Available online at www.sciencedirect.com

ScienceDirect

Bayesian regularized quantile regression: A robust alternative for genome-based prediction of skewed data

Paulino Pérez-Rodríguez[a], *Osval A. Montesinos-López*[b], *Abelardo Montesinos-López*[c], *José Crossa*[a,d,*]

[a]Colegio de Postgraduados, Montecillos, Edo. de México 56230, Mexico
[b]Facultad de Telemática, Universidad de Colima, Colima, Colima 28040, Mexico
[c]Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Guadalajara, Jalisco 44430, Mexico
[d]International Maize and Wheat Improvement Center (CIMMYT), Km 45 Carretera México-Veracruz, Texcoco, Edo. de México 56237, Mexico

ARTICLE INFO

Article history:
Received 30 October 2019
Received in revised form 14 April 2020
Accepted 30 May 2020
Available online 24 June 2020

Keywords:
Laplace distribution
Robust regression
Bayesian quantile regression
Genomic-enabled prediction

ABSTRACT

Genomic prediction (GP) has become a valuable tool for predicting the performance of selection candidates for the next breeding cycle. A vast majority of statistical linear models on which GP is based rely on the assumption of normality of the residuals and therefore on the response variable itself. In this study, we propose to use Bayesian regularized quantile regression (BRQR) in the context of GP; the model has been successfully used in other research areas. We evaluated the prediction ability of the proposed model and compared it with the Bayesian ridge regression (BRR; equivalent to genomic best linear unbiased predictor, GBLUP). In addition, BLUP can be used with pedigree information obtained from the coefficient of coancestry (ABLUP). We have found that the prediction ability of BRQR is comparable to that of BRR and, in some cases, better; it also has the potential to efficiently deal with outliers. A program written in the R statistical package is available as Supplementary material.

1. Introduction

Genome-based prediction (GP) aims to predict the breeding values of selection candidates of a testing set for which there is only genotypic information (e.g., dense molecular markers panels, pedigree information) based on phenotypic and genotypic information from a set of related individuals known as the training population. The concept of genomic enabled prediction based on markers was first introduced almost two decades ago by Meuwissen et al. [1], and since then, many researchers have extended the models to include other sources of information.

Nowadays empirical evidence suggests that genomic prediction models are widely used by the industry on plants and

[*] *Corresponding author at: Colegio de Postgraduados, Montecillos, Edo.* de México 56230, Mexico.
E-mail address: J.CROSSA@cgiar.org (J. Crossa).

Peer review under responsibility of Crop Science Society of China and Institute of Crop Science, CAAS.

https://doi.org/10.1016/j.cj.2020.04.009

animals and in publicly funded breeding research programs [2–4]. Originally, the linear statistical models proposed for GP relied on the assumption that the residuals are distributed as independent and identically distributed random variables with null mean and the same variance, which implies that the response variable is also distributed normally. It is well known that the selection process leads to distributions that are skewed [5], that is, if we select for trait Y and there is another trait of interest (O), then the conditional distribution of Y is non-symmetric (skew), given that the other trait of interest is greater than a given threshold (o), which can be written as Y ∣ O > o. In the genomics context and in many other breeding situations, it is usual to find a subset of observations that belong to the response variable that differs significantly from the rest, that is, these observations are considered outliers [6]. Outlier observations could be due to several reasons such as missing covariates, ignoring the presence of unknown segregating QTL or because of Genotype × Environment interaction [7]. Therefore, if genomic data are analyzed under these circumstances, inferences could potentially lead to incorrect results.

There are several statistical approaches to deal with non-symmetric distributions, some of which are well known: 1) transformations [8] and 2) robust regression approaches. In the case of robust regression, the assumption of normality is relaxed, for example, when including residuals with skew distributions or thick tails [6,7,9–11]. Montesinos López et al. [6] proposed a model with errors assumed to follow a Laplace distribution which leads to a median regression model; it provides good prediction even when outliers are present in the data; this model is a special case of a more general quantile regression model. Nascimento et al. [10] argued that asymmetry can be overcome with regularized quantile regression and used the model in the genomic context obtaining promising results. Therefore, quantile regression has the potential to simultaneously deal with asymmetric responses and the presence of outliers. Quantile regression also provides a powerful model that can be used to study the relationship between the predictors and the response variable for a set of quantiles, thus providing a more complete view of the phenomenon under study [10,12–14]. This is particularly important in the case of genomic selection, for example, because quantile regression can potentially improve the estimation of marker effects for a given quantile and also opens the theoretical possibility of selecting the quantile function that best represents the relationship between predictors and response, although this is challenging [10]. On the other hand, researchers who use quantile regression are typically interested in well-defined quantiles that are fixed [15] in the context of genomic selection; for example, a breeder can be interested in obtaining reliable genomic breeding values for high yielding varieties or those that are less affected by a disease.

In this study we develop a Bayesian regularized quantile regression (BRQR) model that allows simultaneously including dense molecular markers as well as the additive relationship matrix derived from pedigree. We believe that the model proposed in this study has not been published before in the context of genome-based prediction. The proposed model is an extension of the model developed by Li et al. [14] and is a generalization of the model developed by Montesinos-López et al. [6].

The organization of the article is as follows. In the Materials and methods section, we introduce the quantile regression, both from the frequentist and from the Bayesian perspective; then we introduce the BRQR model with markers and pedigree. We use the Bayesian ridge regression model as a reference when comparing the predictive power of the proposed BRQR model. Next we describe a simulation experiment to assess the performance of the proposed model when the distribution of the response is non symmetric and outliers are present. Next, we present an application of BRQR with real data and we studied the predictive ability of the proposed model through random cross-validation. Finally, we describe the results and present a brief discussion of the results. We included as supplementary material the derivation of the full conditional distributions necessary to implement a Gibbs sampler and R codes that implement the proposed algorithms.

2. Materials and methods

2.1. Quantile regression (QR) model

Consider the following linear model:

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + w_i \tag{1}$$

where y_i is the response variable for case $i = 1, \ldots, n$, $\mathbf{x}_i^t = (x_{i1}, \ldots, x_{ip})$ represents a set of covariates that could be associated with the response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ is a vector of regression coefficients and w_i are independent with their θ-th quantile, $\theta \in (0,1)$. The regression coefficients can be estimated as follows: $\hat{\boldsymbol{\beta}}_{QR} = \arg\min_{\boldsymbol{\beta}}\{\sum \rho_\theta(y_i - \mathbf{x}_i^t\boldsymbol{\beta})\}$, where $\rho_\theta(t) = -(1-\theta)tI_{(-\infty,0)}(t) + \theta tI_{[0,\infty)}(t)$ [16]. A regularized estimate of the regression coefficients, widely used in variable selection and high dimensional data sets, is obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}}_{QR} = \arg\min_{\boldsymbol{\beta}}\left\{\sum \rho_\theta\left(y_i - \mathbf{x}_i^t\boldsymbol{\beta}\right) + \lambda J(\boldsymbol{\beta})\right\}$$

with $\lambda \geq 0$ and $J(\boldsymbol{\beta})$ a function of model unknowns, for example, $J(\boldsymbol{\beta}) = \sum_j \mid \beta_j \mid$ corresponds to the LASSO penalty [17], and $J(\boldsymbol{\beta}) = \sum_j \beta_j^2$ corresponds to the L2 penalty.

