Chapter 10

Bayesian Modeling in Genetics and Genomics

Hafedh Ben Zaabza, Abderrahmen Ben Gara and Boulbaba Rekik

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70167

Abstract

This chapter provides a critical review of statistical methods applied in animal and plant breeding programs, especially Bayesian methods. Classical and Bayesian procedures are presented in pedigree-based and marker-based models. The flexibility of the Bayesian approaches and their high accuracy of prediction of the breeding values are illustrated. We show a tendency of the superiority of Bayesian methods over best linear unbiased prediction (BLUP) in accuracy of selection, but some difficulties on elicitation of some complex prior distributions are investigated. Genetic models including marker and pedigree information are more accurate than statistical models based on markers or pedigree alone.

Keywords: accuracy of prediction, breeding value, Bayesian methods, BLUP, pedigree, markers

1. Introduction

Quantitative genetics result from the (connection) combination of statistics and the principles of animal and plant breeding. In quantitative genetics, selection for economically important traits refers to use of phenotypic values of the individual and pedigree information. Genomic is based on the use of dense markers through the whole genome to predict the breeding value of the individuals [1]. Linear models (univariate and multivariate) are of fundamental importance in applied and theoretical quantitative genetics [2]. In animal breeding, two major methods were particularly applied, restricted maximum likelihood (REML) and Bayesian methods. REML has emerged as the method of choice in animal breeding for variance component estimation [3]. Bayesian analysis is gaining popularity because of its more comprehensive assumptions than those of classical approaches and its flexibility in

resolving a wide range of biological problems [4, 5]. In the Bayesian approach, the idea is to combine what is known about the statistical ensemble before the data are observed (prior probability distributions) with the information coming from the data, to obtain a posterior distribution from which inferences are made using the standard probability calculus techniques [2, 6]. In recent years, Bayesian methods were broadly used to solve many of the difficulties faced by conventional statistical methods and extend the applicability of statistics on animal and plant breeding data [7]. Furthermore, Markov chain Monte Carlo (MCMC) has an important impact in applied statistics, especially from Bayesian perspective for the estimation of genetic parameters in the linear mixed effect model [2, 5]. The specific objective of this chapter was to illustrate applications of Bayesian inference in quantitative genetics and genomics. First, Bayesian models in the quantitative genetics theory are examined. Second, and in the context of the genomic selection, we presented the details of statistical modeling, using BLUP and Bayesian analyses. Third, a critical review with a focus on the prior distributions is illustrated. Finally, genomic predictions from several methods used in many countries are discussed.

2. A brief introduction to Bayesian analyses

In Bayesian inference, the idea is to combine what is known about the statistical ensemble before the data are observed (prior probability distributions) with the information coming from the data, to obtain a posterior distribution from which inferences are made using the standard probability calculus techniques.

$$P(\theta/y)\alpha \;\; \mathrm{P(y}/\theta)\mathrm{P}(\theta) \tag{1}$$

P(θ) is the prior distribution, which reflects the relative uncertainty about the possible values of θ before the data are seen. P(y/θ) is the likelihood function of observing the data given the parameter which represents the contribution of y to knowledge about the parameter θ. P(θ/y) is the posterior distribution of the parameter θ given the previous information on the data.

3. Bayesian analyses of linear models

3.1. The mixed linear model

The mixed linear model is of great importance in genetics and is one of the most used statistical models. Arguably, variance components and genetic parameters are important because they give an indication of the ability of species to respond to selection and thus the potential of that species to evolve. Mixed linear model is the simplest method for estimating the variance components for quantitative traits in population. In the "frequentist" view, mixed linear model is one included linearly the fixed and random effects. In the Bayesian context, there is no distinction between fixed and random effects. Detailed Bayesian analyses of models with two or more component variances will be discussed.

3.1.1. The univariate linear additive genetic model

The mixed linear model is one that includes fixed and random effects.

