion of an $n \times n$ matrix compared to Expression (5) where a $p \times p$ matrix needs to be inverted. This is particularly convenient in the context of SNP markers where $p >> n$. If we let $\mathbb{K} = XX'$, Expression (11) can be written more conveniently as:

$$\hat{f}(X)_{Ridge} = \mathbb{K}\hat{\alpha}^{Ridge} \quad \text{where} \quad \hat{\alpha}^{Ridge} = (\mathbb{K} + \lambda I_n)^{-1} Y \tag{12}$$

For each genotype vector $X_i = [X_i^{(1)}, X_i^{(2)}, .., X_i^{(j)}, .., X_i^{(p)}] \in \mathbb{R}^p$, Expression (12) can be written as:

$$\hat{f}(X_i)_{Ridge} = \sum_{j=1}^{n} \hat{\alpha}_j^{Ridge} \mathbb{K}_{ij} = \sum_{j=1}^{n} \hat{\alpha}_j^{Ridge} < X_i, X_j >_{\mathbb{R}^p} \tag{13}$$

where $\mathbb{K}_{ij} = < X_i, X_j >_{\mathbb{R}^p}$ are elements of $\mathbb{K}$, i.e., $\mathbb{K} = (\mathbb{K}_{ij})_{1 \leq i,j \leq n}$, and $< ., . >_{\mathbb{R}^p}$ denotes the inner product between two vectors in $\mathbb{R}^p$. Expression (13) is particularly helpful as it can allow one to understand the kernel "trick" exploited by kernel methods, in the context of epistatic genetic architectures, as shown by the following example.

Consider the school case where we have $p = 2$ markers, i.e., $X_i = [\, X_i^{(1)}, \, X_i^{(2)} \,]$, and n measured phenotypic responses. Moreover, consider the following transformation ϕ applied to X_i: $\phi(X_i) = [\, (X_i^{(1)})^2, \underbrace{\sqrt{2} X_i^{(1)} X_i^{(2)}}_{Interaction\ term}, (X_i^{(2)})^2 \,] = [\, \phi^{(1)}(X_i), \, \phi^{(2)}(X_i), \, \phi^{(3)}(X_i) \,] \in \mathbb{R}^3$ where $\phi^{(2)}(X_i)$ corresponds to the interaction term between $X_i^{(1)}$ and $X_i^{(2)}$. In what follows we define $\phi(X)$ to be the $n \times 3$ transformed marker genotype matrix. Hence, for our school case, two possible models for example are given by:

$$\begin{aligned}
&Model\ 1\ (M_1): \hat{f}(X_i) = \hat{\beta}_1 X_i^{(1)} + \hat{\beta}_2 X_i^{(2)} \\
&\quad \text{where} \quad \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \underset{\beta \in \mathbb{R}^2}{argmin} \{\ ||Y - X\beta||^2_{2,\mathbb{R}^n} + \lambda ||\beta||^2_{2,\mathbb{R}^2}\ \} \\
&\quad \iff \hat{f}(X_i) = \sum_{j=1}^{n} \hat{\alpha}_j^{M_1} \mathbb{K}_{ij} = \sum_{j=1}^{n} \hat{\alpha}_j^{M_1} < X_i, X_j >_{\mathbb{R}^2} \quad (14) \\
&Model\ 2\ (M_2): \hat{f}(X_i) = \hat{\theta}_1 \phi^{(1)}(X_i) + \hat{\theta}_2 \phi^{(2)}(X_i) + \hat{\theta}_3 \phi^{(3)}(X_i) \\
&\quad \text{where} \quad \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix} = \underset{\theta \in \mathbb{R}^3}{argmin} \{\ ||Y - \phi(X)\theta||^2_{2,\mathbb{R}^n} + \lambda ||\theta||^2_{2,\mathbb{R}^3}\ \} \\
&\quad \iff \hat{f}(X_i) = \sum_{j=1}^{n} \hat{\alpha}_j^{M_2} \mathbb{K}_{ij}^{\phi} = \sum_{j=1}^{n} \hat{\alpha}_j^{M_2} < \phi(X_i), \phi(X_j) >_{\mathbb{R}^3} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15) \\
&\quad \text{where} \quad \mathbb{K}^{\phi} = \phi(X)\phi(X)' \ \text{and} \ \hat{\alpha}^{M_2} = (\mathbb{K}^{\phi} + \lambda I_n)^{-1} Y
\end{aligned}$$

However one can notice that $\mathbb{K}_{ij}^{\phi} = < \phi(X_i), \phi(X_j) >_{\mathbb{R}^3} = (< X_i, X_j >_{\mathbb{R}^2})^2 = (\mathbb{K}_{ij})^2$. Indeed we have:

$$\begin{aligned}
< \phi(X_i), \phi(X_j) >_{\mathbb{R}^3} &= \left[(X_i^{(1)} X_j^{(1)})^2 + 2(X_i^{(1)} X_j^{(1)})(X_i^{(2)} X_j^{(2)}) \right. \\
&\quad \left. + (X_i^{(2)} X_j^{(2)})^2 \right] \\
&= \left[X_i^{(1)} X_j^{(1)} + X_i^{(2)} X_j^{(2)} \right]^2 = (< X_i, X_j >_{\mathbb{R}^2})^2
\end{aligned}$$

This means that we only need to square the elements of matrix $\mathbb{K}$ for *Model* 1 to obtain *Model* 2 (i.e., to perform a Ridge regression in $\mathbb{R}^3$ modeling an interaction term). Indeed, in matrix form *Model* 2 can be written as:

$$\hat{f}(X)_{Ridge} = \mathbb{K}^{\phi}(\mathbb{K}^{\phi} + \lambda I_n)^{-1} Y \quad \text{where} \quad \mathbb{K}_{ij}^{\phi} = (\mathbb{K}_{ij})^2 \tag{16}$$

Similarly, for the case of $p = 3$ markers we can perform a Ridge regression in $\mathbb{R}^6$, which models three interaction terms, by just squaring the inner product between genotype vectors in $\mathbb{R}^3$, i.e., $\mathbb{K}_{ij}^{\phi} = (< X_i, X_j >_{\mathbb{R}^3})^2$. This process of implicitly computing inner products, in the space of transformed genotype vectors, by performing computations only in the original space of genotype vectors is known as the kernel "trick." The space of transformed covariates (i.e., space of transformed genotype vectors here), associated to a map ϕ, is commonly known as a feature space in machine learning. A kernel function associated to a feature map ϕ is defined as follows.

Definition of a kernel k:
For $X_i, X_j \in E$, a kernel k is a function which satisfies $k(X_i, X_j) =< \phi(X_i), \phi(X_j) >_F$, where E and F are the space of covariates and feature space respectively.

For example, in our school case we used the quadratic kernel defined by $k(X_i, X_j) = (< X_i, X_j >_E)^2 = < \phi(X_i), \phi(X_j) >_F$ where $F = \mathbb{R}^3$ when $E = \mathbb{R}^2$ (i.e., $p = 2$ markers). Note that there is no one-to-one correspondence between a feature map ϕ and a kernel k. Indeed, more that one feature map can be associated to a unique kernel (see Supplementary Material, lemma 6). In classical Ridge regression we do not have any interaction term and the feature map is the identity (i.e., $\phi = id$) since $\mathbb{K}_{ij} = k(X_i, X_j) =< X_i, X_j >$ in this situation. A necessary and sufficient condition for a function k to be a kernel is that matrix $\mathbb{K} = k(X_i, X_j)_{1 \leq i,j \leq n}$ (known as the Gram matrix) is positive semi-definite. This condition comes from Mercer's theorem (Gretton, 2013) and it gives a practical way to check if a function k defines a kernel.

