G3: Genes|Genomes|Genetics Early Online, published on August 25, 2014 as doi:10.1534/g3.114.013094

Bayesian genomic-enabled prediction as an inverse problem

J. Cuevas, S. Pérez-Elizalde, V. Soberanis, P. Pérez-Rodríguez, D. Gianola, and J. Crossa

J. Cuevas, S. Pérez-Elizalde, and P. Pérez-Rodríguez

Colegio de Posgraduados, 56230, Montecillo, Texcoco, Edo. de México.

V. Soberanis

Universidad de Quintana Roo, Chetumal, Quintana Roo, México.

D. Gianola

Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA

J. Crossa

Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, México, D.F., México. Email(j.crossa@cgiar.org)

ABSTRACT

Genomic-enabled prediction in plant and animal breeding has become an active area of research. Many prediction models address the collinearity that arises when the number (p) of molecular markers (SNPs) is larger than the sample size (n). Here we propose four Bayesian approaches to the problem based on commonly used data reduction methods. Specifically, we use a Gaussian linear model for an orthogonal transformation of both the observed data and the matrix of molecular markers. Since shrinkage of estimates is affected by the prior variance of transformed effects, we propose four structures of the prior variance as a way of potentially increasing the prediction accuracy of the models fitted. To evaluate our methods, maize and wheat data previously used with standard Bayesian regression models were employed for measuring prediction accuracy using the proposed models. Results indicate that, for the maize and wheat data sets, our Bayesian models yielded, on average, a prediction accuracy that is 3% higher than that of standard Bayesian regression models, with less computational effort.

Keywords: Bayesian regression, shrinkage, prior distribution, inverse regression.

INTRODUCTION

Predicting complex traits possibly affected by large numbers of genes (each one having a small effect) that are greatly affected by the environment is a difficult task. Information from dense molecular markers attempts to exploit linkage disequilibrium between at least one marker and at least one putatively causal locus, so as to predict the genetic values of individuals based on their phenotypic data.

There is a vast literature describing methods that use different functions of markers for predicting genetic values (e.g., de los Campos et al., 2012; Gianola, 2013), starting with the seminal work of Meuwissen et al. (2001), which proposes genome-based prediction functions that regress phenotypes on all available markers using a Gaussian linear model with three different prior distributions on marker effects. Several regularized regression models, such as ridge regression (Hoerl and Kennard, 1970), the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), and its Bayesian counterpart (Park and Casella, 2008; de los Campos et al., 2010), have been described and used for genomic-based prediction in animals and plants (de los Campos et al., 2009, 2010; Crossa et al., 2010, 2011; Gónzalez-Camacho et al., 2012; Heslot et al., 2012; Pérez-Rodríguez et al., 2010, 2012).

The basic quantitative genetic model describes the i^{th} response or phenotype (y_i) as the sum of an unknown genetic value (g_i) plus a residual ε_i expressed as a deviation from some general mean (μ); thus the basic model is $y_i = g_i + \varepsilon_i$ $(i = 1, \ldots, n)$. The unknown genetic value can be represented as a function of genotypes with a large number of genes that may involve all gene × gene interactions, if there are any. Since the genes affecting a trait are unknown, this complex function must be approximated by, for example, a regression of phenotype on marker genotypes. Large numbers of bi-allellic markers on $\{x_{i1}, \ldots, x_{ip}\}$ (x_{ij} is the number of copies of one of the two alleles observed in the i^{th} individual at the j^{th} marker) may be used in a regression function for predicting the genetic value of the i^{th} individual. The regression can be formulated as $u(\boldsymbol{x}) = u(x_{i1}, \ldots, x_{ip})$ such that the basic genomic model becomes $y_i = u_i + \varepsilon_i$, where ε_i is a model error that may include errors due to unspecified environmental effects, imperfect linkage disequilibrium between markers and the actual loci affecting the trait, and unaccounted for gene × gene

and gene × environment interactions. In several applications, $u(\boldsymbol{x})$ is a parametric linear regression of the form $u(x_{i1}, \ldots, x_{ip}) = \sum_{j=1}^{p} x_{ij}\,\beta_j$, where β_j is the substitution effect of the allele coded as 'one' at the j^{th} marker.

The linear regression function on markers becomes $y_i = \sum_{j=1}^{p} x_{ij}\,\beta_j + \varepsilon_i$, or, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{X}$ is the $n \times p$ incidence matrix, $\boldsymbol{\beta}$ is the vector of unknown marker effects, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors, typically assumed to follow the normal distribution $N(\mathbf{0}, \boldsymbol{I}_n\sigma_\varepsilon^2)$, where σ_ε^2 is the random error variance component. If the vector of marker effects is assumed random with normal distribution $N(\mathbf{0}, \boldsymbol{I}_p\sigma_\beta^2)$, where σ_β^2 is the variance of marker effects, then the genetic value of the individuals is $\boldsymbol{u} = \boldsymbol{X\beta}$ with a variance-covariance matrix proportional to the genomic relationship matrix $\boldsymbol{G} = \boldsymbol{XX}'/\sum_{j=1}^{p} 2q_j(1-q_j)$ (where q_j is the frequency of allele "1" often estimated from the data at hand); this random effect leads to the Genomic Best Linear Unbiased Predictor (GBLUP) (VanRaden, 2007, 2008), which is extensively used in genomic-assisted prediction. Note that the Bayesian Ridge Regression (BRR) is equivalent to the GBLUP when the distribution of SNP effects is regarded as Gaussian with mean **0** and variance $\boldsymbol{I}_p\sigma_\beta^2$, that is $N(\mathbf{0}, \boldsymbol{I}_p\sigma_\beta^2)$ (Pérez-Rodríguez et al., 2012)

The regression function $u(x_{i1}, \ldots, x_{ip})$ can also be represented by semi-parametric approaches, such as Reproducing Kernel Hilbert Space (RKHS) regressions or different types of neural networks (NN) (Gianola et al., 2006, 2011; Gianola and van Kaam, 2008; de los Campos et al., 2010; González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012). De los Campos et al. (2012) reviewed penalized linear regression models and Bayesian shrinkage estimation methods.

The basic linear model (1) is generally over-parameterized and therefore ill-conditioned because there are many more predictors (markers) than individuals with observations ($p>>n$), as well as strong collinearity among predictors due to linkage disequilibrium; hence $\boldsymbol{X}$ is not a full column rank. These challenges can be overcome using what is called the inverse problem approach (Aster et al., 2005; Tarantola, 2005). The

discrete inverse problem approach for matrices of incomplete rank or that are ill-conditioned due to possible collinearity is discussed at length by Aster et al. (2005), whereas Tarantola (2005) addresses the continuous version of the same problem. Cavalier (2008) summarized the inverse problem as a non-parametric regression, $\boldsymbol{d} = \boldsymbol{T}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}$, of a transformed data vector $\boldsymbol{d}$ on the linear operator $\boldsymbol{T}_{n\times n}$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_n \subseteq \mathbb{R}^n$ as a vector of regression parameters and $\tilde{\boldsymbol{\varepsilon}}$ as a vector of Gaussian errors. Within a Bayesian framework, Knapick et al. (2012) proposed a solution by using the singular value decomposition of $\boldsymbol{T}$ such that the decay of singular values is mimicked in the prior distribution of $\boldsymbol{\theta}$. Thus estimating the unknown transformed marker effects of $\boldsymbol{\theta}$ from this perspective seems of interest but has not been attempted in genome-enabled prediction so far. The inferential solution to the inverse problem is based on the posterior distribution of $\boldsymbol{\theta}$, with the variability of $\tilde{\boldsymbol{\varepsilon}}$ and prior beliefs about $\boldsymbol{\theta}$ conveyed through probabilistic models.

In the prediction context, de los Campos et al. (2010) used the fact that the matrix $\boldsymbol{G}$ (or $\boldsymbol{X}$) may be eigen-decomposed as $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'$, with $\boldsymbol{D}$ containing the eigenvalues of $\boldsymbol{G}$, and $\boldsymbol{U}$ being its corresponding eigenvectors. This transformation allows reducing the original highly dimensional data to fewer dimensions by extracting most of the signals in the "first" components of the decomposition and relegating the noise to the "last" components. As de los Campos et al. (2010) demonstrated, the regression of phenotypes on all markers is equivalent to the regression of phenotypes on a set of rotated covariates. Implementing regression methods with $p >> n$ generally requires either shrinkage estimation or reducing the dimensionality of the data, which is the approach considered here. Recently, Gianola (2013) studied the influence of the prior on the posterior inference of marker effects in the over-parameterized models used in genome-enabled prediction. He concluded that a main driving force in the various Bayesian models (Bayesian alphabet) is the prior and not the data; thus different priors will produce different results mainly because their shrinkage effect varies.

In this paper, we propose using the parametric model (1) within the framework of inverse problems and a Bayesian approach to predict genetic values of individuals when *p>>n*. This proposal is similar to that introduced by de los Campos et al. (2010) for dimension reduction. However, ours has several different features, including an orthogonal

transformation of the observed data (***y***), as well as differential shrinkage as a result of the prior variance assigned to the regression parameters.

