ORIGINAL RESEARCH

Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP

Jeffrey B. Endelman*

Abstract

Many important traits in plant breeding are polygenic and therefore recalcitrant to traditional marker-assisted selection. Genomic selection addresses this complexity by including all markers in the prediction model. A key method for the genomic prediction of breeding values is ridge regression (RR), which is equivalent to best linear unbiased prediction (BLUP) when the genetic covariance between lines is proportional to their similarity in genotype space. This additive model can be broadened to include epistatic effects by using other kernels, such as the Gaussian, which represent inner products in a complex feature space. To facilitate the use of RR and nonadditive kernels in plant breeding, a new software package for R called rrBLUP has been developed. At its core is a fast maximum-likelihood algorithm for mixed models with a single variance component besides the residual error, which allows for efficient prediction with unreplicated training data. Use of the rrBLUP software is demonstrated through several examples, including the identification of optimal crosses based on superior progeny value. In cross-validation tests, the prediction accuracy with nonadditive kernels was significantly higher than RR for wheat (*Triticum aestivum* L.) grain yield but equivalent for several maize (*Zea mays* L.) traits.

THE ABILITY TO PREDICT COMPLEX TRAITS from marker data is becoming increasingly important in plant breeding (Bernardo, 2008). The earliest attempts, now over 20 years old, involved first identifying significant markers and then combining them in a multiple regression model (Lande and Thompson, 1990). The focus over the last decade has been on genomic selection methods, in which all markers are included in the prediction model (Bernardo and Yu, 2007; Heffner et al., 2009; Jannink et al., 2010).

One of the first methods proposed for genomic selection was ridge regression (RR), which is equivalent to best linear unbiased prediction (BLUP) in the context of mixed models (Whittaker et al., 2000; Meuwissen et al., 2001). The basic RR-BLUP model is

$$\mathbf{y} = \mathbf{WGu} + \varepsilon, \qquad\qquad [1]$$

where $\mathbf{u} \sim N\left(0, \mathbf{I}\sigma_u^2\right)$ is a vector of marker effects, $\mathbf{G}$ is the genotype matrix (e.g., {aa,Aa,AA} = {–1,0,1} for biallelic single nucleotide polymorphisms (SNPs) under an additive model), and $\mathbf{W}$ is the design matrix relating lines to observations ($\mathbf{y}$). The BLUP solution for the marker effects can be written as either $\hat{\mathbf{u}} = \mathbf{Z}'\left(\mathbf{ZZ}' + \lambda\mathbf{I}\right)^{-1}\mathbf{y}$ or $\hat{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}'\mathbf{y}$, where $\mathbf{Z} = \mathbf{WG}$ and the ridge parameter $\lambda = \sigma_e^2 / \sigma_u^2$ is the ratio between the residual and marker variances (Searle et al., 2006). Compared with ordinary regression, for which the number of markers cannot exceed the number of observations, RR has no such limit and also has improved numerical stability

Dep. of Crop and Soil Sciences, Washington State Univ., 16650 State Route 536, Mount Vernon, WA 98273. Received 26 May 2011. *Corresponding author (j.endelman@gmail.com).

Abbreviations: θ_{REML}, restricted maximum likelihood solution for θ; BLR, Bayesian Linear Regression; BLUP, best linear unbiased prediction; EXP, exponential model; GAUSS, Gaussian model; GEBV, genomic-estimated breeding value; LL, log-likelihood; ML, maximum likelihood; REML, restricted maximum likelihood; RR, ridge regression; r_{pred}, cross-validation accuracy; r_{train}, training population accuracy; SNP, single nucleotide polymorphism.

when markers are highly correlated (Hoerl and Kennard, 2000).