Some kernels are called universal kernels in the sense that they can approximate any arbitrary function $f^*(.)$, with a finite number of training samples, if regularized properly (Micchelli et al., 2006). One such example is the Gaussian kernel given by $k(X_i, X_j) = e^{-h||X_i - X_j||_2^2}$, where $h > 0$ is a rate of decay parameter for k. This kernel is associated to an infinite-dimensional feature map which allows an implicit modeling of all possible orders of interaction between markers (see Supplementary Material, lemma 7). Hence, the Gaussian kernel is useful for genomic prediction when dealing with complex epistatic genetic architectures.

2.4. RKHS and the Representer Theorem

The concept of RKHS (Smola and Schlkopf, 1998; Cornuéjols and Miclet, 2011; Gretton, 2013) with its implications in statistics and machine learning are well beyond the scope of this article. Here we review the basic definition of a RKHS so as to introduce the representer theorem which exploits the definition of RKHS. The representer theorem has important applications in practice. Indeed, it can allow one to find optimal solution to RERM problems and it shows that solutions to many parametric and machine learning problems have similar form.

Definition of a RKHS:
Let $\phi(X_i) = k(., X_i)$, a RKHS H_k associated to a kernel k can be defined as a space of functions generated by linear combinations of $k(., X_i)$;

$$H_k = \left\{ \sum_{i=1}^{n} \alpha_i k(., X_i);\ X_i \in E,\ \alpha_i \in \mathbb{R},\ n \in \mathbb{N} \right\}$$

such that (i) for all $X_i \in E$, $k(., X_i) \in H_k$ and (ii) for all $X_i \in E$ and every $f(.) \in H_k$, $< f(.), k(., X_i) >_{H_k} = f(X_i)$ (Cornuéjols and Miclet, 2011). The condition (ii) is called the reproducing property of k as it reproduces f in some sense. Hence, from the reproducing property we have $< \phi(X_i), \phi(X_j) >= < k(., X_i), k(., X_j) >= k(X_i, X_j)$. According to Moore-Aronszajn theorem, every RKHS has a unique positive semi-definite kernel (i.e., a reproducing kernel) and vice-versa. In other words, there is one-to-one correspondence between RKHS and kernels. A simplified version of the representer theorem is given as follows.

The Representer Theorem (Kimeldorf and Wahba, 1971):
Fix a set E and a kernel k, and let H_k be the corresponding RKHS. For any loss function $L : \mathbb{R}^2 \rightarrow \mathbb{R}$, the solution $\hat{f}$ of the optimization problem;

$$\hat{f}(.) = \underset{f \in \mathcal{H}_k}{argmin} \left\{ \sum_{i=1}^{n} L(Y_i, f(X_i)) \ + \lambda ||f||^2_{\mathcal{H}_k} \right\} \qquad (17)$$

has the following form:

$$\hat{f}(.) = \sum_{i=1}^{n} \alpha_i k(., X_i) \qquad (18)$$

This result is of great practical importance. For example, if we substitute the representation Equation (18) into Equation (17) when $L(Y_i, f(X_i)) = (\ Y_i - f(X_i)\ )^2$ (aka kernel Ridge regression) then we obtain the following equivalent problem;

$$\hat{\alpha}_{Kernel\ Ridge} = \underset{\alpha \in \mathbb{R}^n}{argmin} \left\{ \ \frac{1}{2} ||Y - \mathbb{K}\alpha||_2^2 + \frac{\lambda}{2} \alpha' \mathbb{K} \alpha \ \right\} \qquad (19)$$

where $\hat{\alpha}_{Kernel\ Ridge}$ can be shown to be given by $\hat{\alpha}_{Kernel\ Ridge} = [\mathbb{K} + \lambda I_n]^{-1} Y$. Moreover, if we follow the same reasoning as for Equation (7) to Equation (10), one can easily show from Equation (19) that $\hat{\alpha}_{Kernel\ Ridge} = \hat{\alpha}_{Bayesian\ Kernel\ Ridge} = \hat{\alpha}_{RR-BLUP}$, where $\hat{\alpha}_{RR-BLUP}$ is the BLUP of α for the following mixed model;

$$Y = \mathbb{K}\alpha + \varepsilon \ \text{ where } \ \alpha \sim \mathcal{N}_n(0, \sigma^2_\alpha \mathbb{K}^{-1}) \text{ and } \varepsilon \sim \mathcal{N}_n(0, \sigma^2_\varepsilon)$$

$$\Leftrightarrow Y = g + \varepsilon \ \text{ where } g = \mathbb{K}\alpha \sim \mathcal{N}_n(0, \sigma^2_g \mathbb{K}) \text{ , with } \sigma^2_g = \sigma^2_\alpha \text{ , and } \varepsilon \sim \mathcal{N}_n(0, \sigma^2_\varepsilon)$$

Hence, the mixed model methodology can be used to solve kernel Ridge regression (i.e., RKHS regression) for which classical Ridge regression (i.e., GBLUP) is a particular case.

In Expression (17) we have $L(Y_i, f(X_i)) = |Y_i - f(X_i)|_\varepsilon$ for SVR (i.e., ε-insensitive loss), proposed by Vapnik (1998), which is given by:

$$|Y_i - f(X_i)|_\varepsilon = \begin{cases} 0 & \text{if } |Y_i - f(X_i)| \leq \varepsilon \\ |Y_i - f(X_i)| - \varepsilon & \text{otherwise} \end{cases}$$

Note that SVR also performs regularization in a RKHS. The parameter λ in Equation (17) correspond to $\frac{1}{2C}$ (where $C > 0$) in the original formulation of SVR (Basak et al., 2007). Moreover, slack variables are also found in the original formulation in order to cope with infeasible constraints. SVR has several interesting properties. For example, the dual optimization problem of finding the Lagrange multipliers (i.e., dual variables: α_j, α^*_j) in SVR, and in support vector machine for classification, is generally a constrained Quadratic Programming (QP) problem which assures a global minimum. Furthermore, only a fraction of the Lagrange multipliers are non-zero, due to the so-called Karush-Kuhn-Tucker (KKT) conditions which state that the product between dual variables and constraints vanishes at the optimal solution. The genotype vectors (i.e., X_j) corresponding to non-zero Lagrange multipliers are called support vectors as they are the only ones which contribute to prediction. This is particularly convenient for data sets with a large number of accessions where we need only the support vectors for prediction. Indeed, the estimated prediction function in SVR can we written as;

$$\hat{f}(X_i)_{SVR} = \sum_{j=1}^{n} \bar{\alpha}_j k(X_i, X_j) + b \text{ with } \bar{\alpha}_j = (\hat{\alpha}_j - \hat{\alpha}^*_j) \ \text{ and } \ b \in \mathbb{R}$$

where only a restricted number of $\bar{\alpha}_j$ are non-zero and have corresponding support vectors X_j (note that both $\hat{\alpha}_j$ and $\hat{\alpha}^*_j$ cannot be non-zero simultaneously Basak et al., 2007; Al-Anazi and Gates, 2012). These vectors are associated to approximation errors which are greater than ε. Thus, the number of support vectors is inversely proportional to the ε parameter. Further details on SVR, and on support vector machine in the general case, can be found in Vapnik (1998), Smola and Schlkopf (1998), Basak et al. (2007), and Al-Anazi and Gates (2012). Finally, note that we do not have the representation Equation (18) for LASSO since the L^1 norm, for this particular case, violates the representer theorem assumptions.

