Genomic-enabled prediction based on molecular markers and pedigree

using the BLR package in R

Paulino Pérez[**], Gustavo de los Campos[*,**], José Crossa, and Daniel Gianola

P. Pérez, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, México D.F., México, and Colegio de Postgraduados, Km. 36.5 Carretera México, Texcoco, Montecillo, Estado de México, 56230, México; G. De los Campos, University of Wisconsin-Madison, 1675 Observatory Dr. WI 53706, US (present address University of Alabama at Birmingham, 1665 University Boulevard, Ryals Public Health Building 414, Birmingham, AL 35294 US); J. Crossa, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, México D.F., México; D. Gianola, University of Wisconsin-Madison, 1675 Observatory Dr. WI 53706, US. *Corresponding author (gcampos@ms.soph.uab.edu). [**] The first two authors made an equal contribution.

Abbreviations: GS, Genomic selection; BLR, Bayesian Linear Regression; RR, ridge regression; BRR, Bayesian Ridge Regression; BL, Bayesian LASSO; BLUP, Best Linear Unbiased Prediction.

Abstract

The availability of molecular markers has made possible the use of genomic selection in plant and animal breeding. However, models for genomic selection pose several computational and statistical challenges and require specialized computer programs not always available to the end user and not implemented in standard statistical software yet. The R BLR (Bayesian Linear Regression) package implements several statistical procedures (i.e., Bayesian Ridge Regression, Bayesian Lasso) in a unified framework that allows including marker genotypes and pedigree data jointly. This article describes the classes of models implemented in the BLR package and illustrates their use through examples using simulated and real data from plant breeding trials. Also addressed are some challenges faced when applying genomic-enabled selection, such as model choice, evaluation of predictive ability through cross-validation, and choice of hyper-parameters.

Introduction

Prediction of genetic values is a central problem in quantitative genetics. Over many decades such predictions have been obtained using phenotypic and family data, the latter usually represented by a pedigree. Accurate predictions of genetic values of genotypes whose phenotypes are yet to be observed (e.g., newly developed lines) are needed to attain rapid genetic progress and reduce phenotyping costs (e.g., Bernardo and Yu, 2007). However, pedigree-based models typically do not account for Mendelian segregation, a term that in an additive model and in the absence of inbreeding explains as much as one half of the genetic variance. This sets an upper limit on the accuracy of estimates of genetic values of individuals without progeny. Dense molecular markers (MM) are now available in the genome of humans and of many plant and animal species. Unlike pedigree data, MM allow tracing back Mendelian segregation events at many points along the genome. Potentially, this information can be used to improve the accuracy of estimates of genetic values of newly developed lines.

Following the ground-breaking contribution of Meuwissen et al. (2001), genomic selection (GS) has gained ground in plant and animal breeding (e.g., Bernardo and Yu, 2007; Hayes et al., 2009; VanRaden et al., 2009; de los Campos et al., 2009; Crossa et al., 2010). In practice, implementing GS involves analyzing large amounts of phenotypic and MM data, and requires specialized computer programs. The main purpose of this article is to show how the R (R Development Core Team, 2009) BLR (Bayesian Linear Regression) package can be used to implement several models for GS. A first version of the algorithms and the R-code implemented in the package was presented in de los Campos et al. (2009). The package was developed further and its performance was significantly improved by the first two authors of this article; it is now freely available through the R website. We provide a brief overview of parametric models for GS and describe the type of models implemented in BLR. We show two applications that illustrate the use of the package and several features of the models implemented in it. The package and data set are available from the R website, http://www.r-project.org.

Parametric models for genomic selection

In parametric models for GS (e.g., Meuwissen et al., 2001), phenotypic outcomes, y_i $(i = 1,...,n)$, are regressed on marker covariates x_{ij} $(j = 1,...,p)$ using a linear model of the form $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, where β_j is the regression of y_i on the jth marker covariate and ε_i is a model residual. In matrix notation, the regression model is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = \{y_i\}$, $\mathbf{X} = \{x_{ij}\}$, $\boldsymbol{\beta} = \{\beta_j\}$ and $\boldsymbol{\varepsilon} = \{\varepsilon_i\}$. The number of molecular markers (p) is usually larger than the number of observations (n) and, because of this, estimation of marker effects via multiple regression by ordinary least squares (OLS) is not feasible. Instead, penalized estimation methods such as ridge regression (RR, Hoerl and Kennard, 1970), or LASSO, or Bayesian methods such as those of Meuwissen et al. (2001) or the Bayesian LASSO (BL) of Park and Casella (BL, 2008) (e.g., Yi and Xu, 2008; de los Campos et al., 2009) can be used to estimate marker effects.

In RR, estimates of the effects of MM are obtained as the solution to the following optimization problem: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \tilde{\lambda} \sum_{j=1}^{p} \beta_j^2 \right\}$. Here, $\tilde{\lambda} \geq 0$ is a regularization parameter controlling the trade-offs between goodness of fit, as measured by the residual sum of squares, and model complexity, with the latter measured by the sum of squared marker effects, $\sum_{j=1}^{p} \beta_j^2$. The first order conditions of the above optimization problem are satisfied by $\left[\mathbf{X}'\mathbf{X} + \tilde{\lambda}\mathbf{I}\right]\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$; equivalently, $\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X} + \tilde{\lambda}\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{y}$. Relative to OLS, RR adds a constant, $\tilde{\lambda}$, to the diagonal of the matrix of coefficients; this makes the solution unique and shrinks estimates of marker effects towards zero, with the extent of shrinkage increasing as $\tilde{\lambda}$ increases (with $\tilde{\lambda} = 0$ the solution to the above problem is the OLS estimate of marker effects).

From a Bayesian perspective, $\hat{\boldsymbol{\beta}}$ can be viewed as the conditional posterior mode in a model with Gaussian likelihood and IID (independent and identically distributed) Gaussian marker effects, that is,

$$\begin{cases} \text{Likelihood}: & p\left(\mathbf{y}\middle|\boldsymbol{\beta},\sigma_\varepsilon^2\right)=\prod_{i=1}^{n} N\left(y_i\middle|\sum_{j=1}^{p} x_{ij}\beta_j,\sigma_\varepsilon^2\right) \\ \text{Prior}: & p\left(\boldsymbol{\beta}\middle|\sigma_\beta^2\right)=\prod_{i=1}^{n} N\left(\beta_j\middle|0,\sigma_\beta^2\right) \end{cases} \qquad [1]$$

Above, σ_β^2 is the variance of marker effects, a priori. The prior distribution of marker effects becomes increasingly informative and concentrated around zero as σ_β^2 decreases. The posterior mean and mode of $\boldsymbol{\beta}$ from [1] is equal to the ridge-regression estimate, with $\tilde{\lambda} = \sigma_\beta^{-2}\sigma_\varepsilon^2$, that is, $E\left(\boldsymbol{\beta}\middle|\mathbf{y}\right)=\hat{\boldsymbol{\beta}}=\left[\mathbf{X}'\mathbf{X}+\sigma_\beta^{-2}\sigma_\varepsilon^2\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{y}$. Therefore, prediction of genetic values from [1] can be obtained as

$$\begin{aligned} E\left(\mathbf{X}\boldsymbol{\beta}\middle|\mathbf{y},\sigma_\varepsilon^2,\sigma_\beta^2\right) &= \mathbf{X}E\left(\boldsymbol{\beta}\middle|\mathbf{y},\sigma_\varepsilon^2,\sigma_\beta^2\right) \\ &= \mathbf{X}\left[\mathbf{X}'\mathbf{X}+\sigma_\varepsilon^2\sigma_\beta^{-2}\mathbf{I}\right]^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \qquad [2a]$$