The introduction of penalty balances model goodness of fit and complexity [18]. From a Bayesian perspective, regularization is done automatically by selecting appropriate prior distributions for the regression parameters. Yu and Moyeed [13] were the first to introduce the Bayesian representation of quantile regression, where the model is fitted using the Metropolis-Hastings algorithm [19,20], which does not scale well in the context of high dimensional data.

Several authors developed hierarchical representations of the model that makes it possible to easily derive conditional distributions of the parameters of interest and implement a Gibbs sampler [14,15,21–23]. Li et al. [14] proposed a Bayesian hierarchical representation of the model and, in a unifying framework, presented a set of regularization approaches for this model. The model we propose allows efficient implementation of a Gibbs sampler [24,25] to sample from the posterior distribution and obtain estimates of the parameters of interest.

2.2. Bayesian regularized quantile regression model (BRQR)

Next we describe the quantile regression model from a Bayesian perspective; see Li et al. [14] for more details. Let $w_i = \xi_1 v_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{v_i}e_i$, with $v_i \mid \tau \sim \exp(-\tau)$, $e_i \sim N(0,1)$, $\xi_1 = (1 - 2\theta)/(\theta(1 - \theta))$ and $\xi_2 = \sqrt{2/(\theta(1-\theta))}$; then model (1) can be rewritten as model (2):

$$y_i = \boldsymbol{x}_i^t\boldsymbol{\beta} + \xi_1 v_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{v_i}e_i \tag{2}$$

In this context, the variance associated with W is $\sigma_w^2 = \frac{1}{\tau^2} \times \frac{1-2\theta+2\theta^2}{\theta^2(1-\theta)^2}$ [13]. Assignation of prior distributions to $\boldsymbol{\beta}$ will lead to different well-known cases of BRQR, e.g., LASSO, Elastic Net LASSO, etc. In this study, we considered an application of BRQR in the context of genomic-enabled prediction with markers and pedigree for asymmetric responses. Nascimento et al. [10,12,26] applied a non-Bayesian version of quantile regression in the genomic prediction and GWAS context using molecular markers. Montesinos-López et al. [6] proposed a regression model with errors distributed as Laplace random variables and applied it in genome-based prediction in wheat with molecular markers. The model proposed by Montesinos-López et al. [6] is a special case of the general BRQR model when $\theta = 0.50$.

2.3. BRQR with markers and pedigree (BRQR **A** + **M**)

Let's assume that we have a matrix of markers **X**, of dimensions $n \times p$, with $x_{ij} \in \{0,1,2\}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ representing the number of copies of a biallelic marker, for example, SNPs, although it can be used with other type of markers, for example, DArT. Let ***A*** be a relationship matrix derived from pedigree, ***Z*** a matrix that connects phenotypes with genotypes of dimensions $n \times q$. A BRQR that includes an intercept, markers and relationship matrix derived from pedigree jointly using the hierarchical representation in [14] is as follows:

$$y_i = \mu + \boldsymbol{x}_i^t\boldsymbol{\beta} + \mathbf{z}_i^t\boldsymbol{u} + \xi_1 v_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{v_i}e_i \tag{3}$$

where $\boldsymbol{u} \mid \sigma_a^2 \sim MN(\mathbf{0}, \boldsymbol{A}\sigma_a^2)$, *MN* stands for "Multivariate Normal", and σ_a^2 is a variance parameter associated with **A**. Let $\boldsymbol{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t$ be the eigen-value decomposition of ***A***, where $\boldsymbol{\Gamma}$ is the matrix whose columns correspond to the eigen-vectors of **A** and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the corresponding eigen-values. Then $\boldsymbol{u} \overset{d}{=} \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}} \mid \sigma_a^2 \sim \mathrm{MN}(\mathbf{0}, \sigma_a^2\boldsymbol{I})$ and "$\overset{d}{=}$" stands for equality in distribution. Using these results, model (3) is statistically equivalent to:

$$y_i = \mu + \boldsymbol{x}_i^t\boldsymbol{\beta} + \tilde{\mathbf{z}}_i^t\tilde{\boldsymbol{u}} + \xi_1 v_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{v_i}e_i \tag{4}$$

where $\tilde{\mathbf{Z}} = \boldsymbol{Z}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}$. Note that once the predictions for $\tilde{\boldsymbol{u}}$ are obtained, the predictions for $\boldsymbol{u}$ are obtained as $\hat{\boldsymbol{u}} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\widehat{\tilde{\boldsymbol{u}}}$.

2.3.1. Sampling model and likelihood function

Assuming a random sample from model (4), the conditional distribution $y_i \mid \mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, v_i$ is normal with mean $\mu + \mathbf{x}_i^t\boldsymbol{\beta} + \tilde{\mathbf{z}}_i^t\tilde{\boldsymbol{u}} + \xi_1 v_i$ and variance $\frac{\xi_2^2}{\tau}v_i$. The joint conditional distribution of y_i and the unobserved random variable v_i is $p(y_i, v_i|\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau) = p(y_i|\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, v_i)p(v_i|\tau)$; therefore the augmented likelihood is given by:

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{v}|\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau) &= \prod_{i=1}^{n} p(y_i|\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, v_i)p(v_i|\tau) \\ &\propto \frac{\tau^{\frac{3n}{2}}}{\xi_2^n|\mathbf{D}|^{\frac{1}{2}}} exp\left\{-\frac{\tau}{2\xi_2^2}\tilde{\mathbf{y}}^t\mathbf{D}^{-1}\tilde{\mathbf{y}}\right\} \\ &\quad \times exp\left\{-\tau\sum_{i=1}^{n} v_i\right\} \end{aligned} \tag{5}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \mu\mathbf{1} - \mathbf{X}\boldsymbol{\beta} - \tilde{\mathbf{Z}}\tilde{\mathbf{u}} - \xi_1\boldsymbol{v}$ and $\mathbf{D} = diag\ (v_1, \ldots, v_n)$.