There is a close connection between marker-based RR-BLUP (Eq. [1]) and kinship-BLUP, in which the performance of breeding lines is predicted based on their kinship to other germplasm (Bernardo, 1994; Piepho et al., 2008). The basic kinship-BLUP model is

$$\mathbf{y} = \mathbf{Wg} + \varepsilon$$

$$\mathbf{g} \sim N\left(0, \mathbf{K}\sigma_g^2\right), \qquad [2]$$

where $\mathbf{g}$ is a vector of genotypic values. In pedigree-based prediction of breeding values, $\mathbf{K}$ is the additive relationship matrix $\mathbf{A}$ derived from the coefficients of coancestry (Bernardo, 2010). These coefficients reflect the average behavior of alleles undergoing Mendelian segregation, but the actual segregation can be captured with the marker-based relationship matrix

$$\mathbf{K}_{RR} = \mathbf{GG}'. \qquad [3]$$

Equation [3] has the property that, for random populations, its expected value is proportional to $\mathbf{A}$ plus a constant (Habier et al., 2007); for this reason it has been called the realized (additive) relationship matrix. Another key property of $\mathbf{K}_{RR}$ is that the genomic-estimated breeding values (GEBVs) it produces ($\hat{\mathbf{g}}$ in Eq. [2]) are equivalent to those from the marker-based RR-BLUP approach ($\mathbf{G}\hat{\mathbf{u}}$ in Eq. [1]) (Hayes et al., 2009).

When using genomic selection to advance lines as varieties, it is not just the breeding (additive) value but the full genotypic value that is of interest (Piepho et al., 2008). Rather than modeling epistatic interactions directly, which is challenging because of the combinatorial complexity, an alternative approach is to capture them through an appropriate kernel function (Gianola and van Kaam, 2008; Piepho, 2009; de los Campos et al., 2010). The realized relationship model (Eq. [3]) is in fact a kernel in genotype space and can be written as $K_{ij} = \left\langle G_{i\bullet}, G_{j\bullet} \right\rangle$, where the angle brackets denote the inner (or dot) product between genotypes i and j. In geometry the inner product measures the similarity of two vectors, so with the additive relationship model the genetic covariance between lines is proportional to their similarity in genotype space.

This geometric formulation enables use of the so-called kernel "trick" in machine learning, which involves replacing the inner product in the original (genotype) space with an inner product in a more complex feature space, technically called a reproducing kernel Hilbert space (Schölkopf and Smola, 2002):

$$K_{ij} = K\left(G_{i\bullet}, G_{j\bullet}\right) = \left\langle \Phi\left(G_{i\bullet}\right), \Phi\left(G_{j\bullet}\right) \right\rangle, \qquad [4]$$

Equation [4] means that the kernel function K, which takes the two genotypes as arguments and returns a single number, equals the inner product between the genotypes in a feature space defined by Φ. Although one can construct kernels by first specifying Φ and then applying Eq. [4], this is unnecessary as the feature space is guaranteed to exist for any positive semidefinite kernel

(Schölkopf and Smola, 2002). To calculate BLUPs that include nonadditive effects, it is sufficient to solve Eq. [2] with $\mathbf{K}$ based on an appropriate kernel function (Gianola and van Kaam, 2008).

The objective of the present research was to develop an R package for genomic prediction based on a maximum likelihood (ML) or restricted maximum likelihood (REML) approach to ridge regression (RR) and other kernels. The result is rrBLUP (available at http://cran.r-project.org/web/packages/rrBLUP [verified 21 Nov. 2011]), which uses a fast spectral algorithm for mixed models with a single variance component besides the residual error (Kang et al., 2008). After demonstrating features of the software, the accuracy of its prediction methods are compared by cross-validation using structured populations of wheat (*Triticum aestivum* L.) (Crossa et al., 2010) and maize (*Zea mays* L.) (Yu et al., 2006).

MATERIALS AND METHODS

The wheat population consisted of 599 inbred lines genotyped at 1279 Diversity Array Technology (DArT) markers and was downloaded as part of the Bayesian Linear Regression (BLR) package for R, version 1.2 (Pérez et al., 2010). Single nucleotide polymorphism markers and phenotypic data for maize ear height, ear diameter, and male flowering time were downloaded from the TASSEL website (Bradbury et al., 2007). For each of the ten maize chromosomes, the diploid marker data were phased and missing alleles imputed using the software BEAGLE, version 3.3.1 (Browning and Browning, 2007). After removing monomorphic markers, 2953 remained. The population size was 279 inbred lines, but due to missing phenotypic data only 276 lines were available for flowering time and 249 for ear diameter.