Alternatively, from [1] and using properties of the multivariate normal distribution, one has:

$$\begin{aligned} E\left(\mathbf{X}\boldsymbol{\beta}\middle|\mathbf{y},\sigma_\varepsilon^2,\sigma_\beta^2\right) &= Cov\left(\mathbf{X}\boldsymbol{\beta},\mathbf{y}'\right)Var\left(\mathbf{y}\right)^{-1}\mathbf{y} \\ &= \mathbf{X}\mathbf{X}'\sigma_\beta^2\left[\mathbf{X}\mathbf{X}'\sigma_\beta^2+\sigma_\varepsilon^2\mathbf{I}\right]^{-1}\mathbf{y} \\ &= \mathbf{X}\mathbf{X}'\left[\mathbf{X}\mathbf{X}'+\sigma_\varepsilon^2\sigma_\beta^{-2}\mathbf{I}\right]^{-1}\mathbf{y} \end{aligned} \qquad [2b]$$

Formulae [2a] and [2b] are equivalent and correspond the Best Linear Unbiased Predictor (BLUP) under the model defined by [1]. Note that [2a] and [2b] require solving systems of *p* and *n* equations, respectively. Therefore, with *p*>>*n* expression [2b] is computationally more convenient. However, unlike [2a], expression [2b] does not yield estimates of marker effects.

In RR-BLUP, estimates of marker effects are penalized to the same extent, and this may not be appropriate if some markers are located in regions not associated with genetic variance whereas others are linked to QTLs (Goddard and Hayes, 2007). To overcome this limitation, methods performing variable selection and shrinkage (e.g., the Least Absolute Value Selection and Shrinkage Operator, LASSO, Tibshirani, 1996) or Bayesian methods using marker-specific shrinkage of effects, such as methods BayesA of Meuwissen et al. (2001) or the BL of Park and Casella (2008), have been proposed.

The BLR package implements Bayesian regression with marker-specific (BL) or marker homogenous (Bayesian RR) shrinkage of estimates effects. The package allows inclusion of covariates other than markers, and regressions on a pedigree as well. In BLR, phenotypes are expressed as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_F\boldsymbol{\beta}_F + \mathbf{Z}\mathbf{u} + \mathbf{X}_R\boldsymbol{\beta}_R + \mathbf{X}_L\boldsymbol{\beta}_L + \boldsymbol{\varepsilon} \qquad [3]$$

where $\mathbf{y}$, the response, is an (n×1) vector (missing values are allowed); μ is an intercept; $\mathbf{X}_F = \{x_{Fij}\}_{n\times p_F}$, $\mathbf{Z} = \{z_{ij}\}_{n\times p_U}$, $\mathbf{X}_R = \{x_{Rij}\}_{n\times p_R}$ and $\mathbf{X}_L = \{x_{Rij}\}_{n\times p_L}$ are incidence matrices for the vectors of effects $\boldsymbol{\beta}_F$, $\mathbf{u}, \boldsymbol{\beta}_R$ and $\boldsymbol{\beta}_L$, whose dimensions are p_F, p_U, p_R and p_L, respectively. These vectors of effects differ with respect to the prior distributions assigned, as discussed later on. Finally, $\boldsymbol{\varepsilon}$ is a vector of model residuals assumed to be distributed as $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, Diag\left\{\frac{\sigma_\varepsilon^2}{w_i^2}\right\}\right)$,

where σ_ε^2 is an (unknown) variance parameter and the w_i's are (known) weights that allow for heterogeneous-residual variances. From these assumptions, the conditional distribution of the data, given the location effects, the residual variance and the weights, is:

$$p(\mathbf{y}|\mu,\boldsymbol{\beta}_F,\mathbf{u},\boldsymbol{\beta}_R,\boldsymbol{\beta}_L)=\prod_{i=1}^{n} N\left(y_i|\mu+\sum_{j=1}^{p_F} x_{Fij}\beta_{Fj}+\sum_{j=1}^{p_U} z_{ij}u_j+\sum_{j=1}^{p_R} x_{Rij}\beta_{Rj}+\sum_{j=1}^{p_L} x_{Lij}\beta_{Kj},\frac{\sigma_\varepsilon}{w_i^2}\right) \qquad [4]$$

Any of the elements on the right-hand side of [3], except μ and $\boldsymbol{\varepsilon}$, can be excluded in BLR; by default, the program runs an intercept model, i.e., $y_i = \mu + \varepsilon_i$. If weights are not provided, all weights are set equal to one.

The Bayesian model is completed by assigning a prior to the collection of model unknowns, $\boldsymbol{\theta} = \{\mu, \boldsymbol{\beta}_F, \mathbf{u}, \sigma_u^2, \boldsymbol{\beta}_R, \sigma_{\beta_R}^2, \boldsymbol{\beta}_L, \boldsymbol{\tau}^2, \lambda, \sigma_\varepsilon^2\}$. The intercept, μ, and the vector of 'fixed' effects, $\boldsymbol{\beta}_F$, are assigned flat priors, that is, $p(\mu,\boldsymbol{\beta}_F) \propto cons\tan t$. This treatment yields posterior means of these unknowns that are similar to those obtained with OLS, provided only μ and β_F are fitted.

The vector $\mathbf{u}$ is modeled using the standard assumptions of the infinitesimal additive model (e.g., Fisher, 1918; Wright, 1921; Henderson, 1975), that is, $p(\mathbf{u}|\sigma_u^2) = N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is a positive-definite matrix (usually a numerator relationship matrix computed from a pedigree) and σ_u^2 is an unknown variance, whose prior is an Inverse-Chi square density with degrees of freedom df_u and scale S_u, that is, $\sigma_u^2 \sim \chi^{-2}(\sigma_u^2|df_u, S_u)$; the hyper-parameters are user provided In the parameterization used in BLR, $E(\sigma_u^2 \mid df_u, S_u) = S_u(df_u - 2)^{-1}$. The multivariate normal prior assigned to **u** induces shrinkage of estimates of effects u_j, towards zero to an extent depending on the ratio $\sigma_u^2 \sigma_\varepsilon^{-2}$ and on the amount of inbreeding, as well as borrowing of information between pairs of levels of the random effect, $\{u_j, u_{j'}\}$, depending on the relationships between individuals in the pedigree.

The vector of regression coefficients $\boldsymbol{\beta}_R$ is assigned a Gaussian prior with variance common to all effects, that is, $p\left(\boldsymbol{\beta}_R \middle| \sigma_{\beta_R}^2\right) = \prod_{j=1}^{p_R} N\left(\beta_{Rj} \middle| 0, \sigma_{\beta_R}^2\right)$. This prior induces estimates that are the Bayesian counterpart of those obtained with Ridge Regression; we refer to this as 'Bayesian Ridge Regression' (BRR). The variance parameter, $\sigma_{\beta_R}^2$, is treated as unknown and is assigned a scaled inverse-χ^2 prior density, that is, $p\left(\sigma_{\beta_R}^2\right) = \chi^{-2}\left(\sigma_{\beta_R}^2 \middle| df_{\beta_R}, S_{\beta_R}\right)$ with degrees of freedom, df_{β_R}, and scale, S_{β_R}, provided by the user.