2.3.2. Prior distributions

In order to complete the specification of the model, we assign prior distribution to the model unknowns. Let $\boldsymbol{\beta} \mid \sigma_\beta^2 \sim MN(\mathbf{0}, \mathbf{I}\sigma_\beta^2)$, $\sigma_\beta^2 \mid df_\beta$, $S_\beta \sim \chi^{-2}(df_\beta, S_\beta)$, $\sigma_a^2 \mid df_a$, $S_a \sim \chi^{-2}(df_a, S_a)$, with $\chi^{-2}(df, S)$ denoting a scaled inverse chi squared distribution with expected value $S/(df - 2)$, with S the scale parameter and *df* the degrees of freedom [27]. For τ, we assign a gamma distribution, that is, $\tau \mid a, b \sim \Gamma(a, b)$, where *a* and b correspond to the shape and rate parameters, respectively, and finally, $p(\mu) \propto 1$. The joint prior distribution of model unknowns is given by:

$$\begin{aligned} p\left(\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, \sigma_\beta^2, \sigma_a^2|H\right) &\propto p(\mu)p\left(\boldsymbol{\beta}|\sigma_\beta^2\right)p\left(\sigma_\beta^2|df_\beta, S_\beta\right) \\ &\quad \times p(\tilde{\boldsymbol{u}}|\sigma_a^2)p(\sigma_a^2|df_a, S_a) \\ &\quad \times p(\tau|a, b) \end{aligned} \tag{6}$$

where $H = \{df_\beta, S_\beta, df_a, S_a, a, b\}$ is the set of hyper-parameters.

2.3.3. Posterior distribution

The joint posterior distribution of all unknown quantities can be obtained by applying the Bayes theorem, so by combining Eqs. (5) and (6) we obtain:

$$\begin{aligned} p\left(\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, \sigma_\beta^2, \sigma_a^2|data, H\right) &\propto p(\mathbf{y}, \boldsymbol{v}|\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau) \\ &\quad \times p\left(\mu, \boldsymbol{\beta}, \tilde{\boldsymbol{u}}, \tau, \sigma_\beta^2, \sigma_a^2|H\right) \end{aligned} \tag{7}$$

The distribution given in (7) is analytically un-tractable, but the hierarchical structure allows us to obtain the conditional distributions necessary to implement a Gibbs sampler [24,25] and draw samples from the joint posterior distribution, from which marginal distributions and other quantities can be inferred. The full conditional distributions necessary to implement the Gibbs sampler $p(\mu|else), p(\boldsymbol{\beta}|else)$, $p(\sigma_\beta^2|else), p(\tilde{\boldsymbol{u}}|else), p(\sigma_a^2|else)$, $p(\boldsymbol{v}|else)$, $p(\tau|else)$, all have closed form and are given in Supplementary materials.

The set of hyper-parameters *H* can be difficult to define. Here we adopted the strategy used in [18,28] for setting the parameters for the scaled inverted chi squared distribution. The same strategy was used in [6] and will lead to slightly informative, but proper prior distributions, so we set $df_\beta = df_a = 5$, $S_\beta = 0.5 \times V_y \times (df_\beta + 2)/MS_x$, $S_a = 0.5 \times V_y \times (df_a + 2)/$

$MS_{\tilde{Z}}$, where V_y is the sample variance for the response vector, MS_x and $MS_{\tilde{Z}}$ are the average sum of squares of the columns of matrices $\mathbf{X}$ and $\tilde{\mathbf{Z}}$ after centering. Finally, for the shape and rate parameters of the gamma distribution, the parameters are set as $a = b = 0.01$, which also leads to a weakly informative prior.

2.3.4. BRQR with markers (BRQR M)

This model is a special case of model (4) obtained by removing the linear predictor associated with the relationship matrix derived from pedigree, that is:

$$y_i = \mu + \mathbf{x}_i^t\boldsymbol{\beta} + \xi_1 \upsilon_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{\upsilon_i}e_i$$

2.3.5. BRQR with pedigree (BRQR A)

Similarly, this model is a special case of model (4) obtained by removing the linear predictor associated with the markers, and thus obtaining:

$$y_i = \mu + \tilde{\mathbf{z}}_i^t\tilde{\mathbf{u}} + \xi_1 \upsilon_i + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{\upsilon_i}e_i$$

2.4. Bayesian ridge regression with markers and pedigree (BRR A + M)

Here we consider a linear regression model with normal errors. This model is widely used in genomic prediction [29,30] and we include it as a reference. The model is given by:

$$y_i = \mu + \mathbf{x}_i^t\boldsymbol{\beta} + \tilde{\mathbf{z}}_i^t\tilde{\mathbf{u}} + e_i \tag{8}$$

where e_i's are independent and identically distributed random variables with mean 0 and variance σ_e^2. We assign exactly the same prior distributions for unknowns as in the case of BRQR, while in the case of the residual variance, we assign a scaled inverted chi squared distribution, that is, $\sigma_e^2 \mid df_e, S_e \sim \chi^{-2}(df_e, S_e)$. Following the rules in [28], we set $df_e = 5$ and $S_e = 0.5 \times V_y \times (df_e + 2)$. Models that include markers only (BRR **M**) or the information between relatives only (BRR **A**) can be derived easily from model (8) and present no conceptual or computational difficulty.

2.5. Software

The Gibbs sampler algorithm to fit BRQR models was implemented in a program written in R [31]. The routine takes as arguments the vector of phenotypes **y**, matrices **X**, **Z**, **A**, the number of MCMC iterations, a burn-in period, and the hyper-parameters *H*, which are initialized with the previously described default values, but can be modified. The output of the routine consists of estimates of posterior means of all model unknowns, and in the case of missing values for the phenotypes, it provides the mean of the predictive distribution obtained through the MCMC algorithm. The routine is included as supplementary material. In the case of the BRR model, it can be fitted in the BGLR library [28].

2.6. Simulation

Here we considered a simulation experiment to assess the performance of the proposed model in the presence of outliers when dealing with skewed errors. In order to simplify simulations, we considered the case where only markers are available. The simulation presented here takes elements from [6,11]. We generated semi-synthetic data through simulation, and we considered a set of 1279 DArT markers coded as 0 and 1, for 599 wheat lines analyzed originally by Crossa et al. [30], and after that, by many others. We simulated the data using the following linear model:

$$y_i = \mu + \sum_{j=1}^{1279} x_{ij}\beta_j + e_i$$

where $i = 1, \ldots, 599$, with $\mu = 4$. We assume that the errors are coming from a skew distribution, and particularly consider the centered skew normal distribution (SN_C) with mean 0, variance 1 and skewness index γ_1 [32], that is, $e_i \sim SN_C(0, \sigma, \gamma_1)$, with $\sigma = \sqrt{1 - h^2}$, where h^2 was set to 0.5 and corresponds to the simulated trait heritability, $\gamma_1 = \sqrt{\frac{2}{\pi}}\rho^3\left(\frac{4}{\pi} - 1\right)\left(1 - \frac{2\rho^2}{\pi}\right)^{-3/2}$, $\rho \in \{0.75, 0.95, 0.9999\}$, which leads to different degrees of positive skewness. Nascimento et al. [10] used a similar strategy when generating the residuals from an exponential distribution. We sampled 50 marker effects from normal distribution with mean zero and variance $(1 - h^2)/50$ and the rest of the marker effects were set to 0. The positions of markers that were sampled from normal distribution were set at random. In order to introduce outliers in the phenotypes, we generated a certain proportion of the residuals from $e_i \sim SN_C(0, 3, \gamma_1)$ at random, and we considered two proportions, 5% and 10%, so effectively we sampled from a 2-component mixture of centered skew normal distributions. We generated 50 datasets using the simulation parameter described above and we fitted the BRQR with $\theta = \{0.25, 0.50, 0.75\}$ and BRR models. For each simulated dataset, we computed the correlation between true and estimated $\boldsymbol{\beta}$'s, true and estimated signals, i.e., $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$, and the variance component associated with the residuals for both models. The variance associated with the residuals can be used to assess model goodness of fit [30]. In the case of BRQR, the variable $w = \xi_1 \upsilon + \frac{\xi_2}{\tau^{\frac{1}{2}}}\sqrt{\upsilon}e$ plays the role of residual, and its variance is given by $\sigma_w^2 = \frac{1}{\tau^2} \times \frac{1 - 2\theta + 2\theta^2}{\theta^2(1 - \theta)^2}$ as mentioned previously; the posterior mean and the posterior standard deviation can be obtained from MCMC output. We also used the MCMC output to compute a modified version of the Deviance Information Criterion (DIC*) which, according to Spiegelhalter et al. [33], can be used to select between candidate models, in this case, the BRQR with the selected quantiles. Models with smaller DIC* are preferred to models with larger DIC*.

2.7. Real datasets

2.7.1. Maize

Here we considered a dataset from the Drought Tolerance Maize (DTMA) project of CIMMYT's Global Maize Program (http://www.cimmyt.org). The dataset has been analyzed several times [11,34]. The dataset consists of 300 tropical inbred lines genotyped using 1152 SNPs markers. The response variable is gray leaf spot (GLS) caused by the fungus *Cercospora zeae-maydis*. The disease is measured on a scale from 1 to 5, where 1 = no disease, 2 = low infection, 3 = moderate infection, 4 = high infection and 5 = totally infected.

The experiment was conducted at three sites: Kakamega (Kenya), San Pedro Lagunillas (Mexico) and Santa Catalina (Colombia). An additive relationship matrix (***A***) derived from the pedigree of the lines is also available. According to Pérez-Rodríguez et al. [11], the skewness index for GLS evaluated in Kakamega is 0.4785, 0.7276 for San Pedro Lagunillas and 0.3111 for Santa Catalina. The values are positive in all cases; therefore the data are skewed to the right, which implies that the right tail is long, relative to the left tail, so that most of the mass of the distribution is concentrated around small values of the GLS disease (Fig. 1).

2.7.2. Wheat

The wheat dataset considered here is a subset of the data analyzed by Pérez-Rodríguez et al. [35]. The response variable is the days to heading in two locations: Agua Fria and Cd. Obregon in Mexico, where wheat plants were grown under standard and drought agronomic management. The dataset corresponds to elite lines from CIMMYT's Global Wheat Program (http://www.cimmyt.org). The data include 306 lines that were genotyped with 1717 diversity array technology (DArT) markers. An additive relationship matrix (***A***) derived from pedigree for the lines is also available, but it was not originally included in the prediction models. The skewness index for DTH evaluated in Cd. Obregon is −1.9926, and 1.3641 in the case of Agua Fria. Fig. 2 shows violin plots for these data, from which it is clear that the data are skewed.

2.7.3. Random cross-validation

We fitted the BRQR and BRR models including: 1) pedigree, 2) markers, and 3) pedigree and markers jointly. In the case of BRQR, we selected three quantiles: $\theta = \{0.25, 0.50, 0.75\}$ [10]. In order to evaluate the predictive ability of the proposed models, we performed cross-validation analyses. We partitioned the data into training, validation, and testing data. For each environment we generated 50 random partitions with 70% of the observations in training, 10% in validation, and 20% in the testing set. The idea is to evaluate the model's predictive ability using observations in the testing set. The BRQR models depend on the unknown parameter θ, which in real applications of genome-based prediction where phenotypes in the testing set are not known, must be set before doing the predictions; therefore, a method to set this parameter is needed. The selection of the appropriate

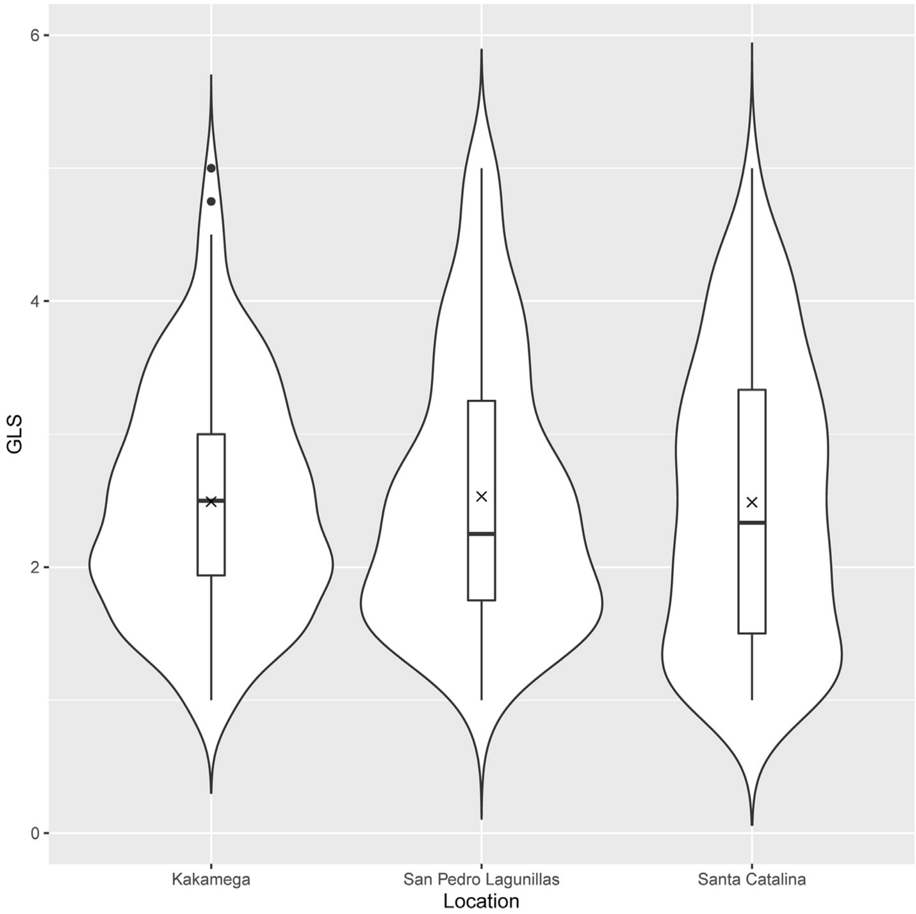

Fig. 1 – Violin plots for GLS in three locations: Santa Catalina (Colombia), Kakamega (Kenya), and San Pedro Lagunillas (Mexico). The mean is represented by the '×' symbol and the median by the horizontal line inside the box. GLS, gray leaf spot.