MATERIALS AND METHODS

Statistical methods

In model (1), when $\boldsymbol{X}$ is not of full rank, the least squares solution for the unknown $\boldsymbol{\beta}$ is neither unique nor stable. These ill-conditioned problems may be analyzed within the framework of an inverse problem theory by using the singular value decomposition of ***X,*** of order $n \times p$ (Aster et al., 2005), as $\boldsymbol{X} = \boldsymbol{USV'}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices of orders $n \times n$ and $p \times p$, respectively, and $\boldsymbol{S}$ is a rectangular $n \times p$ matrix containing n non-negative singular values ordered from largest to smallest, $s_1 \geq s_2 \geq \cdots \geq s_n$. One can write $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2]$, where $\boldsymbol{S}_1$ is the $n \times n$ matrix with singular values along its diagonal, and $\boldsymbol{S}_2$ is a matrix of order $n \times (p\text{-}n)$ and all its entries are null. Consider the linear transformation on both sides of (1):

$$\boldsymbol{U'y} = \boldsymbol{U'USV'\beta} + \boldsymbol{U'\varepsilon}$$

and let $\boldsymbol{d} = \boldsymbol{U'y}$, $\boldsymbol{b} = \boldsymbol{V'\beta}$ and $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{U'\varepsilon}$. Since $\boldsymbol{U'U} = \boldsymbol{I}_n$, (1) becomes:

$$\boldsymbol{d} = \boldsymbol{Sb} + \tilde{\boldsymbol{\varepsilon}} \qquad (2)$$

In (2), $\boldsymbol{d}$ and $\tilde{\boldsymbol{\varepsilon}}$ are vectors each of order $n \times 1$ and $\boldsymbol{b}$ is a vector of order $p \times 1$. The column vector $\boldsymbol{b}$ is partitioned as $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$, where $\boldsymbol{b}_1$ is an $n \times 1$, and $\boldsymbol{b}_2$ has order $(p\text{-}n) \times 1$. Therefore, equation (2) becomes:

$$\boldsymbol{d} = \boldsymbol{S}_1\boldsymbol{b}_1 + \boldsymbol{S}_2\boldsymbol{b}_2 + \tilde{\boldsymbol{\varepsilon}} = \boldsymbol{S}_1\boldsymbol{b}_1 + \tilde{\boldsymbol{\varepsilon}}$$

because $\boldsymbol{S}_2\boldsymbol{b}_2$ is zero for any value of $\boldsymbol{b}_2$. Here only the first n entries of $\boldsymbol{b}$ can be inferred.

Since $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{U'\varepsilon}$, it follows that $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{U'\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{U'U}\sigma_\varepsilon^2) = N(\boldsymbol{0}, \boldsymbol{I}_n\sigma_\varepsilon^2)$. The distribution of the transformed data $\boldsymbol{d}$, given $\boldsymbol{b}$ and σ_ε^2, is

$$\boldsymbol{f}(\boldsymbol{d}|\boldsymbol{b},\sigma_\varepsilon^2) = \prod_{i=1}^{n} N(d_i\,|s_i b_i,\sigma_\varepsilon^2)$$

where d_i is the *i-th* element of $\boldsymbol{d}$ and has the form

$$d_i = s_i b_i + \tilde{\varepsilon}_i\ ,\ (i = 1,\dots,n) \tag{3}$$

since $\boldsymbol{S}_1 = \mathrm{diag}\{s_1,\dots,s_n\}$. To recover the original parameter $\boldsymbol{\beta}$, the $p \times p$ matrix $\boldsymbol{V}$ is partitioned into $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2]$, where $\boldsymbol{V}_1$ contains the first *n* columns of $\boldsymbol{V}$, and $\boldsymbol{V}_2$ contains the remaining $p - n$, so that $\boldsymbol{\beta} = \boldsymbol{V}_1\boldsymbol{b}_1 + \boldsymbol{V}_2\boldsymbol{b}_2$. The *p* original parameters are recovered as $\boldsymbol{\beta} = \boldsymbol{V}_1\boldsymbol{b}_1$, whereas $\boldsymbol{b}_2$ does not contribute to fitting the data, that is:

$$\boldsymbol{\beta} = \sum_{i=1}^{n} b_i\,\boldsymbol{V}_{i,1} \tag{4}$$

where $\boldsymbol{V}_{i,1}$ is the *i-th* column of $\boldsymbol{V}_1$. Hence

$$cov(\boldsymbol{\beta}|\boldsymbol{b_2} = \boldsymbol{0}) = \boldsymbol{V}_1 cov(\boldsymbol{b}_1)\boldsymbol{V}_1' \tag{5}$$

and, for example, independent Gaussian priors can be assigned to each of the elements of $\boldsymbol{b}_1$. Obviously, $\boldsymbol{b}_2$ is an unknown quantity, so a prior must be assigned to this vector. However, as shown above and in Gianola (2013), the likelihood function does not depend on $\boldsymbol{b}_2$, so the data do not provide information about $\boldsymbol{b}_2$. Here the marginal posterior of $\boldsymbol{b}_1$ is the likelihood times its prior, and the marginal posterior of $\boldsymbol{b}_2$ is equal to its prior. Thus, if the Dirac delta function is assigned as a prior to each element of $\boldsymbol{b}_2$, we get (4) as the quantity of interest and can safely leave $\boldsymbol{b}_2$ out of the analysis. Thus $\boldsymbol{b}_1$, $\mathbf{V}_1$ and $\boldsymbol{S}_1$ will herein after be denoted as $\boldsymbol{b}$, $\boldsymbol{V}$ and $\boldsymbol{S}$, respectively, without loss of generality.

The ordinary least squares estimator of $\boldsymbol{\beta}$ found using the generalized Moore-Penrose inverse method is

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \frac{d_i}{s_i}\,\boldsymbol{V}_i \tag{6}$$

Thus $\boldsymbol{\beta}^*$ depends on the ratios $\frac{d_i}{s_i}, i = 1,\dots,n$, between transformed data points and the decaying singular values, which could be interpreted as representing a decreasing signal-noise ratio sequence. So, as singular values decay, there is an increase in the noise

component of the least squares estimate of component $\boldsymbol{\beta}$. This signal-noise ratio is inversely proportional to the singular values, as is clearly shown by the expression:

$$\frac{d_i}{s_i} = b_i + s_i^{-1}\tilde{\varepsilon}_i \qquad (7)$$

The least squares estimator of b_i can be computed from expression (7) as $b_i^* = \frac{d_i}{s_i}$ and serves as a basis for regularizing estimation through truncation or weighting.

A Bayesian inverse regression model

Several methods have been proposed for mitigating ill-conditioned regression problems by shrinking the solution space via restricting the magnitudes of the estimates and their variance. One of the first proposals was the ridge regression estimator (Hoerl and Kennard, 1970). Since model (3) has the form $\boldsymbol{d} = \boldsymbol{Sb} + \tilde{\boldsymbol{\varepsilon}}$, with $\tilde{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_\varepsilon^2)$, the ridge regression estimator is :

$$\boldsymbol{b}_{\boldsymbol{\gamma}} = (\boldsymbol{S}'\boldsymbol{S} + \boldsymbol{\gamma}\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{d} = (\boldsymbol{S}^2 + \boldsymbol{\gamma}\boldsymbol{I})\boldsymbol{S}\boldsymbol{d} \qquad (8)$$

with $\boldsymbol{S}$ as a diagonal matrix of dimensions $n \times n$. Here $\boldsymbol{\gamma} > \mathbf{0}$ is a vector of parameters that reduce ill-conditioning. If their magnitude increases excessively, this can lead to a poor fit of the model to the data. Therefore, the choice of vector $\boldsymbol{\gamma}$ is critical.

Hoerl and Kennard (1970) showed that there is a range of $\boldsymbol{\gamma}$ values where the mean squared error (MSE) of estimates is smaller than the corresponding MSE of the ordinary least squares solution. They minimized MSE by choosing

$$\gamma_i = \frac{\sigma_\varepsilon^2}{b_i^2}, (i = 1, \ldots, n) \qquad (9)$$

This requires knowing the true values of $\boldsymbol{b}$. However, (9) can be used to justify the choice

$$\gamma \approx \frac{n\sigma_\varepsilon^2}{\hat{\boldsymbol{b}}'\hat{\boldsymbol{b}}} \approx \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_b^2} \qquad (10)$$

where $\widehat{\boldsymbol{b}}$, $\hat{\sigma}_b^2 = \widehat{\boldsymbol{b}}'\widehat{\boldsymbol{b}}/n$, and $\hat{\sigma}_\varepsilon^2$ are estimates of $\boldsymbol{b}$, σ_b^2, and σ_ε^2, respectively, thus providing a single (global) shrinkage parameter.

In the Bayesian approach, the prior distribution reflects prior beliefs about b_i, and its variance affects the extent to which the least squares estimate moves towards the prior mean. To construct a model along the lines of de los Campos et al. (2012), we adopted as the conditional prior distribution of b_i

$$p(b_i|\lambda_i) = N(b_i|0, \lambda_i),\ i = 1, \ldots, n \qquad (11)$$

where the b_i coefficients (given λ_i) are conditionally independent. Therefore, the joint posterior distribution of $\boldsymbol{b}$ and σ_ε^2 in model (3) is given by:

$$p(\boldsymbol{b}, \sigma_\varepsilon^2\ |\boldsymbol{d}, \boldsymbol{S}, \boldsymbol{\lambda}) \propto \prod_{i=1}^{n} N(d_i|s_i b_i, \sigma_\varepsilon^2\ ) N(b_i|0, \lambda_i\ )\, p(\sigma_\varepsilon^2) \qquad (12)$$

where λ_i (i=1,…,n) are variance parameters and $p(\sigma_\varepsilon^2)$ is the prior distribution of σ_ε^2. Usually the conjugate prior for σ_ε^2 is a scaled inverse chi-squared distribution with υ_e degrees of freedom and scale parameter Sc_ε, that is,

$$p(\sigma_\varepsilon^2) = \chi^{-2}(\sigma_\varepsilon^2|\upsilon_\varepsilon, Sc_\varepsilon) \qquad (13)$$

From (12), the conditional posterior distribution of $\boldsymbol{b}$ is:

$$p(\boldsymbol{b}|\boldsymbol{d}, \boldsymbol{S}, \boldsymbol{\lambda}, \sigma_\varepsilon^2\ ) = \prod_{i=1}^{n} N\left(b_i \middle| \frac{\lambda_i s_i d_i}{\lambda_i s_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2 \lambda_i}{\sigma_\varepsilon^2 + \lambda_i s_i^2}\right) \qquad (14)$$

Expression (14) indicates that the conditional posterior mean depends on the data (d_i), while the variance is a fixed but unknown quantity.