The vector of regression coefficients $\boldsymbol{\beta}_L$ is treated as in the Bayesian LASSO of Park and Casella (2008); the conditional prior distribution of marker effects, $p\left(\boldsymbol{\beta}_L \middle| \boldsymbol{\tau}^2, \sigma_\varepsilon^2\right)$, is Gaussian with marker-specific prior variances, that is, $p\left(\boldsymbol{\beta}_L \middle| \boldsymbol{\tau}^2, \sigma_\varepsilon^2\right) = \prod_{j=1}^{p_L} N\left(\beta_{Lj} \middle| 0, \sigma_\varepsilon^2 \tau_j^2\right)$. Unlike BRR, this induces marker-specific shrinkage of estimates of effects, whose extent depends on τ_j^{-2}. The variance parameters, τ_j^2, are assigned exponential IID priors, $p\left(\boldsymbol{\tau}^2 \middle| \lambda\right) = \prod_{j=1}^{p_L} Exp\left(\tau_j^2 \middle| \lambda^2\right)$. Finally, in the BLR package, the prior distribution of the regularization parameter λ, $p(\lambda)$, can be:

(a) a mass-point at some value (i.e., fixed λ),

(b) $p\left(\lambda^2\right) \sim Gamma(r,\delta)$t as suggested by Park and Casella (2008), or

(c) $p(\lambda|\max,\alpha_1,\alpha_2) \propto Beta\left(\frac{\lambda}{\max}\middle|\alpha_1,\alpha_2\right)$ (de los Campos et al., 2009).

With the above assumptions, the marginal prior of regression coefficients β_{Lj}, $p(\beta_{Lj}|\lambda) = \int N(\beta_{Lj}|0,\sigma_\varepsilon^2\tau_j^2)Exp(\tau_j^2|\lambda^2)\partial\tau_j^2$, is Double-Exponential (DE). Figure 1 displays the Gaussian and Double-Exponential density functions of random variables with zero mean and unit variance. Relative to the Gaussian, the DE process places a higher density at zero and thicker tails, inducing stronger shrinkage of estimates for markers with relatively small effect and less shrinkage of estimates for markers with sizable effect.

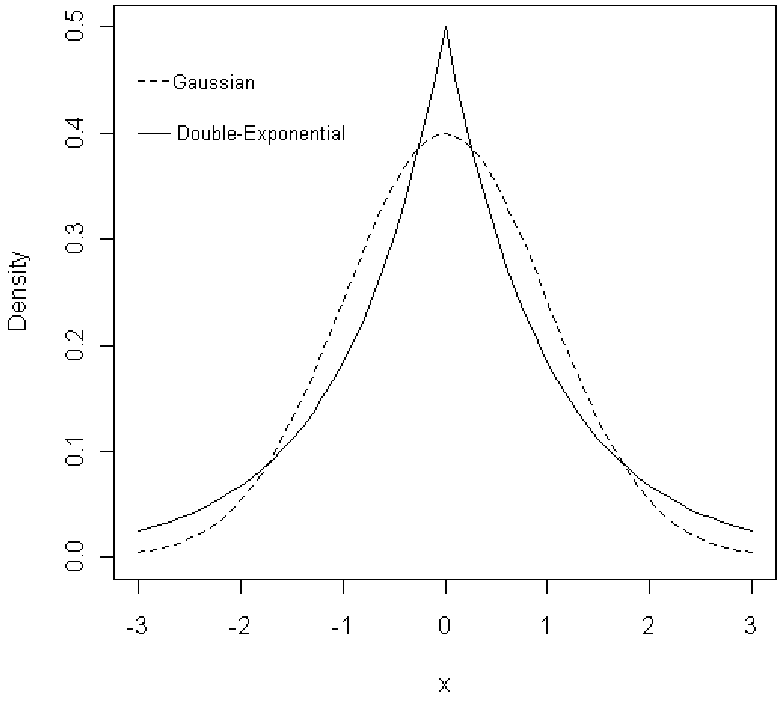

Figure 1. Gaussian and Double-Exponential densities, both for a random variable with zero mean and unit variance.

Finally, the residual variance is assigned a scaled inverse-χ^2 prior density with degrees of freedom, df_ε, and scale parameter provided by the user, S_ε, that is, $p\left(\sigma_\varepsilon^2\right) = \chi^{-2}\left(\sigma_\varepsilon^2 \middle| df_\varepsilon, S_\varepsilon\right)$.

Collecting the aforementioned assumptions, the prior distribution in BLR is:

$$\begin{aligned} p(\boldsymbol{\theta}) \propto\ & N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{A}\sigma_u^2\right)\prod_{j=1}^{p_R} N\left(\beta_{Rj}\middle|0,\sigma_{\beta_R}^2\right)\chi^{-2}\left(\sigma_{\beta_R}^2\middle|df_{\beta_R},S_{\beta_R}\right) \\ & \times\left\{\prod_{j=1}^{p_L} N\left(\beta_{Lj}\middle|0,\sigma_\varepsilon^2\tau_j^2\right)\prod_{j=1}^{p_L} Exp\left(\tau_j^2\middle|\lambda^2\right)\right\} p(\lambda)\chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right) \end{aligned}$$

The prior distribution is indexed by several hyper-parameters; in the Appendix we provide guidelines for choosing these parameters based on prior expectations about the proportion of phenotypic variance that can be attributed to each of the components on the right-hand side of [3].

The posterior distribution of model unknowns is proportional to the product of the likelihood and the prior distribution, that is:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) \propto\ & \prod_{i=1}^{n} N\left(y_i\middle|\mu+\sum_{j=1}^{p_F} x_{Fij}\beta_{Fj}+\sum_{j=1}^{p_U} z_{ij}u_j+\sum_{j=1}^{p_R} x_{Rij}\beta_{Rj}+\sum_{j=1}^{p_L} x_{Lij}\beta_{Kj},\frac{\sigma_\varepsilon}{w_i^2}\right) \\ & \times N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{A}\sigma_u^2\right) \\ & \times \prod_{j=1}^{p_R} N\left(\beta_{Rj}\middle|0,\sigma_{\beta_R}^2\right)\chi^{-2}\left(\sigma_{\beta_R}^2\middle|df_{\beta_R},S_{\beta_R}\right) \\ & \times \prod_{j=1}^{p_L} N\left(\beta_{Lj}\middle|0,\sigma_\varepsilon^2\tau_j^2\right)\prod_{j=1}^{p_L} Exp\left(\tau_j^2\middle|\lambda^2\right)p(\lambda) \\ & \times \chi^{-2}\left(\sigma_\varepsilon^2\middle|df_\varepsilon,S_\varepsilon\right) \end{aligned} \qquad [5]$$

This posterior distribution does not have a closed form; however, a Gibbs sampler can be used to draw samples from it. The Gibbs sampler is as in de los Campos et al. (2009), but extended to accommodate “fixed” effects and BRR.

Using BLR

This section gives two examples that illustrate the use of the BLR package and describe features of the models. It is assumed that the reader is familiar with the R-language/environment. In Example 1, we study the impact of different shrinkage methods (BRR vs BL) using simulated data. Example 2 illustrates how the package can be used to implement cross-validation (CV) using Bayesian methods. Cross-validation can be used for model comparison (e.g., to compare predictive ability of pedigree-based models versus marker-based models) or for selecting model parameters of the model (e.g., λ in the BL). Both examples make use of a wheat data set made available with the package, whose main features are described next.