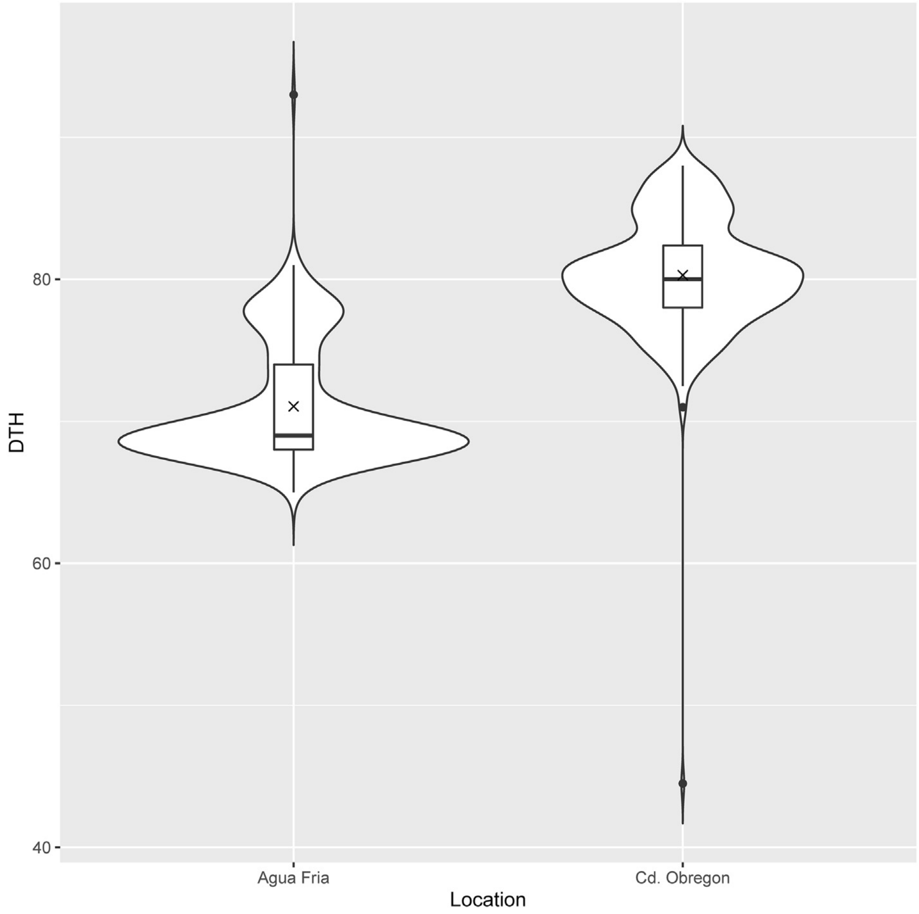

Fig. 2 – Violin plots for days to heading (DTH) in two locations: Agua Fria and Cd. Obregon, Mexico. The mean is represented by the '×' symbol and the median by the horizontal line inside the box.

quantile that "best" represents the relationship between phenotypes and predictors is a challenging task [10]. Here we propose using the validation set to tune this parameter; this approach is widely used in machine learning [36]. BRQR models were fitted with 70% of observations in the training set, and the 10% of observations in the validation set were used to compute the mean squared error, which can be used as a criterion for selecting between models in cross-validation settings. BRR models were fitted using 70% of observations in the training set and 10% of observations in the validation set (80% in total). In all the cases markers were centered and standardized. After fitting the models, we calculated the Pearson's correlation between observed and predicted values in the testing set. Inferences were based on 25,000 MCMC samples obtained after discarding 25,000 samples that were taken as burn-in.

3. Results

3.1. *Simulation*

Table 1 presents the results from the simulation experiment for different degrees of skewness and different proportions of outliers. From the column that corresponds to correlations between 'true' marker effects and estimated marker effects, it can be seen that in general the correlations for BQRR are higher than those obtained with the BRR. The result is even clearer with higher skewness and proportion of outliers. A similar pattern is observed in the case of correlations between 'true' signal ($\mathbf{X}\boldsymbol{\beta}$) and estimated signal ($\mathbf{X}\hat{\boldsymbol{\beta}}$). In the case of DIC*, results were mixed, but in general the index favored models with low quantiles (0.25 and 0.50), which are the ones with the highest correlations between 'true' and estimated marker effects and also the highest correlation between 'true' and estimated signals. DIC* also favored BRQR models, and in all cases the DIC* was larger for the BRR model than for the BRQR models.

3.2. *Random cross-validation*

3.2.1. *Maize*

Results from the random cross-validation obtained as the average Pearson correlation between the observed and predicted values for fitting the BRQR and the BRR models are given in Table 3. The genomic-enabled prediction accuracy measured by the average Pearson correlation shows better prediction accuracy for model BRQR $\theta = 0.25$ using $\boldsymbol{A}$, $\boldsymbol{M}$ and

Table 1 – Correlations between true and estimated marker effects, correlations between true and estimated signals and estimated variance components associated with the residuals and DIC* for different degrees of skewness and different proportions of outliers.