The conditional expectation of $\boldsymbol{b}$ in (14) can also be expressed as

$$E(b_i|d_i, s_i, \lambda_i, \sigma_\varepsilon^2) = \frac{\lambda_i s_i d_i}{\lambda_i s_i^2 + \sigma_\varepsilon^2} = f_i\ b_i^*$$

where $f_i = \frac{\lambda_i s_i^2}{\lambda_i s_i^2 + \sigma_\varepsilon^2}$ is known as the 'weighting factor' that weights the least squares estimate $b_i^* = \frac{d_i}{s_i}$.

Since the magnitude of variance of $b_i^*(i = 1, \dots, n)$ grows with i, Casella (1985) suggested applying a lower weighting factor for estimates with larger variance. The weighting factor assigns weights close to one to the first elements of $\boldsymbol{b}^*$, while the remaining elements receive lower weights, and can take on values close to zero, or zero itself.

Bayesian inverse parametric regression models

Four versions of the Bayesian inverse regression models are described in this section. The first two, Bayesian Inverse Ridge Regression (BIRR) and Bayes A Inverse (BAI) are versions of the standard Bayesian Ridge Regression (BRR) and Bayes A (BA) (Meuwissen et al., 2001), respectively. In these models, the decay of the prior variances is not explicitly considered, whereas the other two proposed models do incorporate a prior decay of the variances.

Bayesian Inverse Ridge Regression (BIRR)

The posterior expectation of b_i, given d_i, s_i, λ_i and σ_ε^2 in (14), can also be written as:

$$E(b_i|d_i, s_i, \lambda_i, \sigma_\varepsilon^2) = (s_i^2 + \frac{\sigma_\varepsilon^2}{\lambda_i})^{-1} s_i d_i$$

which agrees with the ridge estimator in (8). A special case (BIRR) is when all prior distributions of b_i are assigned a constant variance $\lambda_i = \sigma_b^2$. Then

$$p(b_i|\lambda_i) = N(b_i|0, \lambda_i = \sigma_b^2) \;\; i = 1, \dots, n$$

If a scaled inverse chi-squared distribution is assigned to σ_b^2, the joint posterior distribution of $(\boldsymbol{b},\ \sigma_b^2, \sigma_\varepsilon^2,)$ is given by:

$$p(\boldsymbol{b}\ , \sigma_b^2, \sigma_\varepsilon^2\ , |\boldsymbol{d}, \boldsymbol{S}, \upsilon_b, Sc_b, \upsilon_e, Sc_e) \propto$$
$$\{\prod_{i=1}^{n} N(d_i|s_i b_i, \sigma_\varepsilon^2) N(b_i|0, \sigma_b^2)\} \chi^{-2}(\sigma_b^2|\upsilon_b, \upsilon_b Sc_b) p(Sc_b) \chi^{-2}(\sigma_\varepsilon^2|\upsilon_\varepsilon, Sc_\varepsilon) \qquad (15)$$

Note that model BIRR is analogous to the standard Bayesian Ridge Regression (BRR) in the sense that they both assign to the regression parameters a prior distribution with mean zero and constant variances.

Bayes A Inverse (BAI)

Bayes A (Meuwissen et al., 2001) assigns the scaled *t* distribution as a prior to each marker effect, which can be represented as a normal-gamma mixture. We then have a hierarchical model with two levels: in the first one, a normal distribution with zero mean and variance $\lambda_i = \sigma^2_{b_i}$ is used as a conditional prior for $\boldsymbol{b}$ and, in the second level, the same scaled inverse chi-squared prior distribution is used for all $\sigma^2_{b_i}$ (Yi and Xu, 2008). Thus the joint posterior distribution of $(\boldsymbol{b}, \sigma^2_{b_i}, Sc_b, \sigma^2_\varepsilon,)$ is

$$p\left(\boldsymbol{b}, \sigma^2_{b_1} \ldots, \sigma^2_{b_n}, Sc_b, \sigma^2_\varepsilon \,\middle|\, \boldsymbol{d}, \boldsymbol{S}, \upsilon_b, Sc_b, \upsilon_\varepsilon, Sc_\varepsilon\right) \propto$$
$$\{\prod_{i=1}^{n} N(d_i|s_i b_i, \sigma^2_\varepsilon) N\left(b_i|0, \sigma^2_{b_i}\right) \chi^{-2}\left(\sigma^2_{b_i}|\upsilon_b, \upsilon_b Sc_b\right)\} p(Sc_b) \chi^{-2}(\sigma^2_\varepsilon|\upsilon_\varepsilon, Sc_\varepsilon) \qquad (16)$$

According to Yi and Xu (2008), when $\upsilon_b = 0,$ and $Sc_b = 0$, this induces an improper distribution of $\boldsymbol{b}$**.** Therefore, these hyper-parameters should be greater than zero. We assigned $\upsilon_b = 3$ and, for Sc_b, a uniform distribution with support on the interval (0,A), where A>0. It should be noted that model BAI is not equivalent to the standard Bayes A because the prior variance of $\boldsymbol{b}$ under BAI is not the prior variance of the transformation $V'\boldsymbol{\beta}$ under Bayes A.

Bayesian inverse regression with decay model 1 (BIR1)

The mean of the posterior distribution of $\boldsymbol{b}$ can be seen as the product of a filter producing values that tend to shrink the least squares estimate $\boldsymbol{b}^*$ towards 0 through the weighting factor f_i, which depends on singular values of $\boldsymbol{X}$. It is then reasonable to assume a prior that will shrink the values of $\boldsymbol{b}^*$, in accordance with the singular values decay. Suppose we assign $N(\boldsymbol{0}, \boldsymbol{\Psi})$ as prior distribution $N(\boldsymbol{0}, \boldsymbol{\Psi})$ to $\boldsymbol{\beta}$ with $\boldsymbol{\Psi} = \boldsymbol{V\Lambda V'}$

being the eigen-decomposition of $\boldsymbol{\Psi}$, where $\boldsymbol{V}$ is the right orthogonal $p \times p$ matrix of the singular value decomposition $\boldsymbol{X} = \boldsymbol{USV}'$. Then the prior distribution of $\boldsymbol{b}$ is $N(\boldsymbol{0}, \boldsymbol{\Lambda})$ because $\boldsymbol{b} = \boldsymbol{V}'\boldsymbol{\beta}$ and the vector $\boldsymbol{s^2}$ of eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and the vector $\boldsymbol{\lambda}$ of eigenvalues of $\boldsymbol{\Psi}$ share the same $\boldsymbol{V}$ eigenvectors. Therefore, in this case it can be assumed that there is a relationship between s_i and λ_i, $i = 1, \ldots, n$. Note that this relationship is consistent with what de los Campos et al. (2010) presented for semiparametric regression.

Knapick et al. (2012) proposed that the variances of the prior distribution be represented as $\lambda_i = \varphi i^{-1-2\alpha}$ $(i = 1, \ldots, n)$ in an attempt to mimic the decay of the singular values of $\boldsymbol{X}$, as suggested above. Thus φ is a parameter that scales the rate $i^{-1-2\alpha}$ of decay of the prior variance with respect to i. Knapick et al. (2012) suggested fixing α and letting φ be the regularization parameter to be inferred, but without indicating how to do it. Since φ is an unknown scale parameter, it seems appropriate to assign an inverse scaled chi-squared distribution to it with parameters $\upsilon_\varphi, Sc_\varphi$. However, this may produce over-shrinkage, which would be reflected in a poor fit and, thus, low prediction power. To address this problem, we redefined λ_i as

$$\lambda_i = \varphi(i^{-1-2\alpha} + h) \tag{17}$$

where h represents a smoothing parameter. The idea of adjusting the decay of a prior variance with a smoothing parameter was given by Maruyama and George (2011).

Thus the posterior density of $\left(\boldsymbol{b}, \varphi, Sc_\varphi, \boldsymbol{\sigma_\varepsilon^2}\right)$, given $\boldsymbol{d}, \boldsymbol{S}$, is:

$$p\left(\boldsymbol{b}, \varphi, Sc_\varphi, \sigma_\varepsilon^2, \middle| \boldsymbol{d}, \boldsymbol{S}, h, \upsilon_\varphi, \alpha, \upsilon_\varepsilon, Sc_\varepsilon, \right) \propto$$
$$\{\prod_{i=1}^{n} N(d_i|s_i b_i, \sigma_\varepsilon^2) N(b_i|0, \lambda_i)\} \chi^{-2}(\varphi|\upsilon_\varphi, \upsilon_\varphi Sc_\varphi)\, p(Sc_\varphi) \chi^{-2}(\sigma_\varepsilon^2|\upsilon_\varepsilon, Sc_\varepsilon) \tag{18}$$

where $p(Sc_\varphi)$ is a proper prior distribution for Sc_φ.