Consider the linear model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e} \tag{2}$$

y is a **n**×**1** vector of records on a trait; **β** is the vector of fixed effects affecting records; *a* is the vector of additive genetic effects; **e** is a vector of residual effects. **X** and **Z** are incidence matrices relating records to fixed effects and additive genetic effects, respectively. Data are assumed to be generated from the following distribution:

$$\mathbf{y}|\boldsymbol{\beta}, a, \sigma_e^2 \sim N(\mathbf{Xb} + \mathbf{Za}, \mathbf{I}\sigma_e^2)$$
$$\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$$

where, **I** is an identity matrix of order **n**×**n** and σ_e^2 is the residual variance. Independence of various effects was assumed for the sake of simplicity in implementation. We assume a genetic model in which genes act additively within and between loci, and there are effectively an infinite number of loci. Under this infinitesimal model, and assuming further initial Hardy-Weinberg and linkage equilibrium, the distribution of additive genetic values conditional on the additive genetic covariance is multivariate normal.

$$\mathbf{a}|\mathbf{A}, \sigma_a^2 \sim N(0, \mathbf{A}\sigma_a^2)$$

where **A** is the numerator relationship matrix of order **q**×**q; β** is assumed to have a uniform distribution with bounds $\boldsymbol{\beta}_{\mathbf{min}}$ and $\boldsymbol{\beta}_{\mathbf{max}}$.

$$\mathbf{P}(\sigma_i^2|\nu_i, S_i^2) \sim (\sigma_i^2)^{\left(\left(\frac{\nu_i}{2}+1\right)\right)} \exp\left(-\frac{\nu_i S_i^2}{2\sigma_i^2}\right), \quad (i = a, e)$$

where v_e, S_e^2 and v_a, S_a^2 are interpreted as degrees of belief and a priori values for residual and additive genetic covariances. Posterior conditional distributions derived from the likelihood and the prior distributions for these parameters are,

$\mathbf{b_i}|\ \mathbf{b_{-i}}, \mathbf{a}, \sigma_a^2, \sigma_e^2, \mathbf{y} \sim \mathrm{N}(\hat{b}_i, (x_i'x_i)^{-1}\sigma_e^2)$, with $(\boldsymbol{x'_i x_i})$ is the $\boldsymbol{i}$th element of the diagonal of $X'X$

3.1.2. The univariate linear additive genetic model with permanent and genetic group effects

The model equation [8] used to estimate genetic parameters and genetic breeding value for milk yield was as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{ZQg} + \mathbf{Wp} + \mathbf{e} \tag{3}$$

where **y** is the vector of milk yield, **b** is the vector of fixed effects, **a** is the vector of additive genetic effects, **g** is the vector of genetic group effects**, p** is the vector of random permanent

environmental effects, and **e** is the vector of residual effects. **X, Z, W**, and **ZQ** are incidence matrices relating a record to fixed environmental effects in **b**, to a random animal effects in a, to a random permanent environment effects in **p**, and to genetic groups in **g**, respectively. g^* is the vector of genetic group effects, $\hat{a}$ is a vector of breeding values. **A** is the numerator relationship matrix. where $\hat{\mathbf{a}}^* = \mathrm{Q}\,\hat{\mathbf{g}} + \hat{\mathbf{a}}$.

The conditional distribution of observed yield is defined by:

$$\mathbf{y}|\mathbf{b},\ \mathbf{p},\ \mathbf{a}*, \sigma_e^2 \sim N(\mathbf{Xb}\ +\ \mathbf{Za}* + \mathbf{Wp}, \mathbf{I}\sigma_e^2)$$

with the assumption of P(**b**) being a constant; $\boldsymbol{a}^*|\boldsymbol{A}^*, \sigma_a^2 \sim \mathrm{N}(\mathbf{Qg}, \mathbf{A}^*\sigma_{\mathbf{a}}^2)$;

$$\mathbf{p}|\sigma_{\mathbf{p}}^2 \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathbf{p}}^2);\ \text{and}\ \mathbf{P}(\sigma_{\mathbf{i}}^2|\nu_{\mathbf{i}}, \mathbf{S}_{\mathbf{i}}^2) \sim (\sigma_{\mathbf{i}}^2)^{((\frac{\nu_{\mathbf{i}}}{2}+1))} \exp\left(-\frac{\nu_{\mathrm{i}}\mathrm{S}_{\mathrm{i}}^2}{2\sigma_{\mathrm{i}}^2}\right)$$

where S_i^2 are prior values for the variances, $\chi_{\nu_i}^{-2}$ are inverted chi-square distributions, and ν_i are degrees of freedom of parameters.