For each of the 179,101 unique crosses between the 599 wheat lines, the expected mean and standard deviation (SD) for the GEBV of the recombinant inbred progeny were calculated based on the predicted marker effects in environment 1. In the absence of a linkage map, markers were assumed to segregate independently, which is clearly an approximation. (With a linkage map the SD could be simulated more realistically.) If p_{k+} and p_{k-} denote the frequency of the +1 and –1 alleles, respectively, at locus k in the parents, then the mean GEBV of the inbred progeny is $E[\hat{g}_i] = \sum_k E[G_{ik}]\hat{u}_k = \sum_k (p_{k+} - p_{k-})\hat{u}_k$, and the variance (neglecting uncertainty in the marker effects) is

$$\mathrm{Var}[\hat{g}_i] = \sum_k \mathrm{Var}[G_{ik}]\hat{u}_k^2 = \sum_k \left(E[G_{ik}^2] - E[G_{ik}]^2 \right)\hat{u}_k^2$$

$$= \sum_k \left[1 - \left(p_{k+} - p_{k-} \right)^2 \right]\hat{u}_k^2$$

Bayesian LASSO predictions were made with the BLR package for R, version 1.2, and hyperparameters were chosen based on the guidelines of Pérez et al. (2010). For the prior distribution of the residual variance, the degrees of freedom was $\mathrm{df}_\varepsilon = 3$ and the scale was $S_\varepsilon = (\mathrm{Var}[\mathbf{y}]/2)(2 + \mathrm{df}_\varepsilon)$, where $\mathrm{Var}[\mathbf{y}]$ is the variance of the training data. The prior distribution for the LASSO

shrinkage parameter had mode $\left(2\sum_k \bar{G}_{\bullet k}^2\right)^{1/2}$, where $\bar{G}_{\bullet k}$ is the average over the training data and the sum is over markers. The rate and shape hyperparameters were 2×10^{-5} and 0.52, respectively. A total of 10,000 iterations was used, with a burn-in period of 2000 iterations.

Statistical analysis of the cross-validation results was conducted with SAS PROC GLM (SAS Institute, 1994), with partition and method as fixed effects. The REGWQ option was used to control the strong familywise error rate (the probability of false discovery) at 0.05.

RESULTS AND DISCUSSION

Marker vs. Kinship-Based Prediction

At the core of the rrBLUP package is the function mixed.solve, which solves any mixed model of the form

$$y = X\beta + Zu + \varepsilon$$

$$u \sim N\left(0, K\sigma_u^2\right),\qquad [5]$$

where X is a full-rank design matrix for the fixed effects β, Z is the design matrix for the random effects u, K is a positive semidefinite matrix, and the residuals are normal with constant variance. Variance components are estimated by either ML or REML (default) using the spectral decomposition algorithm of Kang et al. (2008). The R function returns the variance components, the maximized log-likelihood (LL), the ML estimate for β, and BLUP solution for u.

It was stated in the introduction that when the realized relationship matrix GG' is used, the marker-based (Eq. [1]) and kinship-based (Eq. [2]) formulations of the prediction problem give equivalent GEBV. This can be verified numerically using mixed.solve and a set of 599 wheat lines from the BLR package for R (Pérez et al., 2010). The BLR variable Y contains the two-year average grain yield in four environments (standardized to zero mean and unit variance), and the genotype matrix is coded as {0,1} in the variable X. To be consistent with the notation in this article, the genotypes were recoded as {–1,1} in G:

```
library(rrBLUP) #load rrBLUP
library(BLR)     #load BLR
data(wheat)      #load wheat data
G <- 2*X  1      #recode genotypes
y <- Y[,1]       #yields from E1

#marker-based
ans1 <- mixed.solve(y=y,Z=G)

#kinship-based
K <- tcrossprod(G) #K = GG'
I <- diag(599)
ans2 <- mixed.solve(y=y,Z=I,K=K)

#Compare GEBV
cor (G%*%ans1$u, ans2$u) #equals 1
```

In the first call to mixed.solve the design matrix equals the genotype matrix, so the random effects are the marker effects. In this case K is an identity matrix, which the software assumes because no K variable is provided. When no design matrix for fixed effects is provided, as in this example, an intercept term is automatically included. In the second call to mixed.solve, an identity matrix is used for Z and the realized relationship matrix GG' is used for K. In this case the random effects are the breeding values, which in the last line of code are compared with the GEBV from the marker-based model. As shown in the comments, the correlation is exactly 1. Each of the two calls to mixed.solve took five seconds on a laptop computer with two gigabytes of memory, running R 2.13.1 (R Development Core Team, 2011).