Wheat data set

This data , used by Crossa et al. (2010), contains a historical set of 599 wheat lines from CIMMYT’s Global Wheat Program that were genotyped for 1447 Diversity Array Technology markers (DArT, Triticarte Pty. Ltd, Canberra, Australia; http://www.triticarte.com.au) and evaluated for grain yield (GY) in four macro-environments. The dataset becomes available in the R environment by running the following R-code:

```
library(BLR)
data(wheat)
```

Function library()loads the package, and data() loads datasets included in the package into the environment. The above code also loads the following objects (type objects() in the R console to list them) into the environment:

- **Y**, a matrix (599×4) containing the 2-year average grain yield of each of these lines in each of the four environments (phenotypes were standardized to a sample variance equal to one in each of the environments);

- **A** (599×599) is a numerator relationship matrix computed from a pedigree that traced back many generations. This relationship matrix was derived using the Browse application of the International Crop Information System (ICIS), as described in http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse (McLaren, 2005);

- **X** (599×1279) is a matrix with DArT genotypes; data are from pure lines and genotypes were coded as 0/1 denoting the absence/presence of the DArT. Markers with a minor allele frequency lower than 0.05 were removed, and missing genotypes were imputed with samples from the marginal distribution of marker genotypes, that is, x_{ij} = Bernoulli($\hat{p}_j$), where $\hat{p}_j$ is the estimated allele frequency computed from the non-missing genotypes. The number of DArT MMs after edition was 1279.

- **sets** (599×1) is a vector that assigns observations to 10 disjoint sets; the assignment was generated at random. This is used later to conduct a 10-fold CV.

Pedigree information was included in addition to marker data.

Example 1: The nature of different shrinkage methods

As stated, BRR or BL use different priors for marker effects; this induces different types of shrinkage of estimates of such effects. The simulation presented in this section aims at illustrating these differences. Data were generated using marker genotypes from the wheat dataset (**X**) with marker effects and model residuals simulated as described below.

Data simulation

Data were simulated under an additive model of the form,

$y_i = \mu + \sum_{j=1}^{1279} x_{ij}\beta_j + \varepsilon_{ij}$, $i = 1,...,599$, where $\mu = 100$ is an effect common to all individuals; $\{x_{ij}\}$ are marker genotypes from a collection of wheat lines described previously; $\{\beta_j\}$ are marker effects; and $\varepsilon_i \sim \mathrm{N}(\varepsilon_i|0,1)$ are IID standard normal residuals. We assumed that most markers (1267) had a relatively small effect and that only a few (12) had a sizable effect. Specifically, marker effects were sampled from the following mixture model:

$$\beta_j = \begin{cases} \text{Uniform}(-0.5,-0.8) & \text{if } j \in \{3,235,467,699,931,1163\} \\ \text{Uniform}(0.5,0.8) & \text{if } j \in \{119,351,583,815,1047,1279\} \\ \mathrm{N}(0,\mathrm{k}) & \text{otherwise.} \end{cases}$$

where $\mathrm{k} = 1279^{-1}$. The R-code used to implement this simulation was:

```
set.seed(12345)
data(wheat)
p<-ncol(X)
nQTL<-12
n<-nrow(X)
b0<-rnorm(p,sd=1/sqrt(p))
isQTL<-seq(from=3,to=p,length=nQTL)
b0[isQTL]<-rep(c(-1,1),times=nQTL/2)*runif(min=.5,max=.8,n=12)
yHat0<-100+X%*%b0
e0<-rnorm(n,sd=1)
y<-yHat0+e0
```

Function set.seed() initializes the random number generator; **X** is the matrix with information on molecular markers. Functions rnorm() and runif()generate random draws from the uniform and normal distributions, respectively. Figure 2 shows realized marker effects obtained with the above R-code. In this example, the sample variance of phenotypes (y) was 1.78, and the ratio of the sample variance of genetic values relative to phenotypic variance was 0.43.

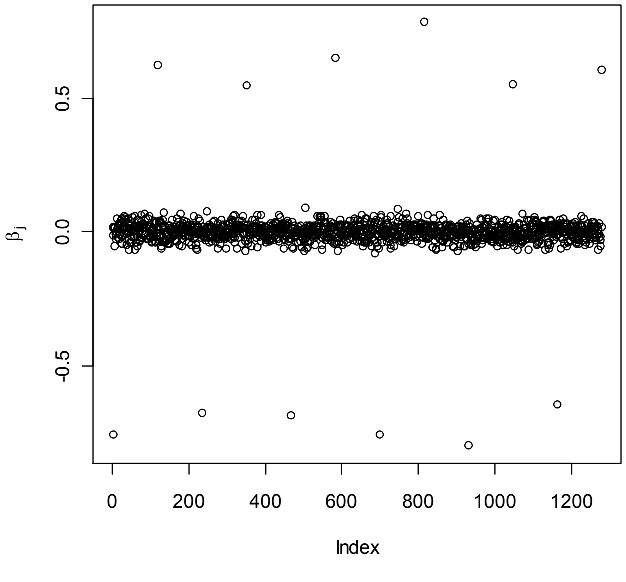

Figure 2. Realized marker effects.

Choice of hyper-parameters

The Appendix provides guidelines on how to choose values of hyper-parameters. It is assumed that the user has prior beliefs about the proportion of phenotypic variance that can be attributed to each of the components of the regression. In the simulation, the variance of phenotypes was about 1.8 and the variance of model residuals was 1. In practice, one does not know the true proportion of phenotypic variance that can be assigned to the genetic signal and model residuals, unless a precise estimate of heritability is available. Suppose our prior belief is that 50% of the phenotypic variance can be attributed to the genetic signal. Using $df_{\varepsilon} = 3$ and $\sigma_{\varepsilon}^{2} = 0.9$ in formula [1A] of the appendix, we set $S_{\varepsilon} = 4.5$. Values of hyper-parameters of the prior distribution of $\sigma_{\beta_R}^{2}$ and λ^{2} can be chosen using formulae [3A] and [4A] in the Appendix. Using these formulae requires computing the sum of squares (over markers) of the average genotype, that is, $\sum_{j=1}^{p} \bar{x}_{j}^{2}$, where $\bar{x}_{j}^{2} = n^{-1}\sum_{i=1}^{n} x_{ij}$. In the wheat dataset, this quantity is approximately 504; using this and $df_{\beta_R} = 3$, $V_R = 0.9$ in formula [3A], we set $S_{\beta_R} = \frac{4.5}{504} \approx .009$. Finally, using $\sigma_{\varepsilon}^{2} V_L = 1$ in formula [4A] of the Appendix, we find $\hat{\lambda} \approx 32$. Choosing $\lambda^{2} \sim G\left(\lambda^{2} \middle| r = 2 \times 10^{-5}, \delta = 0.52\right)$ gives a prior density for λ that has high density and is relatively flat around $\hat{\lambda} = 32$ (Figure 3). When $\lambda^{2} \sim G(r, \delta)$, $p(\lambda|r,\delta) = G\left(\lambda^{2}\middle|r,\delta\right)\left\|\frac{\partial \lambda^{2}}{\partial \lambda}\right\| = G\left(\lambda^{2}\middle|r,\delta\right)2\lambda$; this is the density displayed in Figure 3.