Model	Cor($\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$)	Cor($\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\boldsymbol{\beta}}$)	σ_e^2 or σ_w^2	DIC*
ρ = 0.75, 5% outliers				
BRQR θ = 0.25	0.3130	0.8270	0.8254	1532.12
	(0.0346)	(0.0411)	(0.0926)	(53.57)
BRQR θ = 0.50	**0.3237**	**0.8404**	0.6843	1548.13
	(0.0342)	(0.0367)	(0.0544)	(40.54)
BRQR θ = 0.75	0.3007	0.8143	0.8506	1549.00
	(0.0364)	(0.0394)	(0.0937)	(54.29)
BRR	0.3170	0.8329	0.6788	1602.74
	(0.0361)	(0.0374)	(0.0581)	(46.74)
ρ = 0.95, 5% outliers				
BRQR θ = 0.25	**0.3278**	**0.8417**	0.7697	1495.86
	(0.0341)	(0.0391)	(0.0743)	(47.82)
BRQR θ = 0.50	0.3246	0.8409	0.6763	1540.73
	(0.0351)	(0.0372)	(0.0668)	(52.37)
BRQR θ = 0.75	0.2882	0.7985	0.8824	1568.99
	(0.0382)	(0.0434)	(0.1117)	(63.29)
BRR	0.3166	0.8329	0.6758	1599.13
	(0.0369)	(0.0369)	(0.0792)	(63.18)
ρ = 0.999, 5% outliers				
BRQR θ = 0.25	**0.3413**	**0.8540**	0.7341	1470.60
	(0.0319)	(0.0361)	(0.0613)	(42.58)
BRQR θ = 0.50	0.3265	0.8409	0.6720	1536.28
	(0.0354)	(0.0369)	(0.0631)	(48.07)
BRQR θ = 0.75	0.2823	0.7874	0.8981	1579.77
	(0.0392)	(0.0460)	(0.1140)	(62.88)
BRR	0.3164	0.8318	0.6762	1598.74
	(0.0370)	(0.0369)	(0.0792)	(61.22)
ρ = 0.75, 10% outliers				
BRQR θ = 0.25	0.2995	0.8120	1.0612	1667.28
	(0.0334)	(0.0447)	(0.1171)	(55.47)
BRQR θ = 0.50	**0.3132**	**0.8302**	0.8423	1657.87
	(0.0331)	(0.0383)	(0.0725)	(42.85)
BRQR θ = 0.75	0.2858	0.7982	1.0967	1686.48
	(0.0331)	(0.0410)	(0.1177)	(51.48)
BRR	0.2965	0.8098	0.8650	1734.57
	(0.0365)	(0.0426)	(0.0807)	(47.90)
ρ = 0.95, 10% outliers				
BRQR θ = 0.25	0.3129	0.8267	0.9901	1629.52
	(0.0328)	(0.0417)	(0.1081)	(57.79)
BRQR θ = 0.50	**0.3139**	**0.8306**	0.8366	1652.94
	(0.0341)	(0.0387)	(0.0852)	(53.09)
BRQR θ = 0.75	0.2742	0.7825	1.1426	1707.60
	(0.0346)	(0.0456)	(0.1569)	(68.39)
BRR	0.2939	0.8077	0.8642	1734.01
	(0.0362)	(0.0422)	(0.1012)	(63.20)
ρ = 0.9999, 10% outliers				
BRQR θ = 0.25	**0.3239**	**0.8377**	0.9342	1598.94
	(0.0317)	(0.0395)	(0.0889)	(50.92)
BRQR θ = 0.50	0.3152	0.8298	0.8329	1650.17
	(0.0342)	(0.0394)	(0.0826)	(50.39)
BRQR θ = 0.75	0.2681	0.7712	1.1709	1721.87
	(0.0361)	(0.0488)	(0.1651)	(70.39)
BRR	0.2929	0.8061	0.8694	1736.38
	(0.0366)	(0.0434)	(0.1029)	(62.60)

Results are averages taken from 50 simulated datasets; the standard deviations appear in parentheses. The largest values of the correlations are in bold face.

A + **M** than for the others at the Kakamega site. However, model BRR gave slightly higher prediction accuracy than those obtained by BRQR for sites Santa Catalina and San Pedro Lagunillas. Table 3 presents the results of the mean squared error for the validation set for the BRQR models; for the Kakamega site, the mean squared error in the validation set favored the BRQR with θ = 0.25, which is the one with the highest correlations in the testing set. In the case of Santa Catalina and San Pedro Lagunillas, most of the times the mean squared error favored the models with θ = 0.25, which agrees with the correlations in Table 2. Thus the mean squared error in validation can be a useful tool for selecting the appropriate value of the parameter θ.

3.2.2. Wheat

Table 4 shows the results of the cross-validation analysis for days to heading. In the case of Agua Fria, BRQR models had higher correlations in two of the three scenarios considered. For Cd. Obregon, BRQR models gave the best prediction accuracy in the three scenarios evaluated. Table 5 presents the results of the cross-validation analysis for the mean squared error in the validation set for BRQR models. As expected, low values of the mean squared error in the validation set are associated with high correlations between observed and predicted values in the testing set.

Table 2 – Average of Pearson's correlation and standard deviation (in parentheses) between observed and predicted values in the testing dataset from cross-validation analysis for gray leaf spot.

Predictors*	Model			
	BRQR θ = 0.25	BRQR θ = 0.50	BRQR θ = 0.75	BRR
Kakamega				
A	**0.3098**	0.2888	0.2467	0.2995
	(0.1391)	(0.1473)	(0.1386)	(0.1344)
M	**0.2812**	0.2537	0.2139	0.2502
	(0.1298)	(0.1367)	(0.1330)	(0.1363)
A + **M**	**0.2984**	0.2802	0.2681	0.2796
	(0.1340)	(0.1363)	(0.1317)	(0.1352)
Santa Catalina				
A	0.5583	0.5576	0.5408	**0.5923**
	(0.0850)	(0.0852)	(0.0944)	(0.0944)
M	0.5597	0.5604	0.5521	**0.5786**
	(0.0709)	(0.0728)	(0.0761)	(0.0689)
A + **M**	0.5744	0.5692	0.5702	**0.5927**
	(0.0684)	(0.0763)	(0.0710)	(0.0697)
San Pedro Lagunillas				
A	0.4697	0.4601	0.4467	**0.4875**
	(0.1214)	(0.1328)	(0.1399)	(0.1146)
M	0.4535	0.4569	0.4555	**0.4746**
	(0.1282)	(0.1305)	(0.1315)	(0.1223)
A + **M**	0.4684	0.4617	0.4607	**0.4819**
	(0.1219)	(0.1278)	(0.1239)	(0.1167)

* **A**, additive relationship matrix derived from pedigree; **M**, markers; **A** + **M**, additive relationship matrix derived from pedigree and markers included jointly. The predictions were obtained after fitting the BRQR and BRR models. The average is across 50 random partitions with 20% of observations in the testing dataset. The largest values of the correlations are in bold face.

Table 3 – Mean squared error and standard deviation (in parentheses) for the testing dataset from cross-validation analysis for gray leaf spot.