Bayesian inverse regression with decay model 2 (BIR2)

A variant of BIR1 is obtained by defining $\lambda_i = \varphi s_i^{\alpha}$. In this case, we may assign as prior for φ an inverse scaled chi-squared distribution, with α kept fixed. Thus the posterior is

$$p\left(\boldsymbol{b}, \varphi, Sc_{\varphi}, \sigma_{\varepsilon}^2 \middle| \boldsymbol{d}, \boldsymbol{S}, \alpha\right) \propto$$
$$\{\prod_{i=1}^{n} N(d_i|s_i b_i, \sigma_{\varepsilon}^2) N(b_i|0, \varphi s_i^{\alpha})\} \chi^{-2}\left(\varphi \middle| \upsilon_{\varphi}, \upsilon_{\varphi} Sc_{\varphi}\right) p(Sc_{\varphi}) \chi^{-2}(\sigma_{\varepsilon}^2|\upsilon_e, Sc_e) \qquad (19)$$

with $p(Sc_{\varphi})$ being a proper prior distribution. When the value of α increases, shrinkage tends to increase, and when α decreases, shrinkage decreases and approximates BIRR when α=0. In the examples below, α=1 (the default value).

Gibbs sampler

It is difficult to sample from the joint posterior distributions (15), (16), (18) and (19), since there are no known closed forms. However, it is possible to obtain the closed form for conditional distributions of the parameters (see the Appendix). This allows using Markov Chain Monte Carlo (MCMC) through the Gibbs Sampler (Gelfand and Smith, 1990) algorithm, which samples sequentially from the full conditional distributions until it reaches a stationary process, converging with the joint posterior distribution.

We carried out convergence and diagnostic tests on different data sets. The Gelman-Rubin convergence tests for the four models were satisfactory, using lag-10 thinning results in low autocorrelations in each of the four models. The Raftery-Lewis test suggested a small burn-in and a number of iterations between 10,000 and 20,000 for the data set used below.

With the aim of decreasing the possible impact of MCMC errors on prediction accuracy, we performed a total of 60,000 iterations with a burn-in of 10,000 and a thinning of 10 so that 5,000 samples were used for inference.

EXPERIMENTAL DATA

Maize data set

The maize data represented 21 trait-environment combinations for 300 tropical inbred lines genotyped with 55,000 SNPs each (González-Camacho et al., 2012). A first group of traits included female flowering (FFL) or days to silking, male flowering (MFL) or days to anthesis, and the anthesis-silking interval (ASI). Each trait was evaluated under severe drought stress (SS) and in well-watered (WW) environments. This data set was also used by Crossa et al. (2010) for assessing prediction performance, but using only 1,148 SNPs.

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

Wheat data set

This data set included 306 elite wheat lines from CIMMYT's Global Wheat Program that were also used by Pérez-Rodríguez et al. (2012). These lines were genotyped with 1,717 diversity array technology (DArT) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). Two traits were analyzed: days to heading (DTH) measured in ten different environments, and GY measured in seven different environments.

Comparing models using cross-validation

Model predictions for each of the maize and wheat data sets were done in each of 50 random partitions, with 90% of the individuals in the training set and 10% of individuals in the testing set to be predicted. We used the same 50 random partitions as González-

Camacho et al. (2012) and Pérez-Rodríguez et al. (2012). Pearson's correlation between predicted and observed values and the Predictive Mean Squared Error (PMSE) were used as measures of predictive ability.

Data and software

The 21 maize data sets and the 17 wheat data sets, as well as the R scripts developed to fit the predictive statistical models BIRR, BAI, BIR1 and BIR2 used in this study, are deposited at http://repository.cimmyt.org/xmlui/handle/10883/4036.

RESULTS

Maize data sets

Table 1 shows mean correlations and mean PMSE from 50 random partitions for four models (BIRR, BAI, BIR1 and BIR2), along with the results of Bayesian Ridge Regression (BRR), obtained using the BLR package in R (de los Campos and Pérez-Rodríguez, 2010; Pérez-Rodríguez et al., 2010). The largest correlations and smallest PMSE for each trait-environment combination are in boldface. Models BIRR and BAI gave results similar to those of BRR. Models BIR1 and BIR2, which explicitly considered the decay of the singular values in the prior distribution, in general had higher average correlations and lower average PMSE, indicating better prediction ability than BRR, BIRR and BAI.

Table 1 also shows average correlations and percentage differences from BIR1 for each group of trials: Flowering, GY, GLS, NCLB. The approximate differences from BIR1 or BRR and BIRR were 4% for GY and 9% for NCLB, whereas the differences for GLS and Flowering were 2.6% and 1.2%, respectively. BIR1 and BIR2 had similar performance. BAI and BIR1 had similar results in the Flowering and NCLB groups, but BAI was slightly worse for GY and GLS. This indicates a 3% improvement in the overall average prediction ability of BIR1 and BIR2 over that of BRR, BIRR and BAI. Model BIR1 had higher predictive correlations than BRR in 18 of the 21 trait-environment combinations, whereas

BIR1 had smaller PMSE than BRR in 18 of the trait-environment combinations. Models BIR1 and BIR2 had the lowest PMSE in 16 of the 21 maize data sets. The PMSE of BIRR and BAI were similar.

Table 2 shows the number of times a model had a higher correlation than another model in 50 random partitions for each of the 21 trait-environment combinations. BIR1 was better than BIRR in 17 trait-environment combinations, i.e., in 697 of 1050 comparisons. BIR1 was better than BAI in 16 trait-environment combinations, i.e., in 669 partitions. Also, BIR2 was superior to BIRR and BAI in 640 and 610 partitions, respectively, of a total of 1050 random partitions.

Wheat data sets

Table 3 shows mean correlations of models BIRR, BAI, BIR1 and BIR2, along with results for models BL, BRR, Bayes A and Bayes B, as reported by Pérez-Rodríguez et al. (2012) using the same 50 random partitions. The largest average correlation for each trait-environment combination is in boldface. The table also shows correlation standard deviations (in parentheses). In the case of trait DTH, the average correlations for BIRR, BAI, BIR1 and BIR2 are very similar and better than those for BL, BRR, Bayes A and Bayes B. Within the GY group, BIRR, BAI and BIR1 have similar correlations; however, BIR2, BL, BRR and Bayes B exhibit a smaller global average than BIRR, BAI and BIR1. Table 4 shows the PMSE for both traits, and their values are in agreement with the results in Table 3.

DISCUSSION

The SVD transformation of the data generates a basic probability model such that the joint density of the data (given parameters) is the product of univariate independent random variables, $p(\boldsymbol{d}|\boldsymbol{S},\boldsymbol{b},\sigma_\varepsilon^2)=\prod_{i=1}^{n}N(d_i\,|s_ib_i,\sigma_\varepsilon^2)$. Moreover, the new parameterization allows reducing the dimensionality of the vector of regression parameters

from p to n. Since p-n parameters are not estimable, they do not contribute to data fitting and are considered to have a prior distribution with mean and variance equal to zero (a Dirac delta function). Since the proposed transformation yields n positive singular values, it reduces the number of parameters to n. This gives BIRR, BAI, BIR1 and BIR2 an extra computational advantage over BRR because the Gibbs sampler algorithm is faster due to the posterior distributions simulating a smaller number of parameters with univariate distributions. The proposed inverse Bayesian models implement the idea put forward by Casella (1985) that the increased instability of the OLS estimates b_i^* can be ameliorated by decreasing the value of the weighting factor. Shrinkage is achieved using the weighting factor(f_i), that is, $E(b_i|, d_i, s_i, \lambda_i, \sigma_\varepsilon^2) = f_i\, b_i^*$, which weights the estimates b_i^* such that the instability of b_i^* is controlled with a smaller weighting factor. The weights are highly dependent on the singular values and on the variance of the λ_i prior distribution.

In the following section, we describe the prediction performance of the four inverse Bayesian methods in terms of the magnitude and decay of the prior variance and the behavior of the OLS estimates b_i^*. Finally, we consider the pattern of noise amplification as a result of the decay of the singular values, which also depends on the data d_i.

Shrinkage and prediction in the maize data set

For the 21 maize data sets, the $\boldsymbol{X}$ matrix had a rank equal to n with moderate decay in singular values, indicating moderate ill-conditioning. Figs. 1a-c show the decay of singular values for data set MFL-WW. Fig. 1a shows all singular values, and Fig. 1b excludes the first 10 singular values and depicts the rapid decay in the first subsequent 45 singular values, a smaller rate of decay in singular values 45 to 200 and the rapid decay of singular values after the 200th singular value, reflecting the actual collinearity trend. Fig. 1c shows that the OLS estimator of b_i is very stable for the first 200 singular values and becomes very erratic at the end due to noise, coinciding with the rapid decrease in singular values after the first 200. In short, Figs. 1a-c show that more weight should be given to the first 200 least squares estimates (b_i^*) and less weight should be assigned to the last 50 $b_i^{*\prime}$s, which basically represent noise. The four inverse Bayesian methods had good predictive

power, with differences being due to the variance assigned to the prior distribution of parameter b_i.