2.5. Analyzed Data Sets and Prediction Methods Compared

Three real data sets were analyzed. The first data set was composed of 230 temperate japonica accessions with 22,691 SNP. For the second data set, 167 tropical japonica accessions with 16,444 SNP were available. The third data set was composed of

188 tropical japonica accessions with 38,390 SNP. A total of 15 traits were analyzed for the three data sets. Plant height (PH), flowering time (FL), leaf arsenic content (AR), number of tillers (NT), shoot biomass (SB), maximum root length (RL), number of roots below 30 centimeters (NR), deep root biomass (DR) and root over shoot biomass ratio (RS) were analyzed for the first and second data sets. For the third data set, PH, cycle duration (CD), fertility rate (FE), number of seeds per panicle (NS), straw yield in kilograms per hectare (SY) and number of panicles per square metre (NP) were analyzed. All SNP marker data sets had a minor allele frequency strictly superior to 1%. The three data sets are officially available at http://tropgenedb.cirad.fr/tropgene/JSP/interface.jsp?module=RICE as the "GS-RUSE.zip" folder, or can be downloaded directly at http://tropgenedb.cirad.fr/tropgene/downloads/studies/GS-RUSE.zip.

Four methods; LASSO, GBLUP, RKHS regression and SVR were applied to these data sets and traits, and hence a total of 60 situations were examined. R scripts were written to perform analyses with the four methods and are available on request. The glmnet (Friedman et al., 2010) and kernlab (Karatzoglou et al., 2004) packages were used for LASSO and SVR respectively. R scripts were written to solve GBLUP and RKHS regression. The expectation-maximization (EM) algorithm (Dempster et al., 1977; Foulley, 2002; Jacquin et al., 2014) was used to maximize the restricted likelihoods (REML) of the mixed models, associated to GBLUP and RKHS regression respectively, in order to estimate the associated variance parameters. The Gaussian kernel was used for RKHS regression and SVR. For RKHS regression, values over several grids were tested using cross-validation to tune the rate of decay parameter for each data set. For SVR, the rate of decay was estimated using the heuristic defined in the `sigest` function (Karatzoglou et al., 2004), which is already implemented in the `ksvm` function (Karatzoglou et al., 2004), that allows an automatic selection of this parameter. The regularization parameter C for SVR was estimated as $C = max(|\bar{Y} + 3\sigma_Y|, |\bar{Y} - 3\sigma_Y|)$, where $\bar{Y}$ is the phenotypic mean, as recommended by Cherkassky and Ma (2004). Values higher than 0.5 for the ε parameter in SVR produced no support vectors. Hence lower values were tested for this parameter using cross validation. Values ranging between 0.01 and 0.1 were found to give similar and the best predictive performance for each data set, hence ε was fixed to 0.01. For LASSO, the `cv.glmnet` function (Friedman et al., 2010) was applied with its default values for the `alpha` and `nfolds` parameters (i.e., 1 and 10 respectively). For this function, the squared loss (i.e., `mse` in `cv.glmnet`) was used for cross-validation and its associated `lambda.min` parameter was used as the optimal lambda for prediction.

To evaluate the genomic predictive ability of the four methods, cross-validations were performed by sampling randomly a training and a target population 100 times for each case among the 60 situations. For each random sampling the sizes of the training and target sets were, respectively, two-thirds and one-third times the size of the total population. The Pearson correlation, between the predicted genetic values and the observed phenotypes for the target set, was taken as a measure of relative prediction accuracy (RPA). Indeed, true prediction accuracy (TPA) can be attained only if the true genetic values for the target set are available. The signal-to-noise ratio (SNR) (Czanner et al., 2015) for each method, with respect to each target set, was calculated as the sample variance of the predicted genetic values over the sample variance of the estimated residuals associated to the target phenotypes. Note that the SNR is related to genomic based heritabilities (De los Campos et al., 2013b; Janson et al., 2015). However, there are many different definitions of heritability (Janson et al., 2015) and these are different according to each studied method here. Hence we report only the estimated SNR for each method.

3. RESULTS

Table 1 gives the RPA means with their associated standard errors and the SNR means for the 60 examined situations. The RPA standard errors and SNR means are given within parantheses and square brackets respectively. **Figures 1–5** give the boxplots for the RPA distributions associated to the 60 studied cases.

As can be seen in **Table 1** and in **Figures 1–5**, RKHS regression and SVR performed as well or better than LASSO and GBLUP for most situations. Furthermore, in **Figures 1–5**, RKHS regression and SVR gave RPA values strictly greater than 0, for all data sets and traits, compared to LASSO and GBLUP. Indeed, LASSO gave negative RPA values for PH and NR as can be seen in **Figures 1**, **3**, respectively. In **Figure 5**, both LASSO and GBLUP gave negative RPA values for SY.

In these figures and in **Table 1**, the largest RPA mean differences between parametric and kernel methods can be seen for AR, NT, SB, NR, DR, FE, NS, and SY (see bold values in **Table 1**). For these traits, the RPA mean differences between the parametric and kernel methods varied between 0.03 and 0.21. The highest observed RPA mean difference of 0.21 was between SVR and GBLUP for SY. For CD, one can see in **Table 1** that the RPA and SNR means for GBLUP were simultaneously lower and higher than those of the other methods. This can be explained by a poor consistency of GBLUP, with respect to the DGP for this trait, which leads to an over-estimation of the true SNR. For example, it was shown in the second subsection that the poor consistency of a linear model, due to over-fitting, could minimize substantially the estimated residual variance, thus leading to an over-estimation of the true SNR while inducing a poor predictive ability.

Among the kernel methods, RKHS regression was often more accurate than SVR although only little RPA mean differences can be observed between these methods in **Table 1**. On the other hand, GBLUP was often more accurate than LASSO for the parametric methods. As can be seen in **Table 1** and in **Figures 1–5**, LASSO had a much lower predictive performance than the other methods for most traits. The average of the RPA means for each method, across all traits for the three data sets, were 0.51, 0.50, 0.46, and 0.41 for RKHS regression, SVR, GBLUP and LASSO respectively.

For each analyzed trait, RKHS regression was performed in a reasonable computation time. For example, the computation time of one particular cross-validation for NT was 2.03 s on a personal computer with 8 GB RAM. However, depending on

TABLE 1 | RPA means with their associated standard errors within parantheses (.), and the SNR means within square brackets [.], for the 60 examined situations.