Predictors*	Model		
	BRQR θ = 0.25	BRQR θ = 0.50	BRQR θ = 0.75
Kakamega			
A	**0.7095** (0.1935)	0.7261 (0.2031)	0.7603 (0.2216)
M	**0.7272** (0.2026)	0.7518 (0.2159)	0.8138 (0.2298)
A + *M*	**0.7195** (0.1942)	0.7441 (0.2116)	0.7689 (0.2250)
Santa Catalina			
A	0.8695 (0.1870)	0.8700 (0.1612)	**0.8510** (0.1647)
M	**0.8542** (0.1914)	0.8685 (0.1937)	0.8846 (0.1807)
A + *M*	0.8457 (0.1820)	**0.8399** (0.1700)	0.8589 (0.1764)
San Pedro Lagunillas			
A	**0.8374** (0.2093)	0.8511 (0.2225)	0.8755 (0.2172)
M	**0.8487** (0.2217)	0.8534 (0.2409)	0.8618 (0.2442)
A + *M*	**0.8341** (0.2139)	0.8473 (0.2230)	0.8478 (0.2264)

* **A**, additive relationship matrix derived from pedigree; **M**, markers; **A** + **M**, additive relationship matrix derived from pedigree and markers included jointly. The predictions were obtained after fitting the BRQR. The average is across 50 random partitions with 70% of observations in the training set and 10% in the validation set. The smallest values of the mean squared error are in bold face.

Table 5 – Mean squared error and standard deviation (in parentheses) for the testing dataset from cross-validation analysis for days to heading.

Predictors*	Model		
	BRQR θ = 0.25	BRQR θ = 0.50	BRQR θ = 0.75
Agua Fria			
A	**15.2904** (7.0106)	15.4978 (6.8626)	16.5817 (6.9084)
M	14.6018 (7.0834)	**14.3631** (7.4346)	15.7730 (6.2434)
A + *M*	**14.7788** (7.0613)	15.1833 (6.6531)	15.5591 (6.5374)
Cd. Obregon			
A	14.3934 (16.3496)	**14.2045** (16.4910)	14.4235 (16.5353)
M	14.6522 (12.8591)	**14.5457** (13.6642)	14.79532 (13.5485)
A + *M*	14.3379 (14.4523)	**14.1391** (14.6813)	14.42722 (14.6266)

* **A**, additive relationship matrix derived from pedigree; **M**, markers; **A** + **M**, additive relationship matrix derived from pedigree and markers included jointly. The predictions were obtained after fitting the BRQR. The average is across 50 random partitions with 70% of observations in the training set and 10% in the validation set. The smallest values of the mean squared error are in bold face.

4. Discussion

Here we introduced a model that is an alternative for skew data and robust in the presence of outliers. Results of simulation show that the model is able to estimate the marker effects more precisely than the BRR model and therefore the signal from the data, i.e., the Genomic Estimated Breeding Values, could potentially also be estimated better with this model. The advantage of the model is more evident when the data are more skewed and it works well even in the presence of outliers; similar results were reported by [6,7]. The results for the real datasets were mixed, but in cross-validation, the BRQR model was better in three of the five datasets. The proposed model is a robust alternative to the widely used G-BLUP and A-BLUP models, and its prediction ability is at the same level when no outliers are present. We considered only three quantiles, $\theta = \{0.25, 0.50, 0.75\}$; a similar approach was employed by Nascimento et al. [10]. We also developed an algorithmic approach that allows selecting the quantile to be used to predict individuals in the testing set; this approach is based on cross-validation and is widely used in machine learning [36]. This approach effectively reduces the training data size in the training set, so there is still room for research on generating new algorithms that allows setting this parameter more efficiently. Although we fitted the BRRQ models with smaller sample sizes in comparison with BRR models, the prediction accuracy of the proposed model was similar and, in some cases, better than the accuracy of the competing models. Nascimento et al. [10] compared the accuracy of the regularized quantile regression model against Bayesian LASSO with simulated data that had a heritability of 0.25 and reported gains in terms of prediction accuracy (Pearson's correlation between true and predicted breeding values) that ranged from 55% to 86%. Nascimento et al. [12] compared the predictive ability of Regularized QR models against several Bayesian linear models using flowering time in common bean and reported gains also in terms of prediction accuracy (Pearson's correlation between observed

Table 4 – Average of Pearson's correlation and standard deviation (in parentheses) between observed and predicted values in the testing dataset from cross-validation analysis for days to heading.

Predictors*	Model			
	BRQR θ = 0.25	BRQR θ = 0.50	BRQR θ = 0.75	BRR
Agua Fria				
A	**0.5548** (0.1602)	0.5548 (0.1629)	0.5195 (0.1702)	0.5463 (0.1690)
M	0.4902 (0.2014)	**0.4920** (0.1958)	0.4492 (0.1656)	0.4794 (0.1786)
A + *M*	0.5506 (0.1682)	0.5497 (0.1665)	0.5359 (0.1725)	**0.5578** (0.1661)
Cd. Obregon				
A	**0.6527** (0.1491)	0.6512 (0.1503)	0.6371 (0.1533)	0.6443 (0.1487)
M	0.6417 (0.1286)	**0.6477** (0.1331)	0.6437 (0.1357)	0.6350 (0.1365)
A + *M*	0.6589 (0.1415)	**0.6638** (0.1392)	0.6542 (0.1439)	0.6527 (0.1348)

* **A**, additive relationship matrix derived from pedigree; **M**, markers; **A** + **M**, additive relationship matrix derived from pedigree and markers included jointly. The predictions were obtained after fitting the BRQR and BRR models. The average is across 50 random partitions with 20% of observations in the testing set. The largest values of the correlations are in bold face.

and predicted phenotypes) and reported gains that ranged from 12% to 25%. In our case, the gains in prediction accuracy (when present) are much smaller.

Nowadays, GBLUP and ABLUP models are widely used in GS, and there are many software packages that can fit these models. On the other hand, only a few software packages in the public domain or that were developed commercially are able to fit robust regression models. In the case of BRQR, to our knowledge, BayesQR R package [15] contains the most recent implementations of algorithms to fit this family of models; unfortunately, the BayesQR was not designed to deal with high dimensional data and is not able to fit the model with markers and a relationship matrix derived from pedigree jointly, although for moderate size datasets, other software packages can be used, for example, JAGS [37] or Stan [38].

The computations required to fit BRQR are more complex than those required for fitting, for example, BRR or equivalently G-BLUP. Indeed, our implementation of the Gibbs algorithm without any special optimization or fine tuning to fit BRQR is much slower than the standard Ridge Regression in the BGLR package [28]. For example, when completing 1000 cycles of the Gibbs sampler and fitting the BRQR *M* using the wheat dataset (described previously) and implementing the Gibbs sampler, it took ~35 s on a laptop computer with an intel i7 processor @ 2.8 GHz and 16 Gb of RAM memory. On the other hand, when fitting the BRR in the BGLR package, it took only ~1.5 s.