3.1.2.1. Management and environmental effects

The distribution of a fixed effect is:

$$\mathbf{b_i}|\mathbf{b_{-i}}, \mathbf{a}^*, \sigma_a^2, \sigma_p^2, \sigma_e^2, \mathbf{y} \sim \mathrm{N}(\hat{\boldsymbol{b}}_i, (\mathbf{x}'_i\mathbf{x}_i)^{-1}\sigma_{\mathbf{e}}^2)$$

with $(\boldsymbol{x}'_{\boldsymbol{i}}\,\boldsymbol{x}_{\boldsymbol{i}})\ \hat{b}_i = x'y - x_i'x_{-i}'b_{-i} - x_i'w_p - x_i'za^*$,

where $(\boldsymbol{x}'_{\boldsymbol{i}}\,\boldsymbol{x}_{\boldsymbol{i}})$ is the $\boldsymbol{i}$th element of the diagonal of X′X

3.1.2.2. Permanent environmental effects

The distribution of a permanent effect is:

$$p_i|b_i, p_{-i}, a^*, \sigma_a^2, \sigma_p^2, \sigma_{e'}^2, y \sim \mathrm{N}(\hat{p}_{i'}, (w'_iw_i + \delta)^{-1}\sigma_e^2)$$

with $(w_i'w_i + \delta)\hat{p}_i = w'_i\,y - w'_i\,Xb - (w_iW_{-i} + \delta)p_{-i} - w_iz_a{}^*$,

where $w'_i\,\boldsymbol{w_i}$ is the ith element of the diagonal of $W'W$.

3.1.2.3. Breeding values

The distribution of a breeding value is:

$$\mathbf{a_i^*}|\mathbf{b}, \mathbf{p_{-i}}, \mathbf{a_{-i}^*}, \sigma_{\mathbf{a}}^2, \sigma_{\mathbf{p}}^2, \sigma_{\mathbf{e}}^2, \mathbf{y} \sim \mathrm{N}(\mathbf{a_i^*}(\mathbf{z'_iz_i} + \mathbf{A}_{\mathbf{i},\mathbf{i}}^{*-1}\alpha^{-1})\sigma_{\mathbf{e}}^2)$$

with $(z_i'z_i + A_{i,i}^{*-1}\alpha)\hat{a}_i = z_i'y - z_i'Xb - z_i'W_P - A_{i,i}^{*-1}\alpha a_{-i}^*$,

where $z'_i\,\boldsymbol{z_i}$ is the $\boldsymbol{i}$ th element of the diagonal of $Z'Z$.

3.1.2.4. Variance components

The additive genetic variance is defined by

$$\sigma_a^2|b, p, a^*, \sigma_p^2, \sigma_e^2, \mathbf{y} \sim \tilde{V}_a \tilde{S}_a^2 \chi_{\tilde{v}_a}^{-2}$$

with $\tilde{V}_a = n_a + V_a$, $\tilde{S}_a^2 = (a*' A^{*-1} a^* + V_a S_a^2)/\tilde{V}_a$, and $\boldsymbol{n_p}$ is the number of animals being evaluated.

The variance of permanent environmental effects is given by:

$$\sigma_p^2, |b, p, a^*, \sigma_a^2, \sigma_e^2, \mathbf{y} \sim \tilde{V}_p \tilde{S}_p^2 \chi_{\tilde{v}_p}^{-2}$$

with $\tilde{V}_p = n_p + V_p$, $\tilde{S}_p^2 = (p'p + V_p S_p^2)/\tilde{V}_p$, and n_p is the number of animals being evaluated.

Residual variance:

$$\sigma_e^2|b, p, a^*, \sigma_a^2, \sigma_p^2, \mathbf{y} \sim \tilde{V}_e \tilde{S}_e^2 \chi_{\tilde{v}_e}^{-2}$$

with $\tilde{V}_e = n_e + V_e$,

$$\tilde{S}_e^2 = [(y - Xb - Wp - Za*)'(y - Xb - Wp - Za*) + V_e S_e^2]/\tilde{V}_e,$$

and n_e is the total number of records.

Comparing genetic value predictions based on polygenic model in Tunisian Holstein Population using BLUP and Bayesian analyses, Ref. [8] reported that the rankings of animals with Bayesian methods are similar to those obtained by BLUP method. Spearman's rank correlation between genetic values estimated from Bayesian procedures and genetic values estimated from BLUP methods were high (0.99). Again, Bayesian and best linear unbiased estimator (BLUE) solutions of fixed effects (month of calving, herd-year, and age-parity) showed the same patterns. The same result is reported by Ref. [9]. However, Ref. [8] illustrated different correlation estimates between two methods (Bayesian and BLUP) for cow's and bull's breeding value.