Data set	Trait	Method			
		LASSO	GBLUP	RKHS regression	SVR
Data set 1	PH	0.34 (0.11) [0.11]	0.40 (0.08) [0.14]	0.40 (0.08) [0.16]	0.37 (0.07) [0.21]
230 accessions	FL	0.59 (0.07) [0.42]	0.65 (0.06) [0.93]	0.67 (0.06) [0.73]	0.66 (0.07) [0.75]
22691 SNP	AR	**0.21** (0.11) [0.10]	**0.27** (0.07) [0.35]	**0.35** (0.07) [0.12]	**0.35** (0.08) [0.20]
	NT	**0.34** (0.11) [0.25]	**0.41** (0.09) [0.59]	**0.47** (0.09) [0.24]	**0.46** (0.08) [0.32]
Data set 2	SB	**0.42** (0.09) [0.31]	**0.49** (0.09) [0.72]	**0.53** (0.09) [0.33]	**0.52** (0.10) [0.37]
167 accessions	RL	0.39 (0.09) [0.29]	0.53 (0.09) [0.39]	0.54 (0.08) [0.33]	0.54 (0.09) [0.40]
16444 SNP	NR	**0.25** (0.13) [0.16]	**0.39** (0.09) [0.31]	**0.44** (0.09) [0.17]	**0.42** (0.09) [0.31]
	DR	**0.39** (0.12) [0.29]	**0.45** (0.11) [0.67]	**0.49** (0.10) [0.40]	**0.48** (0.11) [0.21]
	RS	0.55 (0.08) [0.38]	0.54 (0.09) [0.70]	0.57 (0.07) [0.45]	0.57 (0.10) [0.30]
	PH	0.66 (0.07) [0.85]	0.69 (0.06) [1.15]	0.70 (0.05) [0.90]	0.69 (0.06) [0.81]
Data set 3	CD	0.48 (0.11) [0.29]	0.39 (0.09) [0.58]	0.47 (0.09) [0.26]	0.46 (0.09) [0.38]
188 accessions	FE	**0.39** (0.12) [0.28]	**0.43** (0.10) [0.58]	**0.50** (0.09) [0.46]	**0.50** (0.08) [0.47]
38390 SNP	NS	**0.38** (0.12) [0.33]	**0.50** (0.08) [0.44]	**0.54** (0.08) [0.38]	**0.55** (0.09) [0.45]
	SY	**0.18** (0.13) [0.14]	**0.12** (0.09) [0.03]	**0.28** (0.09) [0.10]	**0.33** (0.10) [0.28]
	NP	0.64 (0.08) [0.85]	0.70 (0.06) [0.80]	0.68 (0.06) [0.62]	0.67 (0.06)[0.65]

the trait considered, the computation time for RKHS regression was either lower or higher than that for SVR. For example, the computation times associated to one cross-validation for NT were 2.99 and 2.03 seconds for SVR and RKHS regression respectively. However, the computation times associated to one cross-validation for RL were 2.25 and 3.32 seconds for SVR and RKHS regression respectively. This can be explained by the, well known, slow convergence properties of the EM algorithm in some situations (Naim and Gildea, 2012).

4. DISCUSSION

4.1. Comparison of the Genomic Predictive Abilities of LASSO, GBLUP, SVR and RKHS Regression

Among all the compared methods, RKHS regression and SVR were regularly the most accurate methods for prediction followed by GBLUP and LASSO. On the other hand, LASSO was often the least accurate method for prediction. This can be explained by the fact that, for situations where $p > n$, the predictive performance of LASSO is regularly dominated by Ridge regression (i.e., GBLUP) when covariates are highly correlated (Tibshirani, 1996; Zou and Hastie, 2005). Moreover, Dalalyan et al. (2014) recently showed that the predictive performance of LASSO can be mediocre, irrespective of the choice of the tuning parameter, when covariates are moderately correlated. For SNP marker data it is common to have high numbers of moderately and highly correlated markers due to linkage disequilibrium. Furthermore, there was a limited number of accessions (i.e., n) for the three studied data sets. This may also explain the less accurate performance of LASSO. Indeed, it is well known that the number of non null coefficients for an estimated LASSO model is bounded by $min(n, p)$ (Tibshirani, 2013). Hence, when $p > n$, the number of markers (i.e., covariates) selected as relevant will be bounded by the number of accessions which may be inconsistent with the DGP. Alongside, Onogi et al. (2015) reported that the estimation of parameters via REML for GBLUP could be problematic for small sample size. This was observed for CD in our study where the RPA and SNR means for GBLUP were simultaneously lower and higher than those of the other methods.

Nevertheless, the observed RPA mean differences between the studied methods were somehow incremental for the three data sets. This is most probably due to the fact that our measure of RPA is based on the correlation between observed phenotypes, which are noisy measurements *per se*, and predicted genetic values. Moreover, Gianola et al. (2014) pointed out that differences among methods can be masked by cross-validation noise. Simulation studies conducted by our team (work to be published), based on real data for four traits, indicate that differences between methods based on TPA are often much higher than those based on RPA for the same simulated data set. In other words, small differences in RPA can be an indicator of higher differences in TPA among methods. Results for these simulation studies, with the corresponding simulated data sets, are available at http://tropgenedb.cirad.fr/tropgene/JSP/interface.jsp?module=RICE as the "GS-RUSE.zip" folder.

Still, our results show that kernel methods can be more appropriate than conventional parametric methods for many traits with different genetic architectures. These results are consistent with those of many previous studies (Konstantinov and Hayes, 2010; Pérez-Rodríguez et al., 2012; Sun et al., 2012; Howard et al., 2014). With respect to Morota and Gianola (2014), our results also indicate that kernel methods will have higher