Here we studied how to include markers and a relationship matrix derived from pedigree jointly. In the case of markers, several empirical studies have shown that Reproducing Kernel Hilbert Spaces Methods with Kernel averaging (RKHS) [39], when evaluated in cross-validation settings, exhibit better predictive ability that GBLUP, so a natural step will be to extend BRQR to include Reproducing Kernels. Another topic of interest would be to extend the BRQR model to include multi-traits and multi-environments, and especially, how to include genotype × environment interaction in the BRQR.

Author contributions

Paulino Pérez-Rodríguez, had the initial idea, ran the analyses, and wrote the first version of the article. Osval A. Montesinos-López, Abelardo Montesinos-López, and José Crossa read and wrote several other versions of the article, improved the writing, and checked the tables and references.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

The maize and wheat data set used in this study comes from the Drought Tolerance Maize for Africa Project and from CIMMYT's Global Wheat Program. We are thankful to everyone who generated the data used in this article.

Appendix A. Supplementary materials

Supplementary materials for this article can be found online at https://doi.org/10.1016/j.cj.2020.04.009.

REFERENCES

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

[2] J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J.M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski, R.K. Varshney, Genomic selection in plant breeding: methods, models, and perspectives, Trends Plant Sci. 22 (2017) 961–975.

[3] K. Stock, R. Reents, Genomic selection: status in different species and challenges for breeding, Reprod. Domest. Anim. 48 (2013) 2–10.

[4] P. Juliana, R.P. Singh, J. Poland, S. Mondal, J. Crossa, O.A. Montesinos-López, S. Dreisigacker, P. Pérez-Rodríguez, J. Huerta-Espino, L. Crespo-Herrera, V. Govindan, Prospects and challenges of applied genomic selection-a new paradigm in breeding for grain yield in bread wheat, Plant Genome 11 (2018) 1–17.

[5] B.C. Arnold, R.J. Beaver, Hidden truncation models, Shankhya, Indian J. Stat. 62 (2000) 23–35.

[6] A. Montesinos-López, O.A. Montesinos-López, E.R. Villa-Diharce, D. Gianola, J. Crossa, A robust Bayesian genome-based median regression model, Theor. Appl. Genet. 132 (2019) 1587–1606.

[7] D. Gianola, A. Cecchinato, H. Naya, C.C. Schön, Prediction of complex traits: robust alternatives to best linear unbiased prediction, Front. Genet. 9 (2018) 195.

[8] G.E.P. Box, D.R. Cox, An analysis of transformations, J. R. Stat. Soc. Ser. B 26 (1964) 211–252.

[9] L. Varona, N. Ibañez-Escriche, R. Quintanilla, J.L. Noguera, J. Casellas, Bayesian analysis of quantitative traits using skewed distributions, Genet. Res. 90 (2008) 179–190.

[10] M. Nascimento, F.F. e Silva, M.D.V. de Resende, C.D. Cruz, A.C. C. Nascimento, J.M.S. Viana, C.F. Azevedo, L.M.A. Barroso, Regularized quantile regression applied to genome-enabled prediction of quantitative traits, Genet. Mol. Res. 16 (2017) 1.

[11] P. Pérez-Rodríguez, R. Acosta-Pech, S. Pérez-Elizalde, C.V. Cruz, J.S. Espinosa, J. Crossa, A Bayesian genomic regression model with skew normal random errors, G3-Genes Genomes Genet. 8 (2018) 1771–1785.

[12] A.C. Nascimento, M. Nascimento, C. Azevedo, F. Silva, L. Barili, N. Vale, J. Carneiro, C. Cruz, P.C. Carneiro, N. Serão, Quantile regression applied to genome-enabled prediction of traits related to flowering time in the common bean, Agronomy 9 (2019) 796.

[13] K. Yu, R.A. Moyeed, Bayesian quantile regression, Stat. Probab. Lett. 54 (2001) 437–447.

[14] Q. Li, R. Xi, N. Lin, Bayesian regularized quantile regression, Bayesian Anal. 5 (2010) 533–556.

[15] D.F. Benoit, D. van den Poel, bayesQR: a Bayesian approach to quantile regression, J. Stat. Softw. 76 (2017) https://doi.org/10.18637/jss.v076.i07.

[16] R. Koenker, G. Bassett, Regression quantiles, Econometrica 46 (1978) 33–50.

[17] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B. 58 (1996) 267–288.

[18] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics 193 (2013) 327–345.

[19] S. Chib, E. Greenberg, Understanding the Metropolis-Hastings algorithm, Am. Stat. 49 (1995) 327.

[20] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.

[21] C. Reed, K. Yu, H. Kozumi, G. Kobayashi, Efficient Gibbs Sampling for Bayesian Quantile Regression, Technical report, Brunel University, UK, 2009.

[22] H. Kozumi, G. Kobayashi, Gibbs sampling methods for Bayesian quantile regression, J. Stat. Comput. Simul. 81 (2011) 1565–1578.

[23] R. Alhamzawi, A. Alhamzawi, H.T. Mohammad Ali, New Gibbs sampling methods for Bayesian regularized quantile regression, Comput. Biol. Med. 110 (2019) 52–65.

[24] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (1984) 721–741.

[25] G. Casella, E.I. George, Explaining the Gibbs sampler, Am. Stat. 46 (1992) 167.

[26] M. Nascimento, A.C.C. Nascimento, F.F. e Silva, L.D. Barili, N. M. do Vale, J.E. Carneiro, C.D. Cruz, P.C.S. Carneiro, N.V.L. Serão, Quantile regression for genome-wide association study of flowering time-related traits in common bean, PLoS One 13 (2018), e0190303.

[27] D. Sorensen, D. Gianola, Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics, Springer-Verlag, New York, USA, 2002.

[28] P. Pérez, G. de los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics 198 (2014) 483–495.

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

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

[31] R Core Team, R: a language and environment for statistical computing, https://www.R-project.org/ 2019.

[32] A. Pewsey, Problems of inference for Azzalini's skewnormal distribution, J. Appl. Stat. 27 (2000) 859–870.

[33] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, The deviance information criterion: 12 years on, J. R. Stat. Soc. B. 76 (2014) 485–493.

[34] J. Crossa, P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, C. Magorokosho, Genomic selection and prediction in plant breeding, J. Crop Improv. 25 (2011) 239–261.

[35] P. Pérez-Rodríguez, D. Gianola, J.M. González-Camacho, J. Crossa, Y. Manès, S. Dreisigacker, Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat, G3-Genes Genomes Genet. 2 (2012) 1595–1605.

[36] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, Germany, 2006.

[37] M. Plummer, JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling, in: K. Hornik, F. Leisch, A. Zeileis (Eds.),Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20–22, Vienna, Austria, 2003.

[38] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: a probabilistic programming language, J. Stat. Softw. 76 (2017).

[39] G. de los Campos, D. Gianola, G.J.M. Rosa, Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation, J. Anim. Sci. 87 (2009) 1883–1887.