Although the two approaches are equivalent for calculating GEBV, some analyses depend on knowing the marker effects. For example, when different lines are evaluated in different environments, even though a whole genotype × environment analysis is not possible, one can still study marker × environment interactions (Crossa et al., 2010).

Another application is to design crosses in a breeding program (Bernardo et al., 2006; Zhong and Jannink, 2007). The expected mean for the progeny can be calculated as the mean of the parental GEBV, but the marker effects are needed to compute the variance of the population, which is important for genetic gain. To illustrate, each circle in Fig. 1 shows the expected mean (μ) and standard deviation (σ) for the GEBV of recombinant inbred lines from one wheat cross. Results are shown for all 179,101 unique crosses between the 599 wheat lines, using the predicted marker effects in environment 1. In the upper right corner of the figure are crosses between lines with high GEBV and complementary alleles, for which high levels of transgressive segregation are expected.

For a given selection intensity i, the mean of the selected population is $\mu_s = \mu + i\sigma$, which Zhong and Jannink (2007) called the superior progeny value. The superior progeny values for the crosses in Fig. 1 were calculated for selection intensities ranging from 1.4 (20% selected) to 2.7 (1% selected). The top nine crosses were conserved across this range and are listed in Table 1, with lines identified by their GEBV rank. Exactly one of the top two highest-GEBV lines was found in every pair, but the 1×2 cross does not appear because the two lines share 96% of their alleles and have an expected SD of 0.07.

Kernels with Epistatic Effects

At present there are two kernels other than RR in the rrBLUP package. One is the Gaussian model (GAUSS):

$$K_{ij} = \exp[-(D_{ij}/\theta)^2], \qquad [6]$$

Where

$$D_{ij} = \left[(1/4M)\sum_{k=1}^{M}\left(G_{ik} - G_{jk}\right)^2\right]^{1/2} \qquad [7]$$

is the Euclidean distance between genotypes i and j, normalized to the interval [0,1]. The parameter θ is a scale parameter that influences how quickly the genetic
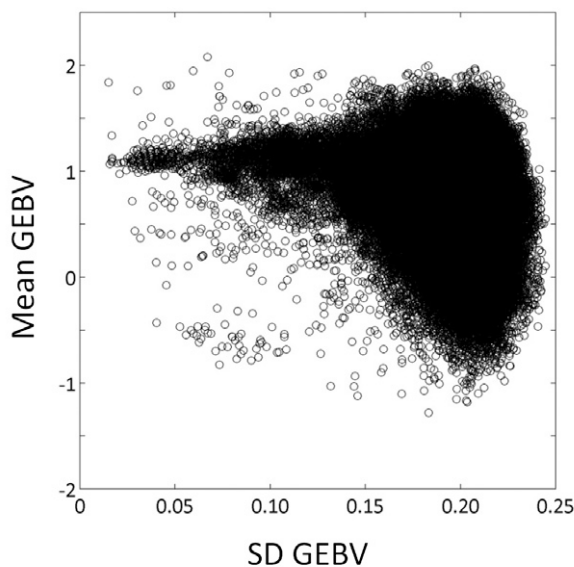
Figure 1. Analysis of line crosses. Each circle is the expected mean and standard deviation (SD) for the genomic-estimated breeding values (GEBVs) of the recombinant inbred progeny from one wheat cross. Results are shown for all 179,101 unique crosses between the 599 wheat lines, using the predicted marker effects in environment 1. In the top right of the figure are crosses between parents with high GEBV and complementary alleles, for which high levels of transgressive segregation are expected.

Table 1. Top nine wheat crosses based on superior progeny value (SPV) in environment 1.

Cross[†]	Kinship[‡]	SPV$_{20\%}$	SPV$_{1\%}$	Mean GEBV[§]	SD GEBV
1×4	0.57	2.261	2.524	1.971	0.207
1×5	0.57	2.260	2.522	1.970	0.207
1×3	0.69	2.256	2.487	2.000	0.183
2×4	0.58	2.245	2.507	1.954	0.208
2×5	0.58	2.243	2.506	1.953	0.208
2×3	0.69	2.236	2.466	1.982	0.181
1×7	0.57	2.227	2.486	1.940	0.205
1×12	0.60	2.210	2.481	1.910	0.214
2×7	0.59	2.209	2.469	1.923	0.205

[†]Line identifier equals the GEBV rank.