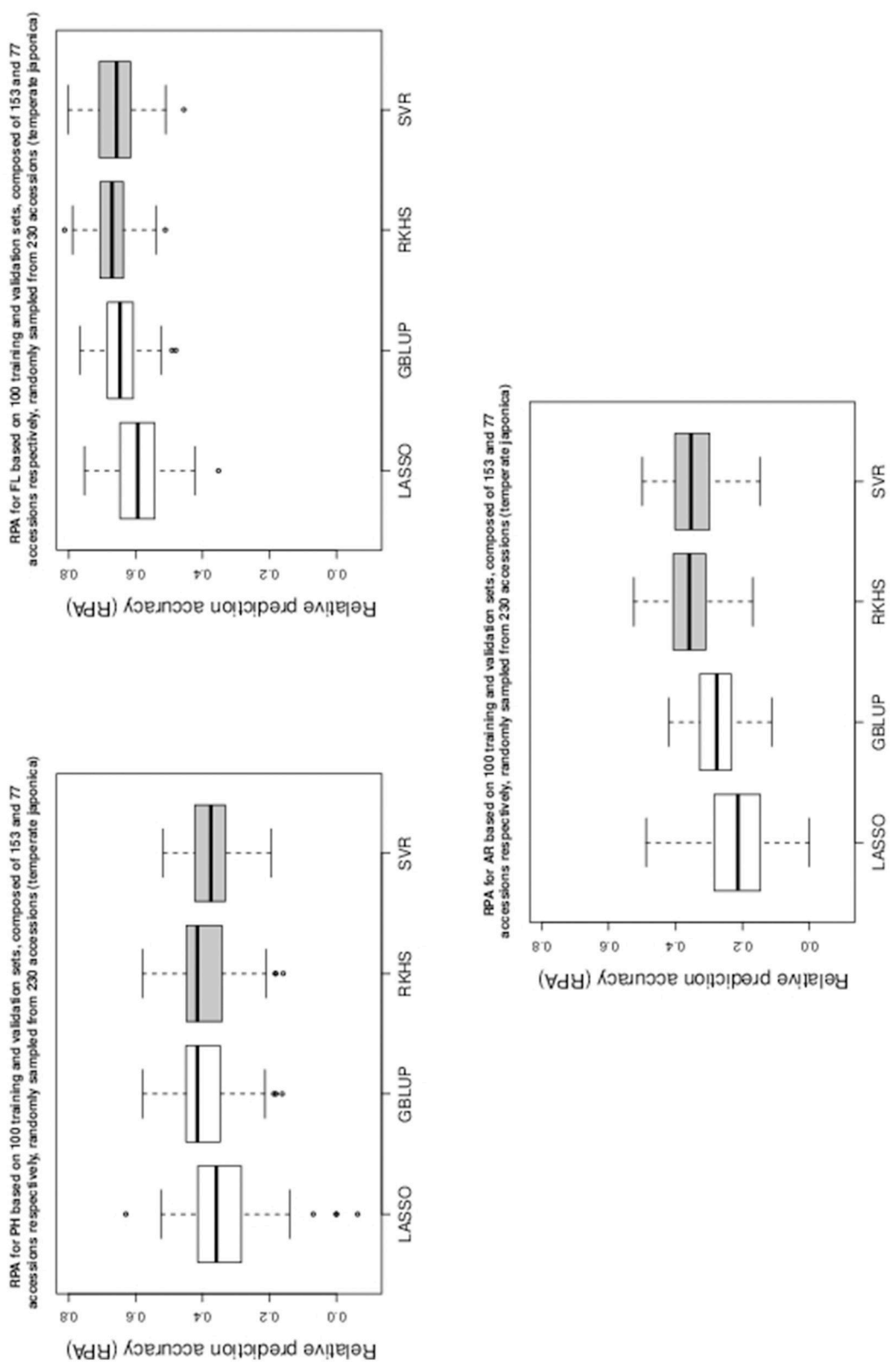

FIGURE 1 | Boxplots of RPA distributions associated to PH, FL, and AR for data set 1.

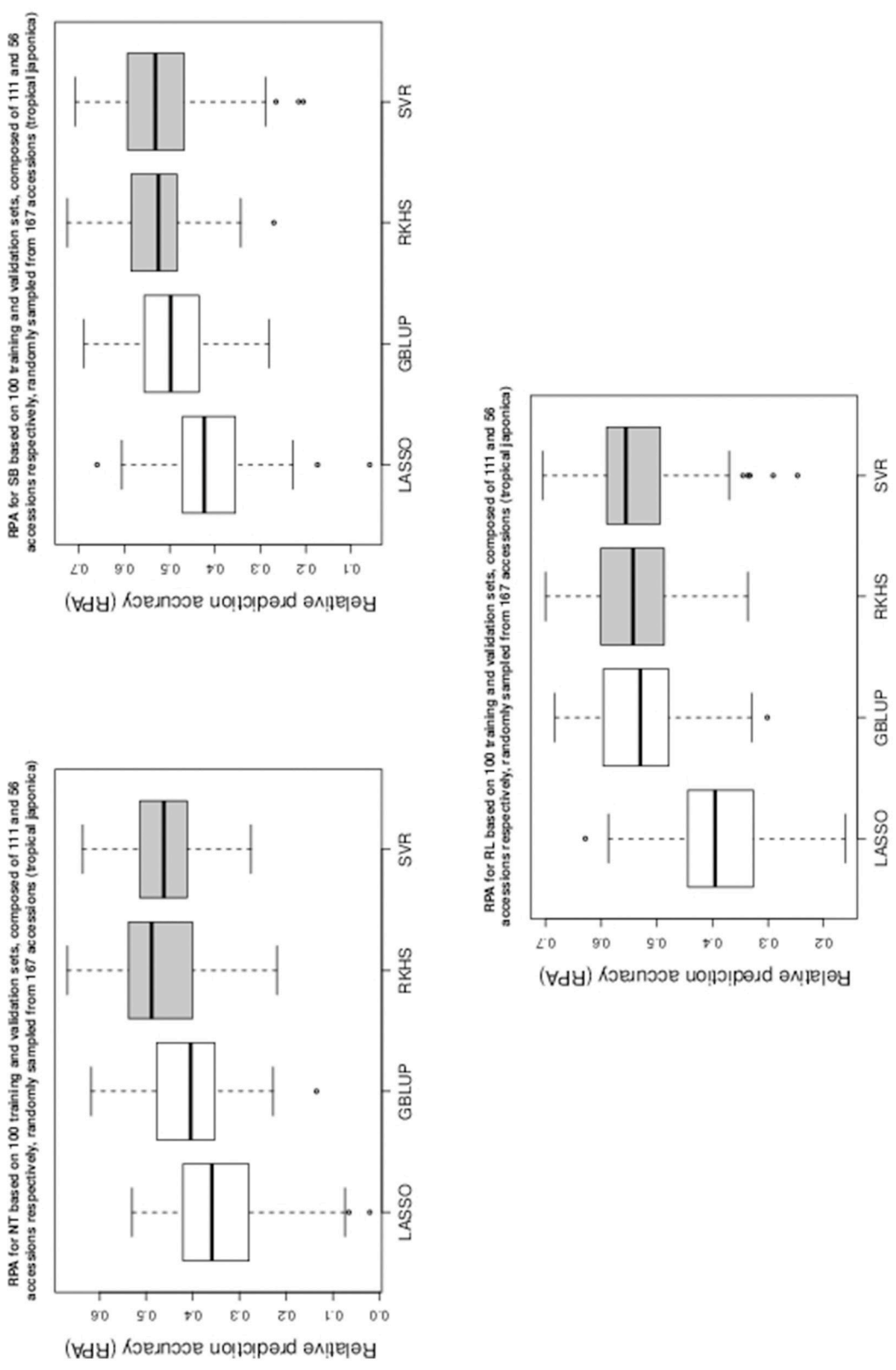

FIGURE 2 | Boxplots of RPA distributions associated to NT, SB, and RL for data set 2.