Figs. 2a-b depict the prior variance for models BIRR, BAI, BIR1 and BIR2 for trait-environment combination MFL-WW. In Fig. 2a, the prior variance of b_i for BIRR is represented by a solid line, whereas λ_i used in BAI are scattered dots, each representing an individual. It is interesting to note that, for MFL-WW, most of the λ_i values for BAI, BIR1 and BIR2 were smaller than those for BIRR, represented by a solid line. This indicates that BAI, BIR1 and BIR2 cause more shrinkage. Fig. 2b depicts the decay of singular values for BIR2 (dashed line) and BIR1 (solid line), both mimicking the current decay of singular values shown in Figs 1a-b. The decay of BIR1 reflects the polynomial function i^{-1}, but smoothed by the parameter *h*, as indicated in (18), with less shrinkage for the first singular values and increasing shrinkage towards the later singular values.

These prior variances are reflected in the weighting factor. For example, Fig. 3a shows the weights for BAI and for BIRR and, in general, there is more shrinkage for BIRR than for BAI. Fig. 3b shows that the weights for BIR1 are slightly larger than those for BIR2. When comparing Figs. 2 and 3, it is clear that as shrinkage increases, the weights come closer to 0. The four inverse Bayesian methods assigned weights that parallel the decay of the singular values (Figs. 1a-b), but for trait MFL-WW, the methods that reduced the weighting factors the most were BIR2 and BIR1, which had larger mean predictive correlations (0.850 and 0.841, respectively) than BIRR, with a mean predictive correlation of 0.822 (Table 1). This may be a reflection of some over-fit caused mainly by larger weights used in the last 50 estimators, where noise is concentrated (Fig. 1c). Trends were different for other traits; for example, for trait-environment combination GLS-3, BIRR, BIR1 and BIR2 had mean predictive correlations of 0.589, 0.586, and 0.584, respectively.

Comparing Fig. 1c (OLS estimates for MFL-WW) and Fig. 4b (OLS estimates for GLS-3), instability of OLS estimates is observed for the last 50 OLS values in MFL-WW but this occurred in only a few of the last OLS estimates for GLS-3. This indicates that trait MFL-WW required more shrinkage than that needed for trait GLS-3. This may explain why, for MFL-WW, model BIR2, which showed more shrinkage than BIRR, had a better predictive correlation than BIRR. For the same reasons, model BIRR was a better predictor

than BIR2 for trait GLS-3. BIR1 had a good predictive correlation for most traits because the decay of the variance was smoothed out by $h = \frac{1}{s_r} \sim 0.10$ (see Eq. 17), which allowed 'intermediate' shrinkage.

Shrinkage and prediction in the wheat data set

The decay pattern of singular values and the noise in the wheat data set were different from those found in the maize data set. The $\boldsymbol{X}$ matrix had a rank equal to n-1, indicating stronger collinearity and, thus, higher decay of singular values than that found in the maize data sets; in addition, data for various trait-environment combinations had outliers and were more complex. This is shown, for example, in Figs. 5a-c for trait-environment combination DHT-1. Fig. 5a shows that the decay of singular values for this trait was more pronounced than that for MFL-WW in maize (Figs. 1a-b). This data set has more collinearity than the maize data, making it difficult to visualize and detect where random noise is manifested in the OLS estimates; this represents a moderate to high ill-conditioning situation. Fig. 5b shows that one singular value close to zero (upper right-hand side of the figure) magnified the OLS estimate, but when that outlier was removed, the noise pattern was more clearly delineated (Fig. 5c). Noise increased systematically towards the last singular values, where the signal was practically lost.

In cases with strong collinearity, Kyung et al. (2010) suggested ridge regression as an effective model. This may explain, in part, why BIRR, BAI and BIR1 had similar performance, whereas BIR2, with more shrinkage than the other models, had a mean predictive correlation similar to those of other models but with differential performance for each trait-environment combination. The greater complexity of the $\boldsymbol{X}$ matrix, as well as of the phenotypes included in the wheat data, suggests that other prediction models, such as the semi-parametric regression described by Pérez-Rodríguez et al. (2012), may be more suitable.

CONCLUSIONS

Models were developed within the framework of inverse regression theory. Inverse solutions induce some parsimony while keeping conditional independence of the transformed phenotypes. Inverse solutions make it possible to visualize that noise is inversely proportional to singular values. The univariate structure allows graphical depiction of shrinkage behavior according to the prior variance and weighting factor. The models developed here seem to deal well with the ill-conditioning and random noise problems arising in genomic prediction, where $p>>n$.

The differences among models depend on several factors, such as the pattern and decay of singular values. It is expected that when the decay of the singular values is moderated, the prediction accuracy of the proposed models will be adequate or improved. For the maize data set, with a large number of markers, the decay of singular values (c) was less than 1 with a low level of ill-conditioning. In these cases, BIR1 and BIR2 seemed to give slightly better predictions than BL, BIRR and BAI. On the other hand, the wheat data set had a higher level of ill-conditioning, with several outliers and a drastic decay of singular values. BIRR, which assigns a constant prior variance (as standard Bayesian ridge regression), tends to over- or under-shrink; this would favor markers with small singular values while penalizing markers with large singular values.

REFERENCES

Aster, R., Borchers, B., Thurber, C. (2005). Parameter Estimation and Inverse Problems. Elsevier Academic Press. ISBN:0-12-065604-3.

Casella, G. (1985). Condition numbers and minimax ridge regression estimators. J. Amer. Statist. 8 1036-1056.

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

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., Araus, J., Makumbi, D., Singh, R., Dreisigacker, S., Yan, J., Arief, V., Bänsiger, M., Braun, H. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. DOI:10.1534/genetics 10.118521.

Crossa, J., Pérez-Rodríguez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. Journal of Crop Improvement 25: 239-261.

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

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

de los Campos, G., Pérez-Rodríguez, P. (2010). BLR: Bayesian linear regression. R package version 1.1. http://www.r-project.org/.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D, Calus, M.P.L. (2012). Whole genome regression and prediction methods applied to plant and animal breeding. Genetics. [doi:10.1534/genetics.112.143313]). PMID: 22745228.

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

Gianola, D., Fernando, R., Stella, A. (2006). Genomic assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761-1776.

Gianola, D., van Kaam, J. B.C.H.M. (2008). Reproducing kernel Hilbert space regression methods for genomic assisted prediction of quantatative traits. Genetics 178(4): 2289-2303.

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

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

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

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

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

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

Kyung, M., Gill, J., Ghosh, M., Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lasso. Bayesian Analysis 5(2): 369-412.

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

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

Park, T., Casella, G. (2008). The Bayesian LASSO. Journal of American Statistical Association 105(482): 681-686.

Pérez-Rodríguez, P., de los Campos, G., Crossa, J., Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. The Plant Genome 3(2): 106-116.

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

Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics ISBN:0-89871-572-5.

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

VanRaden, P. (2007). Genomic measures of relationship and inbreeding. Interbull Bull. 37: 33-36.

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

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

Table 1. Maize data sets. Mean predicted correlation, average, % increase correlation with respect to model BIR1, and mean Predictive Mean Squared Error (PMSE) of five models: Bayesian Regression Ridge (BRR) , Bayesian Inverse Ridge Regression (BIRR), Bayes A Inverse (BAI), Bayesian Inverse Regression models 1 and 2 (BIR1 and BIR2, respectively) for 50 random partitions for each of 21 trait-environment combinations. The largest correlations and smallest PMSE for each trait-environment combination are in boldface.

	Mean correlation					Mean PMSE				
Trait-environment combination[+]	BRR	BIRR	BAI	BIR1	BIR2	BRR	BIRR	BAI	BIR1	BIR2
FFL-WW	0.818	0.819	0.828	0.842	**0.847**	0.262	0.260	0.219	0.201	**0.196**
FFL-SS	0.754	0.755	0.759	0.762	**0.764**	0.342	0.325	0.325	0.324	**0.320**
MFL-WW	0.822	0.822	0.829	0.841	**0.850**	0.263	0.263	0.225	0.203	**0.198**
MFL-SS	0.776	0.777	0.784	0.782	**0.788**	0.318	0.318	0.299	0.298	**0.293**
ASI-WW	**0.582**	**0.582**	0.580	0.578	0.574	0.649	**0.648**	0.649	0.651	0.655
ASI-SS	0.613	0.614	**0.617**	0.614	0.611	0.653	0.652	**0.647**	0.649	0.650
Average	0.727	0.728	0.733	0.736	**0.739**					
% increase	-1.2%	-1.1%	-0.5%	0%	**0.4%**					
GY-SS	0.320	0.326	0.305	0.354	**0.360**	0.890	0.883	0.904	0.866	**0.862**
GY-WW	0.557	**0.558**	0.555	**0.558**	**0.558**	0.677	0.675	0.680	**0.674**	0.675
GY-HI	0.634	0.635	0.664	0.667	**0.674**	0.597	0.595	0.558	0.554	**0.546**
GY-LOW	0.410	0.412	0.419	**0.424**	**0.**423	0.863	0.845	0.847	**0.835**	0.837
Average	0.480	0.483	0.486	0.501	**0.503**					
% increase	-4.2%	-3.7%	-3.1%	0%	**0.4%**					
GLS 1	0.241	0.238	0.260	0.282	**0.287**	0.930	0.928	0.920	0.902	**0.900**
GLS 2	0.419	0.421	0.414	**0.427**	**0.426**	0.827	0.825	0.835	**0.819**	**0.819**
GLS 3	0.588	**0.589**	0.585	0.586	0.584	0.625	**0.624**	0.629	0.627	0.629
GLS 4	0.522	0.522	0.529	**0.533**	0.528	0.735	0.733	0.728	**0.721**	0.729
GLS 5	0.338	0.341	0.348	**0.360**	0.356	0.819	0.815	0.807	**0.801**	0.803
GLS 6	0.257	0.274	0.241	**0.278**	0.276	0.975	0.964	0.986	**0.959**	0.960
GLS 7	0.474	0.475	0.472	**0.480**	0.484	0.761	0.761	0.766	0.756	**0.750**
GLS 8	0.595	**0.596**	0.593	0.593	0.591	**0.618**	0.617	0.621	0.620	0.622
GLS 9	0.522	0.522	0.529	**0.533**	0.530	0.734	0.732	0.728	**0.723**	0.727
Average	0.442	0.442	0.441	**0.453**	0.445					
% increase	-2.9%	-2.4%	-2.6%	**0%**	-0.2%					
NCBL 1	0.649	0.649	**0.697**	0.693	**0.697**	0.592	0.592	**0.521**	0.527	0.523
NCBL 2	0.469	0.473	0.523	**0.526**	0.524	0.731	0.723	**0.673**	0.679	0.680
Average	0.559	0.561	0.61	0.610	**0.611**	0.656	0.658	0.646	**0.638**	0.639
% increase	-9.0%	-8.7%	0.2%	0%	**0.7%**					