[‡]Fraction of shared alleles (identity by state).

[§]GEBV, genomic-estimated breeding value.

covariance decays with distance. The other kernel is the exponential model (EXP): $K_{ij} = \exp(-D_{ij}/\theta)$.

These kernels are available through the rrBLUP function kinship.BLUP, which was designed to predict the genotypic values of one population based on the genotypes and phenotypes of a second, training population. To illustrate its use, consider again the 599 wheat lines from the BLR package, which have been randomly partitioned into 10 sets for use in 10-fold cross-validation (Pérez et al., 2010). The variable sets contain the partition number for each line. To predict the genotypic values of set 1 using the other nine sets as the training population, the R code is

```
train <- which(sets!=1)
pred  <- which(sets==1)
ans.RR<-kinship.BLUP(y=y[train],
  G.train=G[train,],G.pred=G[pred,])
ans.GAUSS<-kinship.BLUP(y=y[train],
  G.train=G[train,],G.pred=G[pred,],
  K.method="GAUSS")

#accuracy with RR
cor(ans.RR$g.pred,y[pred])
#accuracy with GAUSS
cor(ans.GAUSS$g.pred,y[pred])
```

In the first call to kinship.BLUP the kernel method is not specified, so by default the realized relationship model is used. The last two lines of code calculate the correlation $\left(r_{gy}\right)$ between the predicted genotypic value and observed phenotype for the prediction population, which

measures the cross-validation accuracy of the prediction method.

Table 2 shows the accuracies of the two methods for all 10 sets in environments 1 and 2. The results demonstrate that the performance of GAUSS compared to RR depends on both the structure of the population and the phenotype. For 9 out of 10 sets in environment 1, the accuracy with GAUSS was higher than RR. The largest gap was for set 5, where the accuracy with RR was 0.34 vs. 0.51 with GAUSS. Across the 10 sets the mean accuracy with GAUSS was 0.58 vs. 0.51 for RR ($p = 0.009$ by paired t-test). By contrast, in environment 2 there was no significant difference between the prediction methods ($p = 0.2$).

To better understand these differences, Fig. 2 shows the log-likelihood (LL) (solid circles), training population accuracy (r_{train}) (dashed line), and cross-validation accuracy (r_{pred}) (open circles) as a function of the scale parameter θ (see Eq. [6]). The rrBLUP package uses REML (or ML) to identify the optimal scale parameter, and because the genotype distances have been normalized to the unit interval (Eq. [7]), this is also the essential range for θ. The two panels in Fig. 2 correspond to sets 5 and 6 in environment 1, which showed contrasting results in the RR vs. GAUSS comparison: for set 5 the accuracy with GAUSS was higher and vice versa for set 6 (see Table 2). In both cases the REML solution for θ (θ_{REML}) was similar and the r_{train} approached 1 as θ decreased to zero.

The crucial difference lies in r_{pred}. For set 5 r_{pred} exhibited an interior maximum near the θ_{REML} while for set 6 r_{pred} was maximized at $\theta = 1$ and declined steadily as θ decreased. The significance of this observation for understanding Table 2 is that GAUSS behaves like RR when θ is large relative to $\mathbf{D}$. This follows from the Taylor series expansion, $K_{ij} = 1 - \left(D_{ij}/\theta\right)^2 + 1/2\left(D_{ij}/\theta\right)^4 + \cdots$, and the fact that $[D_{ij}^2]$ is equivalent to the additive model $\mathbf{GG'}$ for inbred lines (Piepho, 2009). As θ decreases, the epistatic interactions in the higher order terms (e.g., D_{ij}^4) become more important. When r_{pred} has an interior maximum near θ_{REML}, as in set 5, GAUSS will have higher accuracy than RR. When r_{pred} increases monotonically with θ,

GAUSS will not have higher accuracy than RR; whether GAUSS is lower or equivalent depends on the shape of the LL profile. In the case of set 6, the LL profile peaked at $\theta_{REML} = 0.4$, so RR had higher accuracy. For most sets in environment 2, both LL and r_{pred} increased monotonically with θ (not shown), so GAUSS and RR were equivalent.