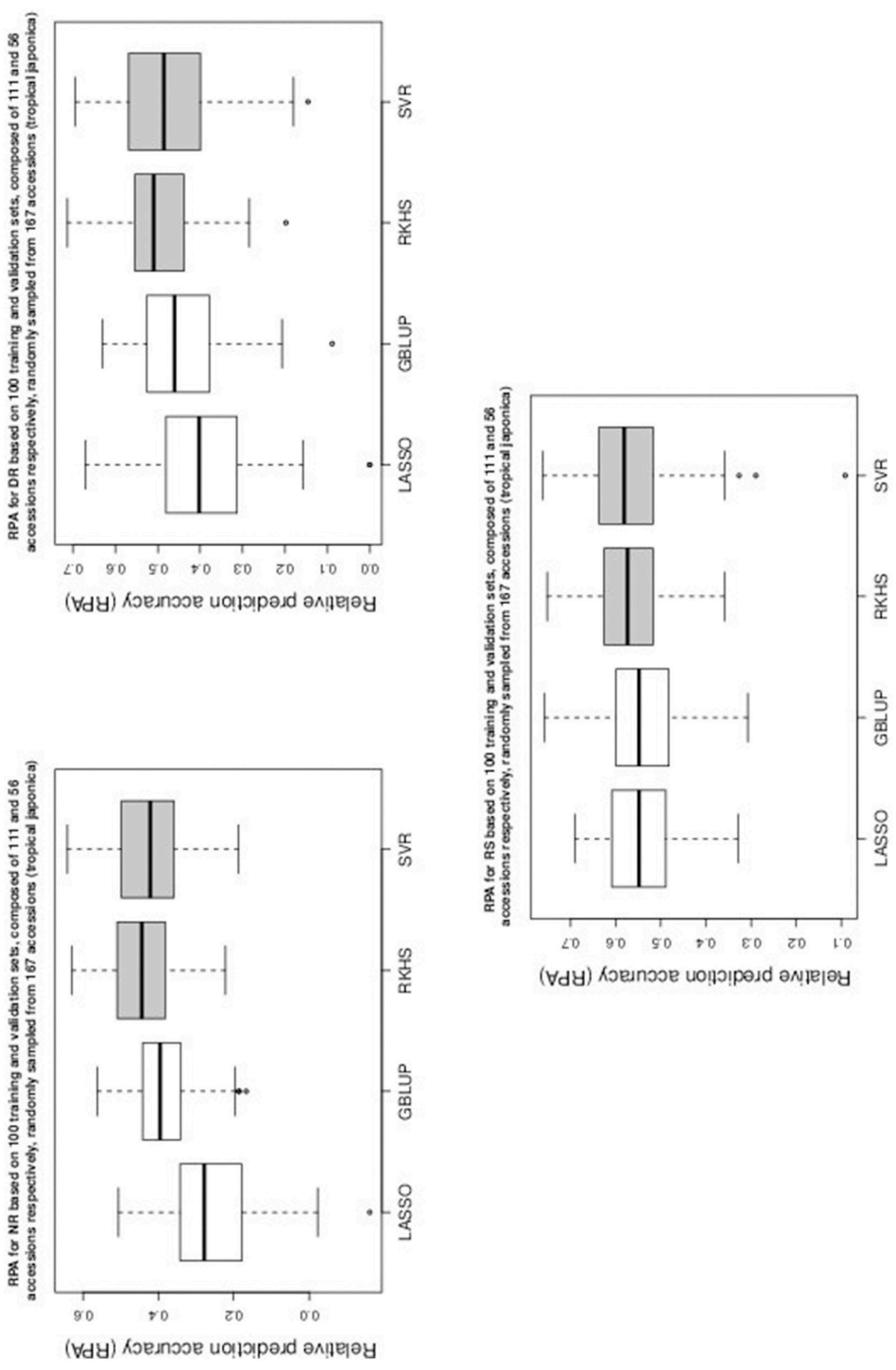

FIGURE 3 | Boxplots of RPA distributions associated to NR, DR, and RS for data set 2.

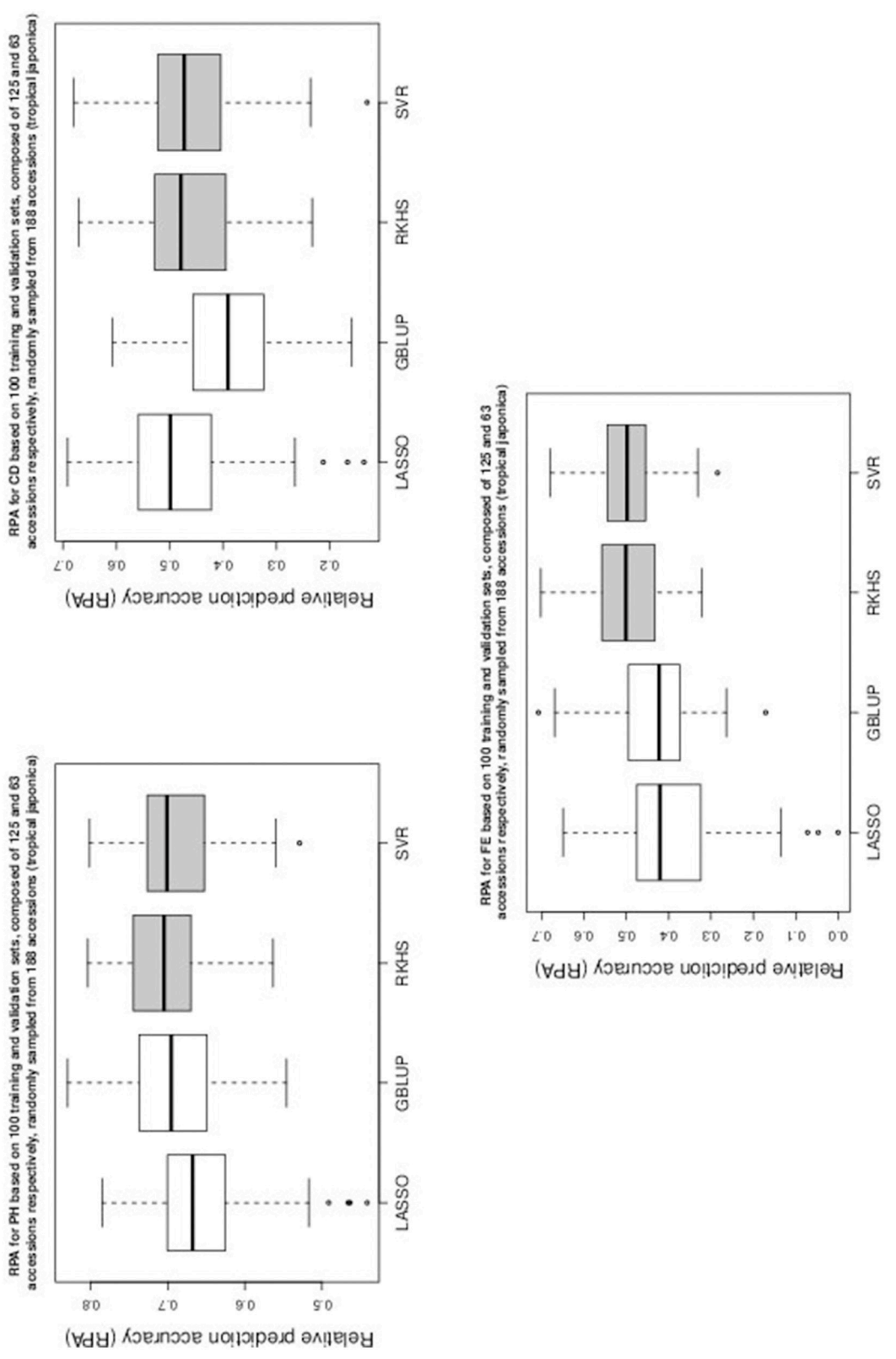

FIGURE 4 | Boxplots of RPA distributions associated to PH, CD, and FE for data set 3.