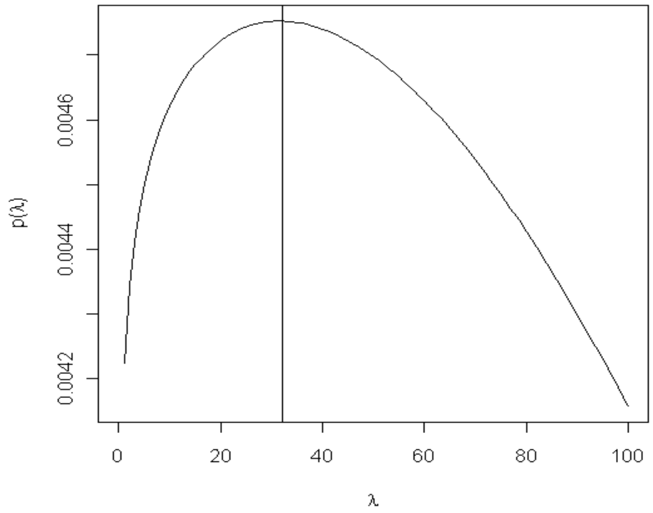

Figure 3. Prior density of λ when λ^2 is assigned a gamma prior with rate equal to 2×10^{-5} and shape equal to 0.52.

Fitting the model

Using the aforementioned values of hyper-parameters, BL and BRR were fitted using the following R-code:

```
prior = list( varE = list(S=4.5,  df=3),
              varBR = list(S=.0009, df=3),
              lambda=list(type='random',value=30,
                          shape=.52,rate=2e-5))
nIter<-60000
burnIn<-10000

fmR<-BLR(y=y,XR=X,nIter=nIter,
         burnIn=burnIn,thin=10,saveAt='R_',prior=prior)
dput(fmR,file='fmR.out')

fmL<-BLR(y=y,XL=X,nIter=nIter,
         burnIn=burnIn,thin=10, saveAt='L_',prior=prior)
dput(fmL,file='fmL.out')
```

In the above code, **y** is the response vector, **X** is the matrix of genotypes, and nIter and burnIn define the number of iterations and burn-in period, respectively, used in the Gibbs sampler. The prior is provided as a list; type 'help(BLR)' in the R console for more details. The BLR function returns a list with posterior means, posterior standard deviations, and Deviance Information Criterion (DIC, Spiegelhalter et al., 2002). Function dput() saves this list to the hard drive. The fitted model can then be retrieved using function dget(). In addition to returning posterior means and posterior standard deviations, as the Gibbs sampler runs, samples of the intercept, of the fixed effects, of the variance parameters and of λ are saved to the disc using the thinning specified by the user (which is set to 10 by default).

Results

We ran the processes in an Intel Xeon 5530 2.4 GHz Quad Core Processor (R was executed in a single thread) with 6GB of RAM memory. With this dataset (599 subjects, 1279 markers) the process took about 1% of RAM memory. BRR took about 5.5 seconds for every 1000 iterations of the sampler; BL takes about twice as much time.

Figure 4 shows the estimated posterior density of the residual variance for each of the models; the prior density (up to a constant) is included as well. The posterior distributions moved away from the prior and were sharp. The estimated posterior means (standard deviation) of σ_{ε}^{2} were 1.00 (.0728) and .929 (.0789) for BRR and BL, respectively. These values are close to the true value of the parameter (one) and suggest that BL over-fitted the data slightly. The posterior SD was 8% larger in BL, and mixing of the residual variance was better in BRR. BL gives a slightly 'less informative' posterior distribution and worse mixing for two reasons. First, due to use of marker-specific variances, the number of unknowns in BL is much larger than in BRR. Second, the BL has an extra level in which the regularization parameter indexing the prior distribution of the marker-specific variances, $\sigma_{\varepsilon}^{2}\tau_{j}^{2}$, is inferred from the data. In BRR, the counterparts of $\sigma_{\beta_R}^{2}$ are df_{β_R} and S_{β_R}, which are specified by the user.

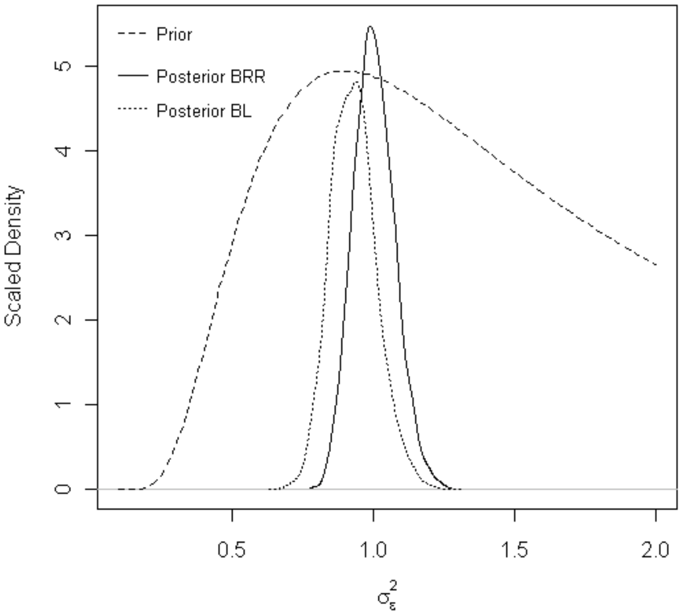

Figure 4. Prior and estimated posterior density of the residual variance, by model (BRR=Bayesian Ridge Regression; BL=Bayesian LASSO).

The posterior mean of λ in BL was 20.0, and a 95th highest posterior density confidence region was bounded by [15.24, 25.2]. These results also indicate that the posterior distribution of λ moved away from the prior (Figure 3), indicating that Bayesian learning takes place.

Measures of goodness of fit and model complexity (pD=estimated effective number of parameters, Spiegelhalter et al., 2002, and DIC) are included in the fitted object. This can be assessed with the following code:

```
fmBR$fit ## Bayesian Ridge Regression
fmBL$fit ## Bayesian LASSO
```

Table 1 provides estimates of the log-likelihood evaluated at the posterior mean of model unknowns, $l(\bar{\boldsymbol{\theta}})$, the posterior mean of the log-likelihood, $\bar{l}(.)$, the estimated effective number of parameters, pD, and DIC for BRR and BL. The Bayesian LASSO fitted the data better, had smaller $\bar{l}(.)$, and had a higher estimated number of effective parameters; DIC, which balances goodness of fit and complexity, favored the BL.

Table 1. Measures of goodness of fit, model complexity and Deviance Information Criterion (DIC) by model.

	Bayesian Ridge Regression	Bayesian LASSO
$l(\bar{\boldsymbol{\theta}})$	-777.8	-748.4
$\bar{l}(.)$	-849.2	-825.7
pD	142.7	154.6
DIC	1841.0	1806.0

$l(\bar{\boldsymbol{\theta}})$= log-likelihood evaluated at the estimated posterior mean of model unknowns; $\bar{l}(.)$=estimated posterior mean of the log-likelihood; pD=estimated effective number of parameters.