These phenomena are relevant to the question of whether GAUSS is prone to overfitting, which Piepho (2009) and Heslot et al. (2012) have raised as a concern. In both studies the residual error with GAUSS was much smaller than with RR, or equivalently the accuracy for the training population was nearly 1. This was also observed with the BLR wheat data, as shown by the dashed line in Fig. 2. To constitute overfitting, however, there must be a tradeoff between higher accuracy for the training set and lower accuracy for the validation set (Dietrich, 1995). The results in Heslot et al. (2012) and the present study show that such a tradeoff is rare provided the scale parameter is chosen properly. Overfitting was observed for set 6 in environment 1, but more typically r_{pred} was either the same or higher with GAUSS compared to RR (see Table 2).

To investigate the matter further, a different data set—279 maize lines genotyped at 2953 SNP markers—was analyzed with the rrBLUP package. The cross-validation accuracies for maize flowering time, ear height, and ear diameter are shown alongside the results for wheat grain yield in Table 3. For wheat grain yield, the accuracy with GAUSS was 6 to 7 percentage points higher than RR in every environment but environment 2 (similar to Crossa et al. [2010]). For all three maize traits there was no significant difference between GAUSS and RR, which provides additional evidence that overfitting (i.e.,

a loss in cross-validation accuracy) is not common with GAUSS. The results also suggest that most (perhaps all) of the genetic variation was additive for the maize traits.

Table 3 includes the cross-validation results with EXP, which was equivalent to GAUSS for all seven traits. Piepho (2009) also found little difference between these two models in his analysis of maize grain yield. Like GAUSS, EXP captures nonadditive effects but the structure of its feature space is different. For the limited plant breeding data analyzed thus far with the two methods, this difference appears to be of little consequence.
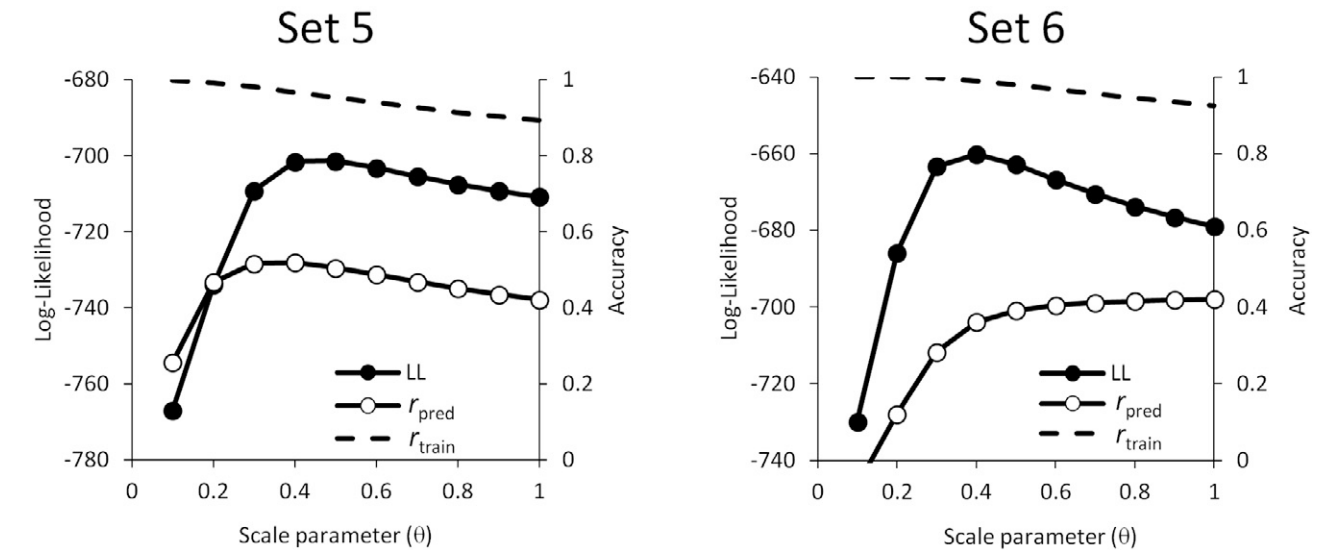
Table 2. Cross-validation accuracies ($r_{\hat{g}y}$) for wheat grain yield.