+ FFL: female flowering; MFL: male flowering; ASI: MFL to FFL interval; GY: grain yield; SS: severe drought stress; WW: well-watered environment; HI: optimum environment; LOW: stress environment; GLS: *Cercospora zeae-maydis*; NCLB: *Exserohilum turcicum*.

1

Table 2. Maize data sets. Number of times one model had a higher correlation than another model in 50 random partitions for each of the 21 trait-environment combinations. The models were Bayesian Inverse Ridge Regression (BIRR), Bayes A Inverse (BAI), Bayesian Inverse Regression models 1 and 2 (BIR1 and BIR2, respectively).

	Number of times one model had a larger correlation than another model over 50 random partitions				
Trait-environment combination[+]	BIR1>BIRR	BIR1>BAI	BIR2>BIRR	BIR2>BAI	BIR2>BIR1
FFL-WW	38	32	40	35	28
FFL-SS	36	27	35	40	33
MFL-WW	41	36	41	38	33
MFL-SS	39	30	39	35	30
ASI-WW	24	23	19	16	16
ASI-SS	25	20	19	18	17
GY-SS	33	41	36	41	31
GY-WW	26	33	25	34	22
GY-HI	42	33	43	30	26
GY-LOW	32	29	31	27	25
GLS 1	35	33	37	33	29
GLS 2	31	34	30	33	19
GLS 3	21	28	18	26	20
GLS 4	32	29	28	23	17
GLS 5	33	32	31	29	21
GLS 6	26	32	26	32	21
GLS 7	34	37	31	34	31
GLS 8	18	24	19	23	24
GLS 9	30	29	29	24	24
NCBL 1	46	19	46	26	33
NCBL 2	46	15	44	21	23
Average	33	29	32	29	25

+ FFL: female flowering; MFL: male flowering; ASI: MFL to FFL interval; GY: grain yield; SS: severe drought stress; WW: well-watered environment; HI: optimum environment; LOW: stress environment; GLS: *Cercospora zeae-maydis*; NCLB: *Exserohilum turcicum*. Environment 1-9.

1

2

1

Table 3**.** Wheat data sets. Mean predictive correlations (SD in parentheses) for grain yield (GY) and days to heading (DTH) by environment for eight models: Bayesian LASSO (BL), RR-BLUP (BRR), Bayes A, Bayes B (Pérez-Rodríguez et al., 2012), Bayesian Inverse Ridge Regression (BIRR), Bayes A Inverse (BAI), Bayesian Inverse Regression models 1 and 2 (BIR1 and BIR2, respectively). Average correlation for each model and % increase correlation with respect to model BIR1. The largest values for each trait-environment combination are in boldface.

Trait	Environment	BL	BRR	Bayes A	Bayes B	BIRR	BAI	BIR1	BIR2
DTH	1	0.59(0.11)	0.59(0.11)	0.59(0.11)	0.56(0.11)	**0.60**(0.11)	0.60(0.11)	0.60(0.11)	0.59(0.11
	2	0.58(0.14)	0.57(0.14)	0.61(0.12)	0.57(0.13)	0.62(0.14)	0.61(0.14)	0.62(0.14)	**0.65**(0.12)
	3	0.60(0.13)	0.60(0.12)	0.62(0.11)	0.60(0.12)	0.62 (0.12)	0.62(0.12)	0.62(0.12)	**0.63**(0.12)
	4	0.02(0.18)	**0.07**(0.17)	0.06(0.17)	0.06(0.17)	0.06(0.18)	0.01(0.18)	**0.06**(0.18)	0.04(0.20)
	5	0.65(0.09)	0.64(0.10)	0.66(0.09)	0.66(0.09)	**0.65** (0.09)	**0.65**(0.09)	**0.65**(0.09)	**0.65**(0.09)
	8	0.36(0.15)	0.37(0.15)	0.36(0.15)	0.35(0.14)	0.38(0.15)	**0.39**(0.15)	0.38(0.15)	0.36(0.15)
	9	0.59(0.12)	0.59(0.11)	0.53(0.12)	0.52(0.11)	**0.60**(0.11)	**0.60**(0.11)	0.60(0.11)	0.57(0.11)
	10	0.54(0.14)	0.52(0.14)	0.56(0.13)	0.54(0.14)	0.55(0.13)	0.54 (0.14)	0.55(0.13)	**0.56**(0.12)
	11	0.52(0.15)	0.52(0.16)	0.53(0.13)	0.51(0.13)	0.54(0.15)	**0.54**(0.16)	0.54(0.15)	0.52 (0.13)
	12	0.45(0.19)	0.42(0.18)	0.45(0.18)	0.45(0.18)	0.47 (0.18)	0.47(0.18)	0.48(0.18)	**0.49** (0.19)
	Average	0.49	0.49	0.50	0.48	0.51	0.50	**0.510**	**0.51**
	% increase	-4.0%	-4.3%	-2.6%	-5.8%	-0.1%	-1.1%	**0%**	**-0.7%**
GY	1	0.48(0.13)	0.43(0.14)	0.48(0.13)	0.46(0.13)	**0.48**(0.13)	**0.48**(0.14)	**0.48**(0.13)	0.47(0.13)
	2	0.48(0.14)	0.41(0.17)	0.48(0.14)	0.48(0.14)	0.47 (0.15)	0.47(0.15)	0.47(0.15)	**0.48** (0.13)
	3	0.20(0.21)	0.29(0.22)	0.20(0.22)	0.18(0.22)	0.24 (0.22)	**0.25** (0.21)	0.24(0.22)	0.21 (0.20)
	4	0.45(0.15)	0.46(0.13)	0.43(0.15)	0.42(0.15)	**0.46** (0.14)	**0.46**(0.14)	0.46(0.14)	0.43 (0.14)
	5	0.59(0.14)	0.56(0.16)	0.75(0.11)	0.74(0.12)	0.58(0.15)	0.58(0.15)	0.59(0.15)	0.58 (0.13)
	6	0.70(0.10)	0.67(0.11)	0.73(0.08)	0.71(0.08)	**0.71** (0.1)	**0.71**(0.10)	**0.71** (0.1)	**0.71** (0.1)
	7	0.46(0.14)	0.50(0.14)	0.42(0.14)	0.40(0.15)	**0.48**(0.14)	**0.48**(0.15)	0.48(0.14)	0.42(0.15)
	Average	0.48	0.474	**0.499**	0.484	0.489	0.49	0.489	0.473
	% increase	-1.7%	-2.9%	**2.1%**	-0.8%	0.2%	0.5%	0%	-3.5%

1

2

Table 4**.** Wheat data sets. Mean Predicted Mean Squared Error (PMSE) between observed and predicted values for grain yield (GY) and days to heading (DTH) of wheat lines in environments (1-11) for eight models: Bayesian LASSO (BL), RR-BLUP (BRR), Bayes A, Bayes B (Pérez-Rodríguez et al., 2012), Bayesian Inverse Ridge Regression (BIRR), Bayes A Inverse (BAI), and Bayesian Inverse Regression models 1 and 2 (BIR1 and BIR2, respectively). The smallest values for each trait-environment combination are in boldface.