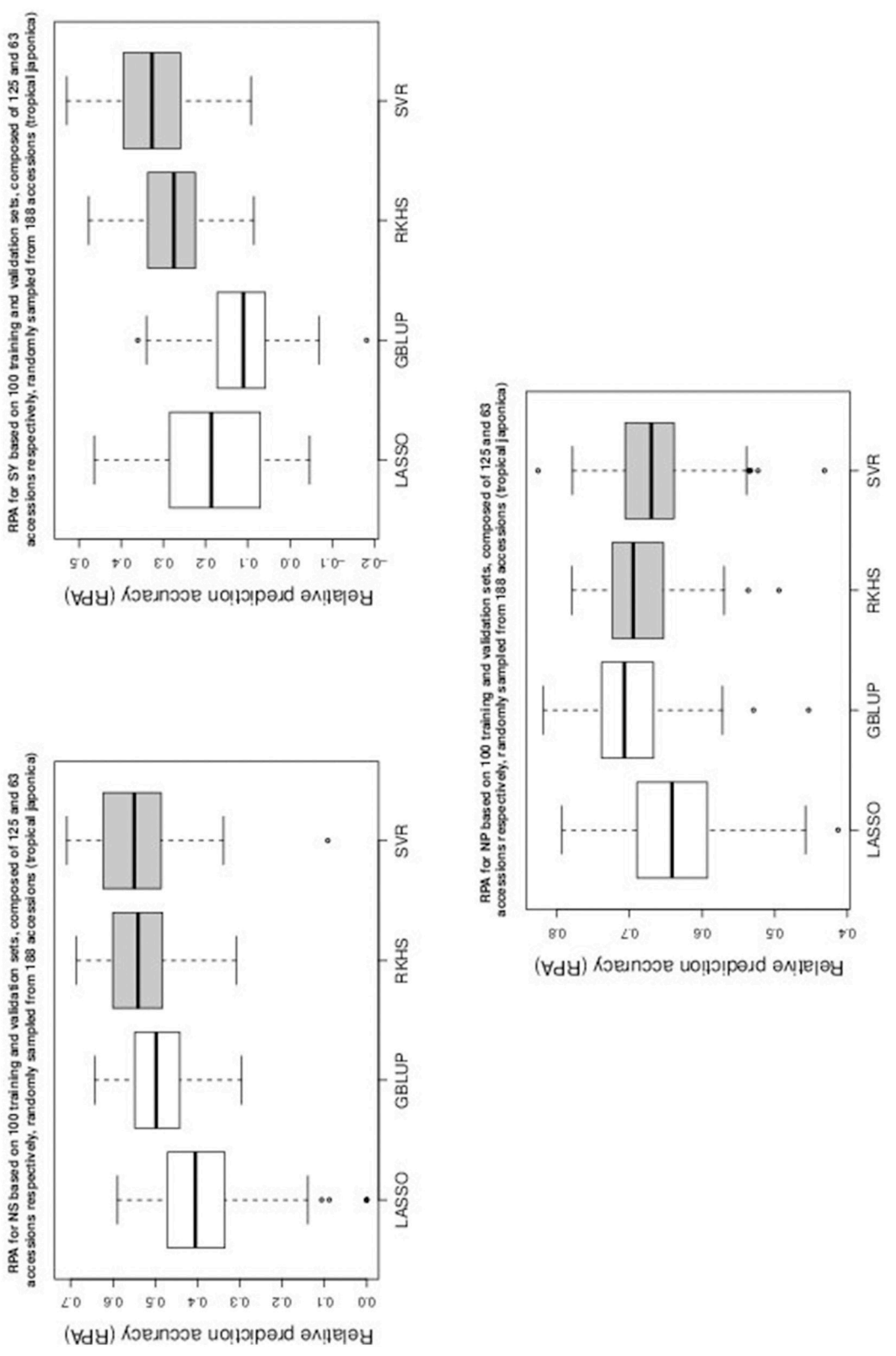

FIGURE 5 | Boxplots of RPA distributions associated to NS, SY, and NP for data set 3.

predictive performance, than conventional parametric methods, for traits potentially having moderate to complex epistatic genetic architectures. For example, the large RPA mean differences for SY, between the studied parametric and kernel methods, is probably due to an epistatic genetic architecture associated to this trait as pointed out by Liu et al. (2006). The same reasoning can be applied to AR for which epistatic mechanisms might potentially be involved (Norton et al., 2010). In this study, SVR and RKHS regression had similar predictive abilities. However, one advantage of RKHS regression over SVR lies in the fewer number of parameters to be estimated, which can be automated quite easily. Thus, RKHS regression can be performed more easily than SVR by low experienced users. Indeed, as pointed out by Cherkassky and Ma (2004), SVR application studies are usually performed by "practitioners," who have a good understanding of the SVM methodology, since the main issue in having good SVM models lies in the proper setting of the meta-parameters.

4.2. Comparison and Connections Between Kernel Methods and Other Methods in Frequentist and Bayesian Frameworks

In comparison with Howard et al. (2014), we did not compare the studied kernel methods to neural networks (NN). Nevertheless, these authors showed that NN did not perform better than these methods in their simulation study. As pointed out by Howard et al. (2014), it is well-known that NN can be prone to over-fitting which reduces predictive performance. Moreover, NN are plagued with the problem of local minima in comparison to support vector machines which are not (Smola and Schlkopf, 1998). Yet, connections between NN with a single layer of hidden units (i.e., neurons) and kernel machines exist (Cho and Saul, 2009). In our study we reviewed the equivalence between well-known regularized, mixed and Bayesian linear models. As a matter of fact, for parametric models where one can specify likelihoods, inferences from frequentist (i.e., maximum likelihood based approaches) and Bayesian procedures will be practically the same if n (i.e., number of accessions) becomes sufficiently large for a fixed p. This is a consequence of the so-called Bernstein-von Mises theorem (Ghosal et al., 1995; Ghosal, 1997). Moreover, we showed in this study that many parametric methods can be framed as kernel methods, with simple kernels, due to their equivalent primal and dual formulations. For instance, this was shown for Ridge regression, Bayesian Ridge regression, RR-BLUP and GBLUP which are mathematically equivalent methods for prediction.

Framing parametric methods as kernel machines with simple kernels has important implications in the sense that many kernel methods can be specified, and solved conveniently, in existing classical frequentist (e.g., embedding kernels in mixed models) and Bayesian frameworks. This was first pointed out by Gianola et al. (2006) and several following works (De los Campos et al., 2010; Endelman, 2011; Morota et al., 2013; Pérez and de los Campos, 2014) developed kernel methods in these frameworks. We also developed a simple and user-friendly R function within the mixed model framework, named `Kernel_Ridge_MM.R`, which allows users to perform RR-BLUP of marker effects, GBLUP and RKHS regression, with a Gaussian, Laplacian, polynomial or ANOVA kernel, in a reasonable computation time. In our study we used only the Gaussian kernel which performed well for RKHS regression. However, other kernels such as the polynomial or ANOVA kernel can be used. For instance, the ANOVA kernel was found to perform well in multidimensional regression problems (Hofmann et al., 2008). A modified version of this function named `Tune_kernel_Ridge_MM.R`, which allows users to tune the rate of decay parameter for RKHS regression based on K-folds cross validation, has also been developed for Windows, Linux and parallelized for HPC Linux clusters. Finally, an R package named `KRMM`, associated to these functions, has also been developed. The `KRMM` package and all scripts are publicly available at https://sourceforge.net/u/ljacquin/profile/. As conclusion, we recommend the use of kernel methods for genomic prediction, and selection, since the genetic architectures associated to quantitative traits are rarely known and can be very complex and complicated to model. Therefore, it seems more advisable to use data-driven prediction models, which can account for multiple orders of interaction, to assess the genetic merits of individuals.