4. Genomic selection

A massive quantity of genomic data is now available in animal and plant breeding with the revolutionary development in sequencing and genotyping. The cost of genotyping is dramatically reduced. Consequently, practices of genomic selection are nowadays possible with the high number of single nucleotide polymorphism (SNP) markers available. Therefore, it is feasible to perform analysis of the genome at a level that was not possible before [10–13]. The concept of genomic selection was introduced by Ref. [1]. The latter suggested that a set of markers covering the whole genome explain the all genetic variances and each marker is likely to be associated with a quantitative trait locus (QTL), and each QTL is in linkage disequilibrium with the

markers. The number of effects per QTL to be estimated is very small. The estimated effects of all markers are summed in order to obtain the genetic value of the individual. Using simulation, Ref. [1] showed in simulation that with a high-density SNP marker, it is possible to predict the breeding value with an accuracy of 0.85 (where accuracy is the correlation between the estimated breeding value and true breeding value). The challenge in genomic evaluation is to find the best prediction method to obtain accurate genetic values of candidates. Many genomic evaluation methods have been proposed [14, 15]. The main objective of this section is to compare Bayesian methods to other methods used in genomic selection based on their predictive abilities. The study reported by Ref. [1] was considered an influential paper on dairy cattle breeding programs. First, the methods suggested correspond well to the data structures where the number of SNPs substantially exceeds the number of observations. Second, the methods of Ref. [1] constitute a logical evolution of the BLUP methodology, which is the reference method in animal genetics by considering specific variances of SNPs in the different loci. Third, the Bayesian approaches used in Ref. [1] that take into account unknown effects (measuring prior uncertainty) in a model, and combined with the ability of the Monte Carlo Markov chain, can be used in the majority of parametric statistical models.

4.1. Genomic BLUP (GBLUP)

The GBLUP method assumes that effects of all SNPs are sampled from the same normal distribution; the effects of all markers are assumed to be small with equal variance. Genomic BLUP was defined by the model:

$$y = 1\mu + Zg + e \tag{4}$$

where **y** is the data vector; μ is the overall mean; 1 is a vector of **n** ones; **Z** is a matrix of incidence, allocating records to the markers' effects; **g** is a vector of SNP effects assumed to be normally distributed $g \sim \mathrm{N}(0, G\sigma_g^2)$, where σ_g^2 is the additive genetic variance and **G** is the genomic relationship matrix; **e** is the vector of normal error, $e \sim \mathrm{N}(0, \sigma_e^2)$ where σ_e^2 is the error variance. The genomic relationship matrix was defined as $G = \frac{X'X}{\sum_{i=1}^{m} p_i(1-p_i)}$, where X is matrix for specified SNP genotype coefficient at each locus, p_i is the rare allele frequency for $\mathrm{SNP_i}$.

4.2. Bayesian approaches

In Bayesian estimation, the information from the data is combined with the information from the prior distribution of the variances of the markers. Several Bayesian statistical analyses have been used in genomic evaluation, which differ in the hypotheses of distributions of marker effects. At the level of the modeling of the variances of the effects of the markers, Meuwissen et al. [1] proposed different distributions a priori between the Bayes A and Bayes B methods.

4.2.1. Bayes A

Bayes A method assumes that variance of marker effects differ among loci (e.g., $\sigma_{g_j}^2$ is different across the **j**) [16]. The variances are modeled according to the scaled inverted chi-square distribution: The a priori distribution of the variances of the SNP effects is written:

$P(\sigma^2_{g_j}) \sim \chi^{-2}(\nu, \mathrm{S})$, where S is the scale parameter and ν is the number of degrees of freedom. This has the advantage, if we consider a normal distribution of the data, to lead to an a posteriori conditional distribution of χ^{-2}.

$$P(\sigma^2_{g_j}|g_j) \sim \chi^{-2}(\nu + n_j, \mathrm{S} + \mathrm{g'_j}g_j),$$

where, $\mathrm{n_j}$ is the number of marker effects at segment j. The posterior distribution combines both the information provided by the data and the a priori distribution.

4.2.2. Bayes B

In a genomic evaluation context, Bayes B method [1, 17] assumes different variances of SNP effects, with many SNP contribute per zero effects, and a few contribute per a large effects on the trait. Meuwissen et al. [1] propose a model in which a proportion π (arbitrarily fixed at 0.95) of the markers having zero effect. The a priori distribution of the variances of the effects to the markers is then written:

$\sigma^2_g = 0$ with a probability π, $P(\sigma^2_{g_j}) \sim \chi^{-2}(\nu, \mathrm{S})$ with a probability $(1 - \pi)$, Gibbs sampling cannot be used to estimate the effects and variances of the Bayes B model because of the high probability on some markers of being of zero variance. We therefore use a Metropolis-Hastings algorithm which allows the simultaneous estimation of $\sigma^2_{g_j}$ and g_j. On the basis of the results of Ref. [1] and many subsequent works, the Bayes B method is often considered the "benchmark" in terms of genomic prediction efficiency, but it is extremely costly in computational time. However, Meuwissen [18] propose an alternative to the Bayes B method which relies on a fast algorithm.