Trait	Environment	BL	BRR	Bayes A	Bayes B	BIRR	BAI	BIR1	BIR2
DTH	1	13.02	13.18	12.72	13.23	12.81	12.85	12.8	**12.66**
	2	11.89	12.37	**10.65**	11.28	11.64	11.88	11.59	10.63
	3	8.18	8.44	**7.31**	7.59	8.09	8.04	8.08	7.44
	4	**21.59**	22.27	21.79	21.67	21.61	21.57	21.61	22.7
	5	8.86	9.23	8.48	8.37	8.78	8.81	8.78	8.68
	8	14.72	15.22	**14.54**	14.58	14.69	14.68	14.69	14.75
	9	21.38	21.44	23.71	23.93	21.36	21.40	**21.35**	22.22
	10	7.72	8.51	7.27	7.57	7.67	7.91	7.62	**7.27**
	11	6.83	7.12	**6.59**	6.74	6.78	6.74	6.77	6.77
	12	13.60	14.42	13.56	**13.46**	13.66	13.62	13.60	13.42
GY	1	0.07	0.09	0.07	0.07	**0.07**	**0.07**	0.07	0.07
	2	0.06	0.08	0.06	0.06	**0.06**	**0.06**	**0.06**	**0.06**
	3	0.06	0.07	0.06	0.06	**0.06**	**0.06**	**0.06**	**0.06**
	4	0.22	0.24	0.23	0.23	**0.22**	**0.22**	**0.22**	0.23
	5	0.39	0.44	**0.26**	0.27	0.40	0.40	0.40	0.39
	6	0.13	0.15	**0.12**	0.13	0.13	0.13	0.13	0.13
	7	**0.40**	0.41	0.43	0.44	**0.40**	**0.40**	**0.40**	0.43

3

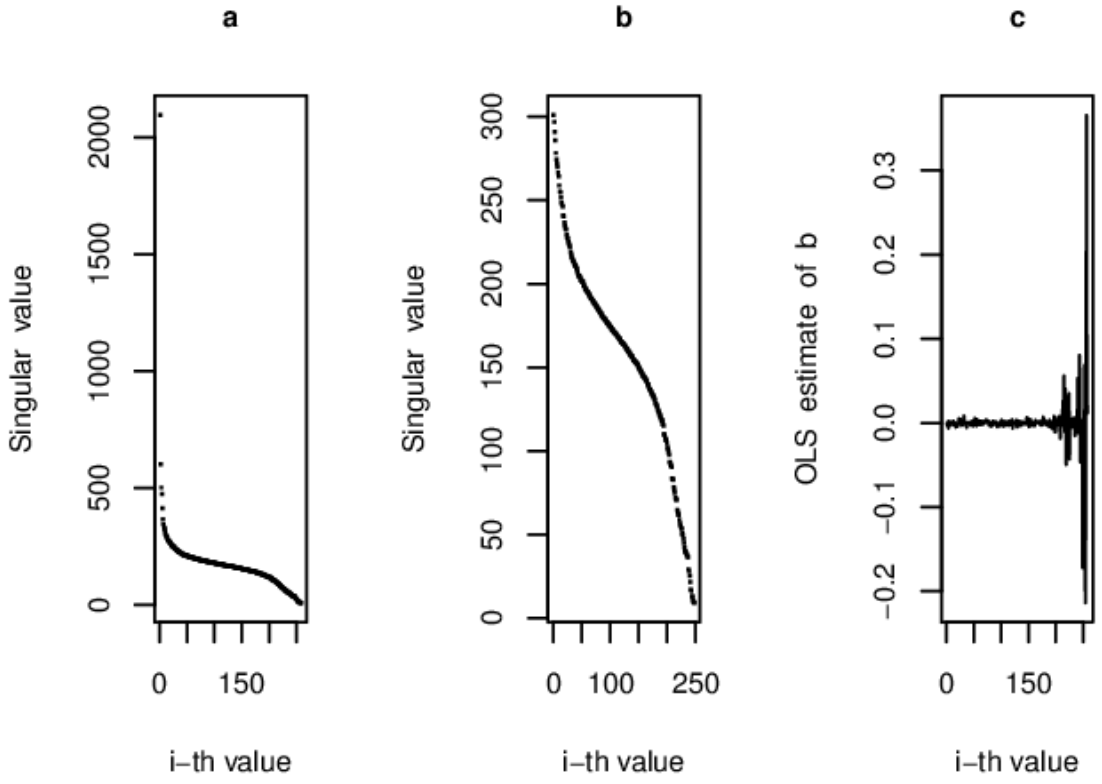

Figure 1a-c. Decay of the singular values and noise pattern for maize trait-environment combination MFL-WW: (**a**) decay of all singular values, (**b**) decay of all singular values except the first 10, (**c**) noise in the ordinary least squares (OLS) estimates for all singular values.

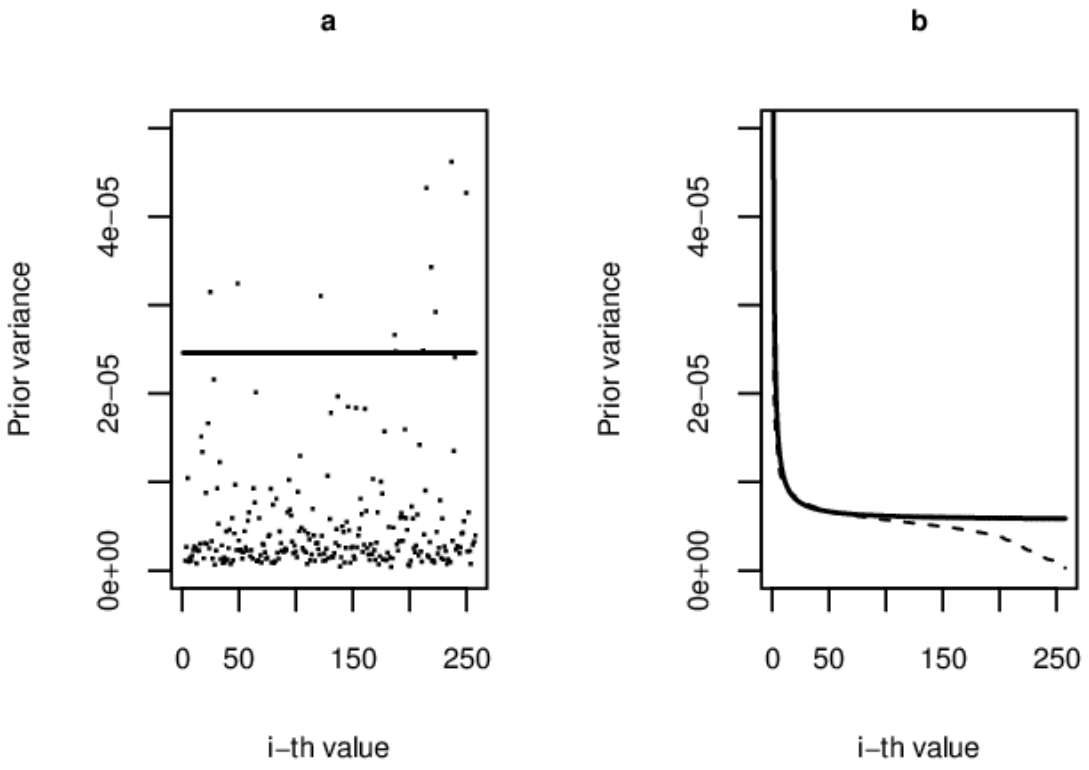

Figure 2a-b. Maize trait-environment combination MFL-WW: the prior variances for the *i*th singular values for four inverse Bayesian regression models: (**a**) BIRR (line) and BAI (dots); (**b**) BIR1 (solid line) and BIR2 (dashed line).

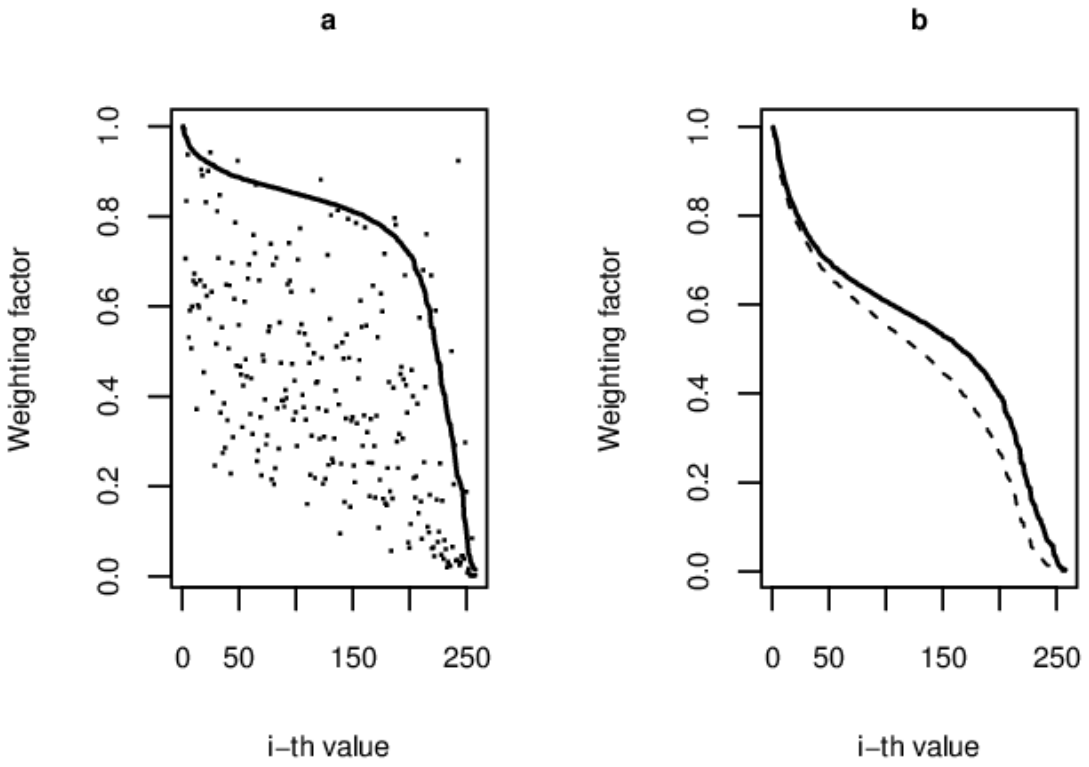

Figure 3a-b. Maize trait-environment combination MFL-WW: weighting factor values for four inverse Bayesian regression models: (**a**) BIRR (solid line) and BAI (dots); (**b**) BIR1 (solid line) and BIR2 (dashed line).