The mean-squared error of estimates of marker effects, $MSE\left(\boldsymbol{\beta}_{.},\hat{\boldsymbol{\beta}}_{.}\right)=p_{.}^{-1}\sum_{j=1}^{p}\left(\beta_{.j}-\hat{\beta}_{.j}\right)^{2}$, where (.) can be either L or R, were 0.0046 and 0.0039 for BRR and BL, respectively. The mean-squared error of estimates of genetic values, $MSE\left(\mathbf{X}\boldsymbol{\beta}_{.},\mathbf{X}\hat{\boldsymbol{\beta}}_{.}\right)=n^{-1}\sum_{j=1}^{n}\left(\mathbf{x}_{.i}\boldsymbol{\beta}_{.}-\mathbf{x}_{.i}\hat{\boldsymbol{\beta}}_{.}\right)^{2}$, were 0.3123 and 0.2759 for BRR and BL, respectively. Therefore, BL outperformed BRR in this simulation. The difference between methods is more pronounced at the level of marker effects (14% difference in MSE between methods for genetic values). Figure 5 shows scatter plots of the estimates of marker effects (left) and of genetic values (right) obtained with BL (horizontal axis) and BRR (vertical axis). The BRR shrinks estimates of markers with sizable effects to a larger extent than BL (see left panel in Figure 5). On the other hand, the two models yielded similar predictions of genetic values (see right panel in Figure 5). This occurs because with *p*>>*n* one can arrive at similar predictions of genetic values either with a model where genetic values are highly dependent on a few markers with sizable effect (something that occurs in LASSO and, to a lesser extent, in BL) or with a model where a large number of markers make a small contribution to genetic values (something that occurs in RR-BLUP and BRR). Which model yields better prediction of genetic values will depend on the underlying architecture of the trait and on the available marker data.

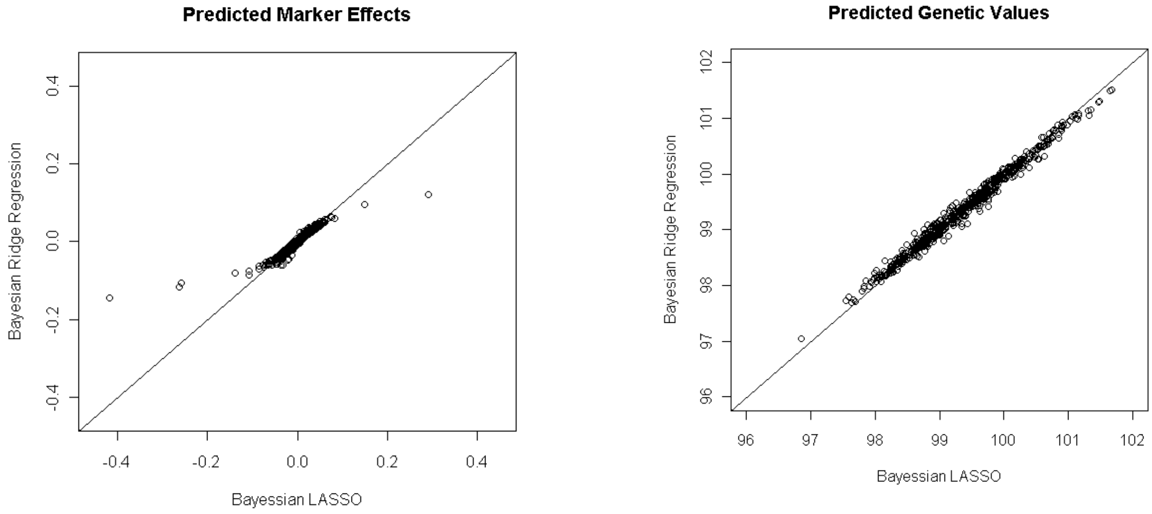

Figure 5. Estimates of marker effects (left panel) and of genetic values (right panel) obtained with the Bayesian Ridge Regression (BRR, vertical axis) versus those obtained with the Bayesian LASSO (BL, horizontal axis).

Example 2: Assessing predictive ability by cross-validation

Predicting genetic values of lines with yet-to-be observed phenotypes is a central problem in plant breeding programs. Such predictions can be used, for example, to decide which of the newly developed lines will be included in field trials or which of these lines will be parents for the next breeding cycle. Either of the models described above, BL or BRR, can be used to obtain these predictions. The rate of genetic progress will depend on how accurate such predictions are, i.e., on the ability of the model to predict future outcomes. Cross-validation (CV) methods can be used to assess predictive ability. CV can also be used for tuning-up values of certain parameters. In this section, we illustrate how the regularization parameter of the BL, λ, can be chosen using CV methods, and compare the performance of this approach with that obtained with the fully-Bayesian approach that consists of assigning a prior to λ.

Model

The linear model [NOTE: the term “data equation is often used in social scienes but it is bad. An equation typically involves a system which needs to be solved for unknowns as opposed to a formulation of how one thinks observations are generated in nature: a model] is $\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_L\boldsymbol{\beta}_L + \mathbf{u} + \boldsymbol{\varepsilon}$. We chose values of hyper-parameters using formulae presented in the Appendix. As in Example 1, it was assumed a priori that 50% of the phenotypic variance (which equals to one because phenotypes were standardized to a unit variance) could be attributed to genetic values. With this, and using $df_\varepsilon = 3$ in formula [1A] in the Appendix, we obtained $S_\varepsilon = 2.5$. Further, we assumed a priori that one half of the variance of genetic values can be accounted for by the regression on markers, $\mathbf{X}_L\boldsymbol{\beta}_L$, and that the regression on the pedigree, $\mathbf{u}$, accounted for the other half. With this, and using $df_u = 3$ and $\bar{a} = 1.98$ in [2A], we obtained $S_u = 0.63$. Finally, using $\sum_{j=1}^{p}\left(n^{-1}\sum_{i=1}^{n} x_{Lij}\right)^2 = 504$ in formula [4A] in the Appendix, we obtained $\hat{\lambda} = \sqrt{2\sigma_\varepsilon^2 V_L^{-1}\sum_{j}^{p_R}\bar{x}_{Lj}^2} = \sqrt{2\frac{1}{2}\frac{4}{1}504} \approx 45$. Choosing $\lambda^2 \sim G\left(\lambda^2 \middle| r = 1\times10^{-5}, \delta = 0.52\right)$ gives a prior for λ that has a maximum and is relatively flat in the neighborhood of 45.

The following R-code illustrates how this CV was carried out for the first trait. The vector **sets** assign lines to folds of the CV. The code involves two loops: the outer loop runs over folds of the CV; the inner loop fits models over a grid of values of λ. For every fold in the outer loop, the phenotypes of approximately 60 lines are declared as missing values; the fitted model yields a prediction of the performance of these lines.

```
library(BLR)
data(wheat)
y<-Y[,1]
sets<-(rep(1:10,60)[order(runif(600))])[-1] # Assignment of lines to folds
lambda<-seq(from=3,to=60,by=3)
postMeanLambda<-numeric()
varE<-matrix(nrow=folds,ncol=(length(lambda)+1),NA)
colnames(varE)<-c(lambda,'random')
yHatCV<-matrix(nrow=599,ncol=(length(lambda)+1),NA)
colnames(yHatCV)<-colnames(varE)
prior<-list( varE=list(df=3,S=2.5),    varU=list(df=3,S=0.63),
             lambda =list(type='fixed',value=30,rate=1e-5,shape=0.52))

for(fold in 1:10){ # LOOP FOR FOLDS
    for(j in 1:length(lambda)){ # LOOP FOR VALUES OF LAMBDA
        prior$lambda$type<-'fixed'
        prior$lambda$value<-lambda[j]
        yNa<-y
        whichNa<-which(sets==fold)
        yNa[whichNa]<-NA
        prefix<-paste('lambda_',lambda[j],'_fold_',fold,sep='')
        fm<-BLR(y=yNa,XL=X,GF=list(ID=(1:nrow(A)),A=A),prior=prior,
                nIter=30000,burnIn=5000,thin=thin,saveAt=prefix)
        yHatCV[whichNa,j]<-fm$yHat[whichNa]
        varE[fold,j]<-fm$varE
    }
    prior$lambda$type<-'random'
    prior$lambda$value<-30
    prefix<-paste('randomLambda_fold_',fold,sep='')
    fm<-BLR(y=yNa,XL=X,GF=list(ID=(1:nrow(A)),A=A), prior=prior,
            nIter=60000,burnIn=10000,thin=thin,saveAt=prefix)
    yHatCV[whichNa,(length(lambda)+1)]<-fm$yHat[whichNa]
    postMeanLambda[fold]<-fm$lambda
    varE[fold,(length(lambda)+1)]<-fm$varE
}
```