Set[†]	Environment 1		Environment 2	
	RR[‡]	GAUSS[§]	RR	GAUSS
1	0.49	0.61	0.37	0.37
2	0.44	0.52	0.49	0.51
3	0.41	0.44	0.48	0.49
4	0.64	0.69	0.42	0.43
5	0.34	0.51	0.31	0.31
6	0.43	0.36	0.59	0.60
7	0.64	0.71	0.54	0.55
8	0.54	0.66	0.62	0.63
9	0.57	0.62	0.42	0.44
10	0.65	0.69	0.56	0.53
Mean	0.51	0.58**	0.48	0.49

**Means significantly different at the 0.01 probability level in Environment 1.

[†]Prediction set; the other nine sets were used for training.

[‡]RR, ridge regression.

[§]GAUSS, Gaussian model.

Figure 2. Performance of the Gaussian model (GAUSS). The figure depicts the effect of the Gaussian scale parameter (θ in Eq. [6]) on the restricted log-likelihood (LL), the training population accuracy (r_{train}), and the cross-validation accuracy (r_{pred}) when predicting sets 5 or 6 in environment 1. For set 5 the restricted maximum likelihood solution for θ (θ_{REML}) = 0.5, and for set 6 θ_{REML} = 0.4. In both cases r_{train} approached 1 as $\theta \to 0$, but the trends for r_{pred} were different. For set 5 r_{pred} exhibited an interior maximum near θ_{REML}, while for set 6 r_{pred} increased monotonically with θ. Because GAUSS is approximately ridge regression (RR) when θ is large, the contrasting behavior in this figure illustrates why GAUSS had higher r_{pred} than RR for set 5 but vice versa for set 6 (see Table 2).

Table 3. Tenfold cross-validation accuracy ($r_{\hat{g}y}$) for maize and wheat traits.

Method[†]	Wheat yield 1	Wheat yield 2	Wheat yield 3	Wheat yield 4	Maize flowering time	Maize ear height	Maize ear diameter
GAUSS	0.58[a‡]	0.49[a]	0.45[a]	0.54[a]	0.73[a]	0.51[a]	0.53[ab]
EXP	0.57[a]	0.49[a]	0.45[a]	0.54[a]	0.73[a]	0.54[a]	0.54[a]
RR	0.51[b]	0.48[a]	0.38[b]	0.48[b]	0.73[a]	0.51[a]	0.52[b]
BL	0.51[b]	0.48[a]	0.38[b]	0.47[b]	0.73[a]	0.52[a]	0.53[ab]

[†]GAUSS, Gaussian model; EXP, exponential model; RR, ridge regression; BL, Bayesian LASSO.

[‡]Within each trait, accuracies with the same letter were not significantly different at the 0.05 probability level.

For the sake of comparison, Table 3 also shows the accuracy of the additive Bayesian LASSO model, which was equivalent to RR for all seven traits.

CONCLUSIONS

The objective of this research was to create software that makes ridge regression and other kernel methods accessible to plant breeders interested in genomic selection. At the core of the rrBLUP package is the function mixed.solve, which can be used to solve both the marker-based and kinship-based versions of the genomic prediction problem. The function kinship.BLUP provides a more intuitive interface for kinship-based prediction and includes several genetic models, including an additive relationship matrix and the nonadditive Gaussian kernel.

References

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34:20–25.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48:1649–1664.

Bernardo, R. 2010. Quantitative traits in plant breeding. Stemma Press, Woodbury, MN.

Bernardo, R., L. Moreau, and A. Charcosset. 2006. Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. Crop Sci. 46:1972–1980.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Available at http://www.maizegenetics.net/tassel (verified 21 Nov. 2011). Bioinformatics 23:2633–2635.

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097.

de los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. Camb. 92:295–308.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724.

Dietrich, T. 1995. Overfitting and undercomputing in machine learning. ACM Comput. Surv. 27:326–327.

Gianola, D., and J.B. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. Camb. 91:47–60.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12.

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

Hoerl, A.E., and R.W. Kennard. 2000. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 42:80–86.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomic 9:166–177.

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

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

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

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. Plant Gen. 3:106–116.

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165–1176.

Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

SAS Institute. 1994. SAS 9.2 for Windows. SAS Institute, Cary, NC.

Schölkopf, B., and A.J. Smola. 2002. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA.

Searle, S.R., G. Casella, and C.E. McCulloch. 2006. Variance components. John Wiley & Sons, Hoboken, NJ.

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. Camb. 75:249–252.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.

Zhong, S., and J.-L. Jannink. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics 177:567–576.