4.2.3. Bayesian lasso

Legarra et al. [19] proposed a model of Bayesian lasso (BL) with different variances for residual and SNP effects which they termed BL2Var. It is therefore assumed that a large number of SNPs have an effect practically zero and that very few have large effects. Tibshirani [20] showed that the distribution of the lasso estimators can be written:

$$\mathrm{P}(\sigma^2_{g_j}|\sigma^2, \lambda) \sim \frac{\lambda}{2} \exp\left(-\lambda|g_j|\right)$$

He suggests that the lasso estimators can be interpreted as an a posteriori mode of a model in which the regression parameters would be independent and identically distributed according to a prior double exponential distribution. Park and Casella [21] propose to use a complete Bayesian approach by assuming an a priori distribution of regression coefficients such as:

$$\mathrm{P}(\sigma^2_{g_j}|\sigma^2, \lambda) \sim \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|g_j|\right)$$

where σ^2 represents the variance of residual effects of the model and the variance of the SNP effects. Applications of the Bayesian lasso to the genomic selection proposed by Refs. [22, 23] use the same variance σ^2 to model both the distribution of effects of SNPs and residuals. De los

Campos et al. [22] showed that the Bayesian lasso is close in terms of precision of prediction to the Bayes B method but with a significant reduction in the complexity of the calculations. In addition, these authors suggested using Bayesian lasso against the large number of markers included in regression models, which is typically larger than the number of records.

4.2.4. The Bayes C method

Bayesian methods such as Bayes A and Bayes B [1] have been widely used for genomic evaluation. Similar methods exist, with similar performances, developed in order to reduce computation times and to simplify statistical modeling. The Bayes C method [24] differs from Bayes B by assuming the variance associated with SNPs common to all markers. In Bayes C, as in Bayes B, the probability π that an SNP has a nonzero effect is assumed to be known. The model is similar to the Bayes B model but for a homogeneous variance of effects on all loci: $\sigma_g^2 = 0$ with a probability $1 - \pi$;$(\sigma_g^2) \sim \chi^{-2}(\nu,\ \mathrm{S})$. The main problem with the Bayes C method is that SNPs with a nonzero effect is assumed to be known. With the Bayes A method, the parameter π is equal to 1, which implies that all the markers have an effect. For the Bayes B method, π is strictly less than 1 in order to take into account the hypothesis that some SNPs may have a zero effect but is fixed arbitrarily while the intensity of the selection of variables is controlled by this parameter. Habier et al. [25] propose to modify the Bayes C method by estimating the parameter π: the parameter π is assumed to be unknown. Thus, the a priori distribution of π becomes uniform over [0, 1]. SNP modeling is the same as with Bayes C. $\mathrm{P}(g_j|\pi, \sigma_{\mathrm{g}}^2) = 0$ with a probability $1 - \pi$; $\mathrm{P}(g_j|\pi, \sigma_{\mathrm{g}}^2) \sim N(0, \sigma_{\mathrm{g}}^2)$ where $P(\sigma_g^2) \sim \chi^{-2}(\nu, \mathrm{S})$ with a probability π. The various parameters of this model are estimated by MCMC methods, Markov Chain Monte Carlo [6, 26] as proposed by Ref. [25]. It is written as a function of the additive genetic variance σ_a^2. $\sigma_g^2 = \frac{\sigma_a^2}{(1-\pi)\sum_{j=1}^{p} 2p_j(1-p_j)}$, where $\mathrm{p_j}$ is the allelic frequency of SNP j.