Results

Figure 6 gives the estimated mean-squared error of predictive residuals (PMSE, vertical axis) versus values of the regularization parameter (λ), by environment. The vertical and horizontal dashed lines give the average (across 10 folds of the CV) estimated posterior mean of λ and the estimated PMSE obtained when a prior was assigned to λ (i.e., the fully-Bayesian LASSO). In all environments except E3, the curve relating PMSE and λ was U-shaped, with an optimum λ (minimum PMSE) near 20. However, the absolute value of the slope of the curve was higher for low values of λ, indicating that over-fitting, something that occurs with small values of λ, is more problematic. This may occur because the conditional expectation function in these models included two random components: the regression on markers and the regression on the pedigree. As values of λ increase (i.e., placing a higher penalty on the regression on markers), the contribution of the regression on the pedigree to the conditional expectation increases as well, preventing lack of fit. Environment 3 constitutes an extreme example of this; here the curve relating PMSE and λ looks L-shaped.

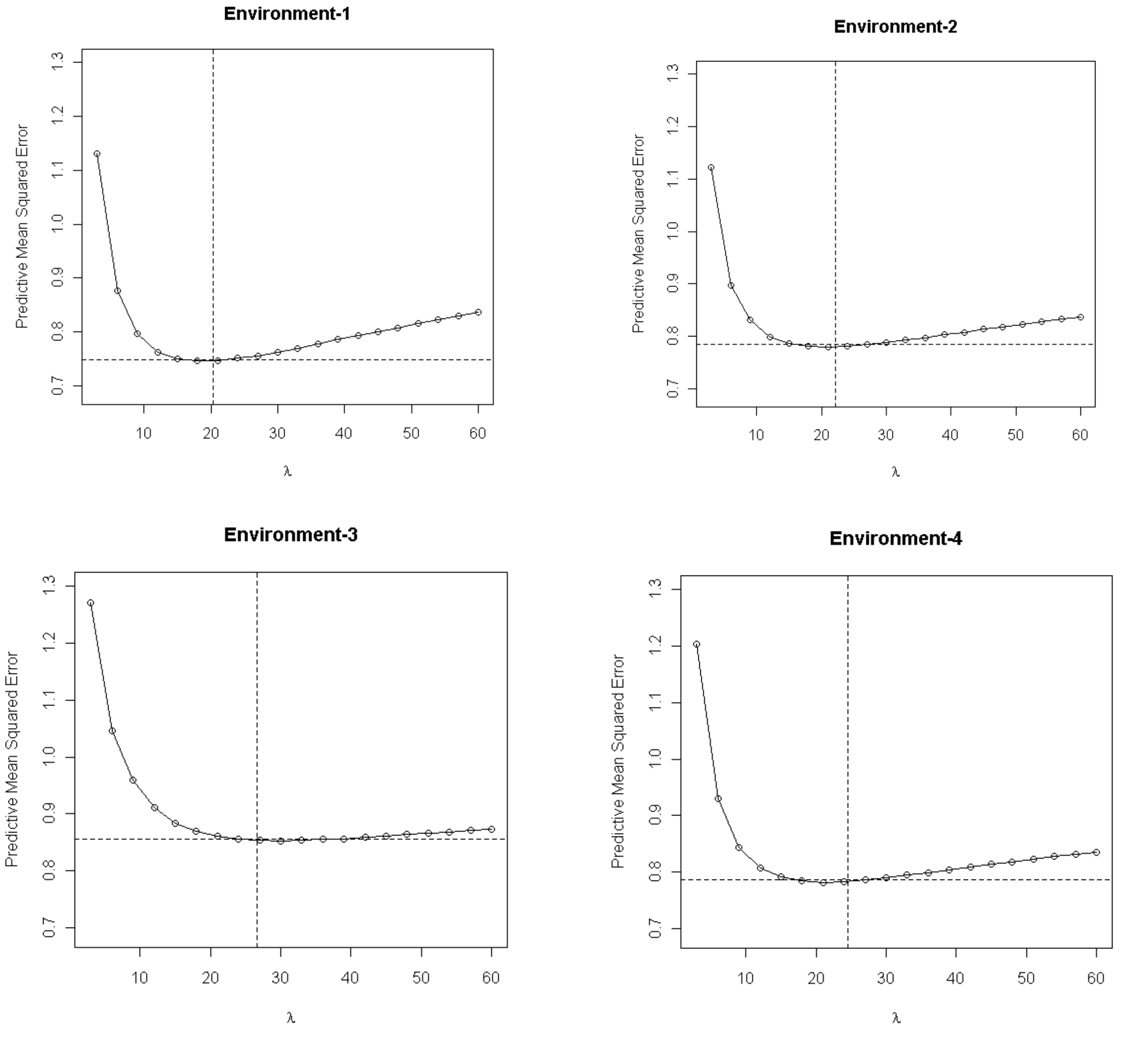

Figure 6. Predictive mean squared error (PMS, vertical axis) versus values of the regularization parameter of the Bayesian LASSO (horizontal axis), by environment. The vertical and horizontal dashed lines give the average (across 10 folds of the cross-validation) estimated posterior mean of λ and the estimated PMSE obtained when a prior was assigned to λ.

The posterior modes of λ were always considerably smaller than the prior mode (45), indicating that Bayesian learning took place. In all environments, the fully-Bayesian treatment yielded posterior means of λ that were in the neighborhood of the optimal values found when models were run over a grid of values of this unknown. Also, predictive ability of models with random λ was as good as the best obtained when CV was used for choosing λ. These results suggest that,at least for these traits and this population, the fully-Bayesian treatment, which consists of inferring λ from the data, yields good results.

Concluding remarks

The BLR package allows fitting high-dimensional linear regression models including dense molecular markers, pedigree information and covariates other than markers. The interface allows the user to choose models (e.g, BL versus BRR) and prior hyper-parameters easily. The algorithms implemented are relatively efficient and models with a modest number of markers (e.g., ~1000) can be fitted in a standard PC easily. The routines implemented in the package have also been used successfully in problems with larger numbers of molecular markers. For example, Weigel et al. (2009) used an earlier version of the package to fit models using data from the 50k Bovine Illumina Bead Chip, and our experience indicates that the software can be used with an even larger number of markers. Computational time is expected to increase linearly with the number of markers; the user also needs to be aware that marker information is loaded in the memory. Therefore, as the number of marker increases, so do the memory requirements.