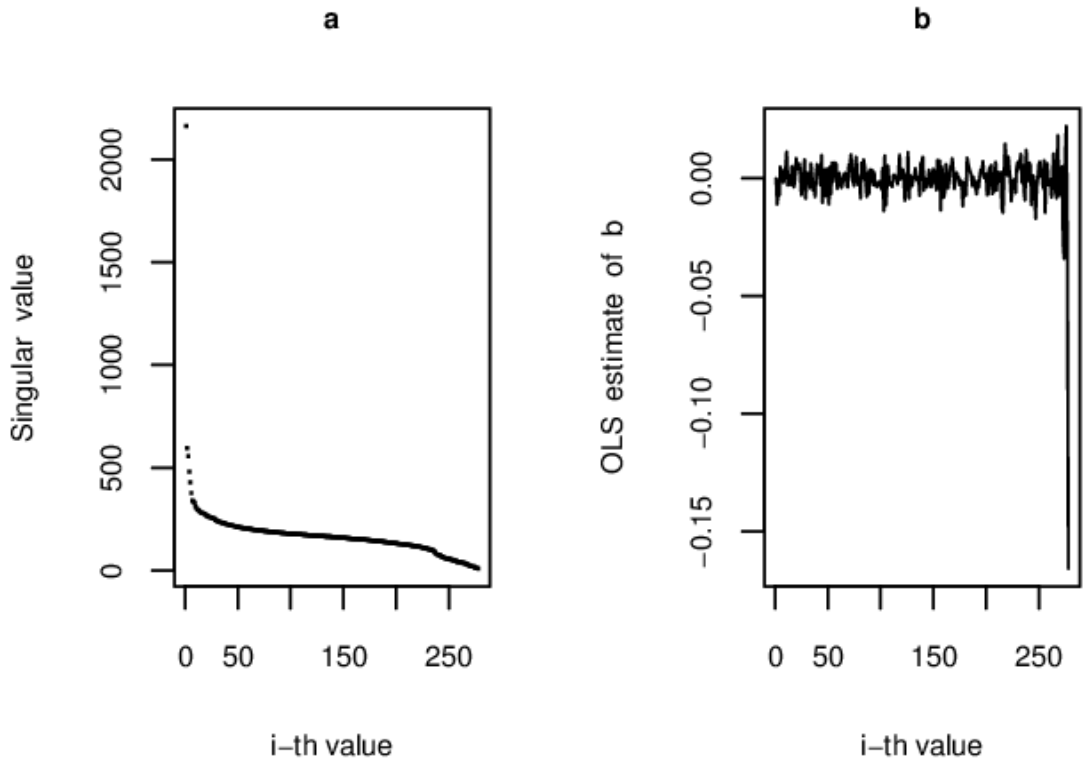

Figure 4a-b. Decay of singular values and noise pattern for maize trait-environment combination GLS-3: (**a**) decay of all singular values; (**b**) noise in the ordinary least squares (OLS) estimates for all singular values.

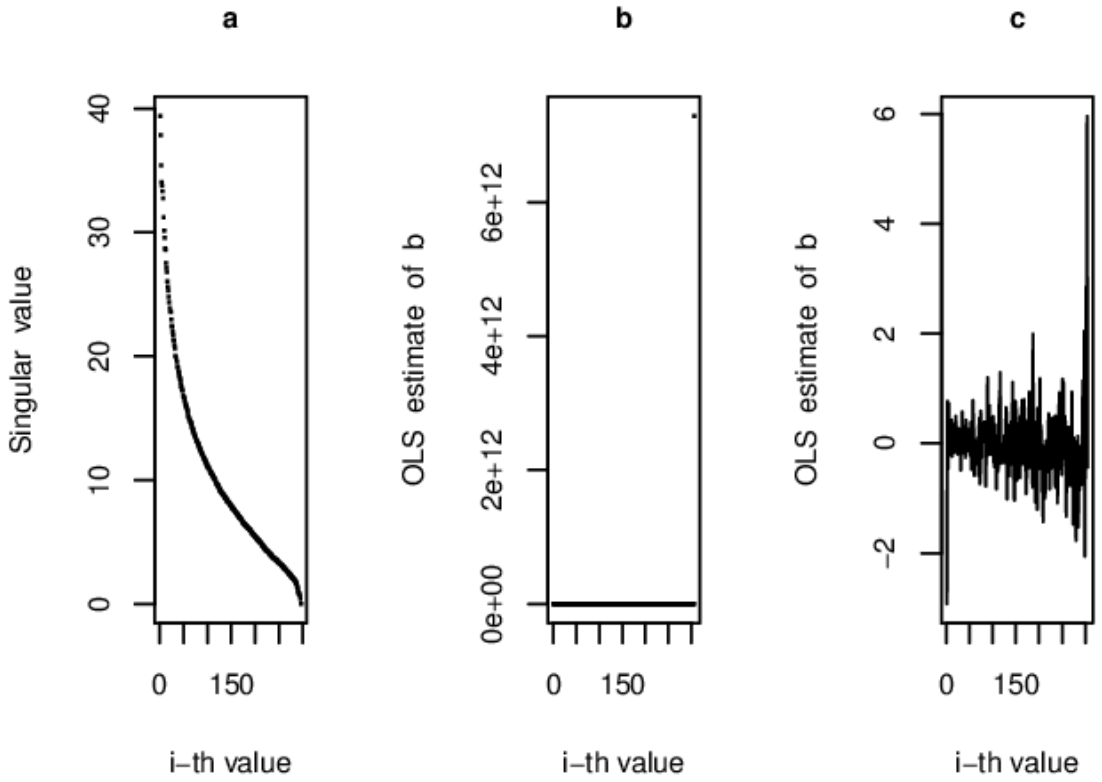

Figure 5a-c. Decay of singular values and noise pattern for wheat trait-environment combination DTH-1: (**a**) decay of all singular values except the first 10; (**b**) noise in the least squares estimates including when all singular values are included; (**c**) noise in the ordinary least squares (OLS) estimates for all singular values except the last one.

APPENDIX

Recalling that model (1) is represented as a deviation from the overall mean μ, closed form distributions can be derived from their joint distributions (15), (16), (18), (19). The conditional distribution of μ is univariate normal:

$$p(\mu|\boldsymbol{y},\boldsymbol{\beta},\sigma_\varepsilon^2) \propto N(\mu|\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\boldsymbol{\beta}),\frac{1}{n}\sigma_\varepsilon^2)$$

The conditional distribution of σ_ε^2 is:

$$p(\sigma_\varepsilon^2|\boldsymbol{d},\boldsymbol{b},\sigma_b^2,\upsilon_e,Sc_e) = \chi^{-2}(\sigma^2|n+\upsilon_e,(\boldsymbol{d}-\boldsymbol{Sb})^t(\boldsymbol{d}-\boldsymbol{Sb})+Sc_e)$$

(small values are given to $\upsilon_e = 0.0001, Sc_e = 0.0001$).

Specifically for the BIRR model, the conditional distributions for σ_b^2 and Sc_b are:

$$p(\sigma_b^2|\boldsymbol{d},\boldsymbol{b},\ \upsilon_b,Sc_b) = \chi^{-2}(\sigma_b^2|n+\upsilon_b,\boldsymbol{b}^t\boldsymbol{b}+\upsilon_b Sc_b)$$

$$p(Sc_b|\boldsymbol{d},\boldsymbol{b},\sigma_b^2,\upsilon_b) = Ga\left(Sc_b\middle|\frac{\upsilon_b}{2}+1,\frac{\upsilon_b}{\sigma_b^2}\right)$$

For the BAI model, the conditional distribution of $\sigma_{b_i}^2$ is:

$$p(\sigma_{b_i}^2|\boldsymbol{d},b_i,\sigma_\varepsilon^2,Sc) = IG\left(\sigma_{b_i}^2\middle|\frac{\nu+1}{2},\frac{\nu Sc+b_i^2}{2}\right)$$

The Bayesian model is completed by giving υ a value of 3 to avoid infinite variance in the distribution of $\sigma_{b_i}^2$. The conditional distribution of Sc is

$$p\left(Sc\middle|\boldsymbol{d},\boldsymbol{b},\upsilon,\boldsymbol{\sigma_b^2}\right) = Ga\left(Sc\middle|\frac{n\upsilon}{2},\left(\frac{\upsilon}{2}\right)\sum_{i=1}^{n}\frac{1}{\sigma_{b_i}^2}\right)$$

For BIR 1 and BIR2, the conditional distributions of φ Sc_φ are:

$$p(\varphi\middle|\boldsymbol{d},b_i,\upsilon_\varphi,Sc_\varphi) = IG\left(\varphi\middle|\frac{\upsilon_\varphi+n}{2},\frac{\boldsymbol{b}'\Omega\boldsymbol{b}+\upsilon_\varphi Sc_\varphi}{2}\right)$$

$$p\left(Sc_\varphi\middle|\boldsymbol{d},\boldsymbol{b},\varphi,\upsilon_\varphi\right) = Ga\left(Sc\middle|\frac{\upsilon_\varphi}{2}+1,\frac{\upsilon_\varphi}{2\varphi}\right),$$

respectively.

For BIR1, $\boldsymbol{\Omega}$ is a diagonal matrix of order $n \times n$, with elements in the diagonal $\frac{1}{i^{-1-2\alpha}+h}$ for i=1,2,…,n. For BIR2, $\boldsymbol{\Omega}$ is equal to matrix $\boldsymbol{S}^{-1}$. For BIR1, the default value of smoothing parameter h is $\frac{1}{s_r}$.