AUTHOR CONTRIBUTIONS

LJ wrote the manuscript, developed all scripts and the `KRMM` package. LJ performed the analyses. T-VC and NA read and approved the manuscript.

FUNDING

This work was funded by Agropolis Foundation Grant n° 1201-006.

ACKNOWLEDGMENTS

The authors thank Brigitte Courtois and Louis-Marie Raboin for providing data set 2 and 3.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00145

This file contains the proofs of lemmas in the main text.

REFERENCES

Al-Anazi, A., and Gates, I. (2012). Support vector regression to predict porosity and permeability: effect of sample size. *Comput. Geosci.* 39, 64–76. doi: 10.1016/j.cageo.2011.06.011

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Inform. Process. Lett. Rev.* 11, 203–224.

Bishop, C. M., and Tipping, M. E. (2003). Bayesian regression and classification. *Nato Sci. Ser. Sub Ser. III Comput. Syst. Sci.* 190, 267–288.

Cherkassky, V., and Ma, Y. (2004). Practical selection of svm parameters and noise estimation for SVM regression. *Neural Netw.* 17, 113–126. doi: 10.1016/S0893-6080(03)00169-2

Cho, Y., and Saul, L. K. (2009). "Kernel methods for deep learning," in *Advances in Neural Information Processing Systems* (San Diego, CA), 342–350.

Cornuéjols, A., and Miclet, L. (2011). *Apprentissage Artificiel: Concepts et Algorithmes*. Paris: Editions Eyrolles.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Czanner, G., Sarma, S. V., Ba, D., Eden, U. T., Wu, W., Eskandar, E., et al. (2015). Measuring the signal-to-noise ratio of a neuron. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7141–7146. doi: 10.1073/pnas.1505545112

Dalalyan, A. S., Hebiri, M., and Lederer, J. (2014). On the prediction performance of the lasso. *arXiv preprint arXiv:1402.1700*.

De los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genet. Res.* 92, 295–308. doi: 10.1017/S0016672310000285

De los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013a). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

De los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013b). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* 39, 1–38.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–451. doi: 10.1214/009053604000000067

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Flint, J., and Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* 19, 723–733. doi: 10.1101/gr.086660.108

Foulley, J.-L. (2002). Algorithme em: théorie et application au modèle mixte. *J. la Société Française Statistique* 143, 57–109.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Vol. 1. Springer Series in Statistics. Berlin: Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33:1. doi: 10.18637/jss.v033.i01

Ghosal, S. (1997). "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference* (Varanasi: Banaras Hindu University).

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *Ann. Stat.* 23, 2145–2152. doi: 10.1214/aos/1034713651

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gianola, D., and van Kaam, J. B. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Gianola, D., Weigel, K. A., Krämer, N., Stella, A., and Schön, C.-C. (2014). Enhancing genome-enabled prediction by bagging genomic blup. *PLoS ONE* 9:e91693. doi: 10.1371/journal.pone.0091693

Gordon, G., and Tibshirani, R. (2012). Accelerated first-order methods. *Optimization* 10:725.

Gretton, A. (2013). Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn.* Lecture Conducted from University College London.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Stat.* 36, 1171–1220. doi: 10.1214/009053607000000677

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* 4, 1027–1046. doi: 10.1534/g3.114.010298

Huang, W., Richards, S., Carbone, M. A., Zhu, D., Anholt, R. R., Ayroles, J. F., et al. (2012). Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15553–15559. doi: 10.1073/pnas.1213423109

Jacquin, L., Elsen, J.-M., and Gilbert, H. (2014). Using haplotypes for the prediction of allelic identity to fine-map QTL: characterization and properties. *Genet. Select Evol.* 46:45. doi: 10.1186/1297-9686-46-45

Janson, L., Barber, R. F., and Candès, E. (2015). Eigenprism: inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*.

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *J. Stat. Softw.* 11, 1–20. doi: 10.18637/jss.v011.i09

Kimeldorf, G., and Wahba, G. (1971). Some results on tchebycheffian spline functions. *J. Math. Anal. Appl.* 33, 82–95. doi: 10.1016/0022-247X(71)90184-3

Konstantinov, K., and Hayes, B. (2010). "Comparison of blup and reproducing kernel hilbert spaces methods for genomic prediction of breeding values in australian holstein friesian cattle," in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Vol. 224, (Leipzig: CD-ROM Communication).

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. Ser. B (Methodol.)* 34, 1–41.

Liu, G.-F., Yang, J., and Zhu, J. (2006). Mapping QTL for biomass yield and its components in rice (*oryza sativa* L.). *Acta Genet. Sin.* 33, 607–616. doi: 10.1016/S0379-4172(06)60090-5

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

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *J. Mach. Learn. Res.* 7, 2651–2667.

Moore, J. H., and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* 85, 309–320. doi: 10.1016/j.ajhg.2009.08.006

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet.* 5:363. doi: 10.3389/fgene.2014.00363

Morota, G., Koyama, M., Rosa, G. J. M., Weigel, K. A., and Gianola, D. (2013). Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet. Sel. Evol.* 45:17. doi: 10.1186/1297-9686-45-17

Naim, I., and Gildea, D. (2012). Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427*.

Norton, G. J., Deacon, C. M., Xiong, L., Huang, S., Meharg, A. A., and Price, A. H. (2010). Genetic mapping of the rice ionome in leaves and grain: identification of qtls for 17 elements including arsenic, cadmium, iron and selenium. *Plant Soil* 329, 139–153. doi: 10.1007/s11104-009-0141-8

Onogi, A., Ideta, O., Inoshita, Y., Ebana, K., Yoshioka, T., Yamasaki, M., et al. (2015). Exploring the areas of applicability of whole-genome prediction methods for asian rice (*oryza sativa* L.). *Theor. Appl. Genet.* 128, 41–53. doi: 10.1007/s00122-014-2411-y

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3* 2, 1595–1605. doi: 10.1534/g3.112.003665

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* 6, 15–32. doi: 10.1214/ss/1177011926

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Number 12. Cambridge: Cambridge University Press.

Saunders, C., Gammerman, A., and Vovk, V. (1998). "Ridge regression learning algorithm in dual variables," in *(ICML-1998) Proceedings of the 15th*

International Conference on Machine Learning (San Francisco, CA: Morgan Kaufmann), 515–521.

Schaeffer, L. (2010). Linear models in animal breeding. Course at the centre for genetic improvement of livestock. *Univ. Guelph* 97–98. Available online at: http://www.aps.uoguelph.ca/~lrs/ABModels/DATA/EuropeNotes.pdf

Smola, A., and Schlkopf, B. (1998). A tutorial on support vector regression. *Neuro COL T2 Technical Report Series NC2-TR-1998-030*.

Sun, X., Ma, P., and Mumm, R. H. (2012). Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. *PLoS ONE* 7:e50604. doi: 10.1371/journal.pone.0050604

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* 58, 267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* 7, 1456–1490. doi: 10.1214/13-EJS815

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

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.