In models for genomic selection, with $p>>n$, marker effects cannot be estimated uniquely from the likelihood. A unique solution can be obtained by using penalized estimation methods or, in a Bayesian framework, by assigning informative priors to marker effects. Because of lack of identification at the level of the likelihood, the choice of prior is expected to play a role. As illustrated in Example 1, different priors yield different estimates of marker effects. Although models per se cannot solve the intrinsic identification problem, one can use model comparison criteria, such as DIC, or the principle of parsimony, or CV methods, to choose among prior distributions of marker effects. Although the choice of prior affects estimates of marker effects, the influence of the prior on estimates of genetic values may be less important (see Example 1, or the simulation study presented in de los Campos et al., 2009). This occurs because one can arrive at similar predictions either with a model where genetic values are highly dependent on some markers or with a model where all markers make a small contribution to genetic values.

Finally, estimates of marker effects obtained with BL or BRR could be used to assess the relative contribution of each region to genetic variability. However, one needs to be aware that the estimated marker effect reflects not only linkage between markers and genes affecting the trait, but also the density of markers in the region. If a region containing a QTL has a high density of markers, the effect of the QTL is expected to be ‘distributed’ across linked markers; conversely, if in the same region markers are sparse, estimated marker effects are expected to be larger in absolute value.

References

Crossa, J., de los Campos, G, Pérez, P., Gianola, D., Atlin, G., Burgueño, J., Araus, J.L., Makumbi, D., Yan, J., Arief, V., Banziger, M., Braun, H-J. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics (submitted).

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, et al. 2009. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigrees. Genetics, 182: 375-385.

de los Campos, G., and P. Pérez. 2010. BLR: Bayesian Linear Regression. R package version 1.1. http://www.r-project.org/.

Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 52: 399-433.

Bernardo, R., and J. Yu. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Science 47: 1082-1090.

Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R.L. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183: 347-363.

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124: 323-330.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. of Dairy Science 92: 433-443.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-447.

Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55-67.

McLaren, C.G. 2005. The International Rice Information System. A platform for meta-analysis of rice crop data. Plant Physiology, 139: 637-642.

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

Park, T., and G. Casella. 2008. The Bayesian LASSO. J. Am. Stat. Assoc. 103:681-686.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 64: 583-639.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. J. Royal. Statist. Soc. B. 58: 267-288.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, et al. 2008. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. of Dairy Science 92:16-24.

Weigel, K., G. de los Campos, A.I. Vazquez, O. González-Recio, H. Naya, X.L. Wu, N. Long, G.J.M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92: 5248-5257.

Wright, S. 1921. Systems of mating. I. The biometric relations between parents and offspring. Genetics 6: 111-123.

Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045-1055.

APPENDIX

The prior distribution of the BLR model is indexed by several hyper-parameters. We provide guidelines for choosing their values based on beliefs as to what proportion of the variance of phenotypes can be attributed to each of the random components of the model, that is, u_i, $\sum_{j=1}^{p_R} x_{Rij}\beta_{Rj}$, $\sum_{j=1}^{p_R} x_{Lij}\beta_{Lj}$ and ε_i. Other authors (e.g., Meuwissen et al., 2001) have discussed how to choose hyper-parameters in the context of models for genomic selection. However, the derivation used by those authors assumes that genotypes are random and marker effects are fixed quantities (e.g., Gianola et al., 2009), while in fact the opposite is true in the Bayesian models that have been proposed. Here, we derive formulae that are consistent with the standard treatment of marker and marker effects in Bayesian models for GS, where marker genotypes are observed quantities and marker effects are random unknowns. Unlike the formulae in Meuwissen et al. (2001), the ones presented here do not require making any assumption about the extent of linkage disequilibrium between markers.

Residual variance. The prior distribution of the residual variance is indexed by two parameters, $\{S_\varepsilon, df_\varepsilon\}$. One can choose these parameters so that the prior mode of σ_ε^2, $\frac{S_\varepsilon}{df_\varepsilon + 2}$ matches our prior beliefs about the variance of model residuals. In practice, df_ε can be chosen to be a small value, usually greater than two, to guarantee a finite prior expectation, e.g., $df_\varepsilon = 3$, then the prior scale can be:

$$S_\varepsilon = V_\varepsilon (df_\varepsilon + 2), \qquad [1A]$$

where V_ε is chosen to reflect our expectation of the variance of model residuals. Formulas similar to that presented in this Appendix can be derived using formulas for the prior expectation.

Variance of the infinitesimal effect. From the prior distribution, the variance of u_i is $a_{ii}\sigma_u^2$, where a_{ii} is the i$^{\text{th}}$ diagonal element of **A**, which in the absence of inbreeding equals one. If we let $\bar{a}$ be the average diagonal value of **A** and V_u be our prior expectation about $\bar{a}\sigma_u^2$, then we can choose $df_u = 3$ and set the prior scale to be:

$$S_u = \frac{V_u(df_\varepsilon + 2)}{\bar{a}}. \qquad [2A]$$

Prior variance of marker effects. The contributions of $\mathbf{X}_R$ and $\mathbf{X}_L$ to phenotypes is $\sum_{j=1}^{p.} x_{.ij}\beta_{.j}$. Here, (.) stands for R or L, depending on whether BRR or BL is being used. In regressions for GS, marker genotypes are fixed and marker effects are random. Furthermore, at the level of the marginal distribution, marker effects are IID; therefore, $Var\left(\sum_{j=1}^{p.} x_{.ij}\beta_{.j}\right) = V_R = \sum_{j=1}^{p.} x_{.ij}^2 Var(\beta_{.j})$. For the 'average' genotype, this formula becomes $V_R = \left\{\sum_{j=1}^{p.} \bar{x}_{.j}^2\right\} Var(\beta_{.j})$, where $\bar{x}_{.j}$ denotes the average value of the jth column of $\mathbf{X}_R$ or $\mathbf{X}_L$.

In BRR, the prior distribution of marker effects is Gaussian and $Var(\beta_{Rj}) = \sigma_\beta^2$. As before, df_{β_R} can be chosen to have a relatively small value, e.g., $df_{\beta_R} = 3$, and then the prior scale can be set to be

$$S_{\beta_R} = \frac{V_R \times \left(df_{\beta_R} + 2\right)}{\sum_j^{p_R} \bar{x}_{Rj}^2} \qquad [3A]$$

where V_R is set to reflect our expectation of the variance of phenotypes that can be attributed to the regression on $\mathbf{X}_R$.

In the BL, the marginal prior density of marker effects is Double-Exponential, and the prior variance of marker effects is $Var\left(\beta_{Lj}\right) = 2\sigma_\varepsilon^2 \lambda^{-2}$. Plugging this in $Var\left(\sum_{j=1}^{p} x_{.ij}\beta_{.j}\right) = V_L = \left\{\sum_{j=1}^{p} \bar{x}_{.j}^2\right\} Var\left(\beta_{.j}\right)$, we obtain $V_L = \left\{\sum_{j=1}^{p} \bar{x}_{.j}^2\right\} 2\sigma_\varepsilon^2 \lambda^{-2}$. Solving for λ, we get:

$$\hat{\lambda} = \sqrt{2\sigma_\varepsilon^2 V_L^{-1} \sum_j^{p_R} \bar{x}_{Rj}^2} \qquad [4A]$$

where $\sigma_\varepsilon^2 V_L^{-1}$ is a noise-to-signal variance ratio. With [4A] we can choose a target value for the regularization parameter. Then we can choose hyper-parameters so that the prior has a mode and is relatively flat, in the neighborhood of $\hat{\lambda}$.

The above formulas constitute guidelines for choosing values of hyper-parameters. In practice, if Bayesian learning takes place, one would expect that the posterior distribution moves away from the prior. Furthermore, with small n it is always useful to check the sensitivity of inferences with respect to the choice of hyper-parameters.