4.3. A critique

The extreme speed with which events are running handicaps the process of linking new development to extant theory, and the understanding of statistical models suggested up until now [27]. The latter authors criticize the theoretical and statistical concepts followed by Ref. [1] in three levels. The first is the connection between parameters (additive genetic variances with Bayesian view) from infinitesimal models with those from marker-based models. The second is the relationship between molecular marker genotypes and similarity between relatives. The third is the connection between infinitesimal genetic models and marker-based regression models. Gianola et al. [27] argued that the methods Bayes A and Bayes B proposed by Ref. [18] require specifying parameters. The latter used formulas for obtaining the variance of SNP effects, based on some knowledge of the additive genetic variance in the population. Their development begins on the assumption that the effects of the markers are fixed and in other development, they consider them as random without a clear demonstration. Meuwissen et al. [1] explained that affecting a priori a value $\sigma_g^2 = 0$ with a probability π means that the specific SNP does not have an effect on the trait. By contrast, Ref. [27] illustrated that a parameter

having zero variance does not obligatory imply that the parameter takes zero value. The parameter could have any value, but with certainty. Gianola et al. [27] suggested the use of a nonparametric method as developed by Refs. [22, 28] because these methods do not impose hypotheses about mode of inheritance as Bayesian A and Bayesian B methods.

5. Applications in genomics

Major dairy breeding countries are now using genomic evaluation [27]. Several results have been reported around the world. Several authors reported that the reliabilities of genomic estimated breeding values (GEBV) were substantially greater than breeding values from estimated breeding values (EBV) based on pedigree information [29]. The accuracy of selection was different between countries [12]. The accuracy was dependent on the size of reference population, the heritability of the trait studied, the statistical models and approaches used for prediction of genetic values for quantitative traits, and the method achieved to estimate the accuracy [12, 27, 29]. Ref. [14] found the reliability of GEBV bulls of the Canadian and American Holstein population. A genotyping of 39,416 molecular markers of 3576 Holstein bulls was used to establish the prediction equations.

The prediction methods contained a linear model, in which marker effects are assumed to be normal, and a nonlinear model with a heavier tailed prior distribution to account for major genes as described by [1]. VanRaden et al. [14] reported that the combination of the polygenic effects based on pedigree information with the genomic predictions can improve the reliability to 23% greater than the reliability of polygenic effects only. The same study showed that the nonlinear model had a little advantage in reliability over the linear model for all traits except for fat and protein percentages. Genomic breeding values of 25 traits in New Zealand dairy cattle were estimated by Ref. [30]. The reference population consisted of 4500 bulls genotyped using the BovineSNP50Beadchip, containing 44,146 SNPs. Harris and Johnson [31] reported an increase in accuracy was found by using Bayesian approaches compared to BLUP methods. In Ref. [31], genomic breeding values (GBVs) for young bulls with no daughter information had accuracies ranging from 50 to 67% for milk traits, live weight, fertility, somatic cell, and longevity, versus an average 34% for progeny test. Meuwissen et al. [1] compared least squares method with BLUP and two Bayesian methods (Bayesian A and Bayesian B). The latter authors estimated the effects of 50,000 marker haplotypes from a limited number of observations (2200). Using least squares method, it is not possible to estimate all effects simultaneously. For this reason, different steps have been adopted to incorporate the effects of markers. First, they performed regression on markers for every segment of 1 cm each. Second, they calculated a Log-likelihood, which assumed to be normal at every segment of chromosome. Third, they summed all segments corresponding to a likelihood peak into multiple regression models. Using BLUP analyses, Ref. [1] considered that all SNP effects were independent and identically distributed with a known variance. Bayes A method was as BLUP at the level of the data, but differs in the variance of the chromosome segments, which assumed to have an inverted chi-square distribution. A mixture prior distribution of genetic variances was used in Bayes B method. **Table 1** shows the accuracy of selection obtained by Ref. [1] from the GBLUP methods, the least squares regression and the

Methods	ρ	b
Least squares	0.318	0.285
GBLUP	0.732	0.896
Bayes A	0.798	0.827
Bayes B	0.848	0.946

Table 1. Comparing estimated versus the breeding value [1].

Bayes A and Bayes B approaches. The predictive abilities of the different methods are estimated by calculating the correlation (ρ) between true and estimated breeding values and the regression (b) of true on estimated breeding value.

The least squares method is the least efficient because it overestimates effects on QTL [32]. The Bayes B approach is the most accurate both in terms of correlation and regression. However, the regression coefficient obtained by the Bayesian methods was still less than 1, and probably due to the hypothesis of a priori distribution χ^{-2} for Bayes A and Bayes B being different from the simulated distribution of the variances. Goddard and Hayes [11] compared the correlation of 0.85 as reported by Ref. [1] to results obtained on real data by Refs. [14, 33, 34]. VanRaden et al. [14] produced a mean correlation over several characters of 0.71 from a reference population of more than 3500 bulls. Studies have shown the superiority of genomic evaluation [35] or marker-assisted selection in France [36] on classical infinitesimal model of quantitative genetics. Several authors have applied the first genomic evaluation methods described by Ref. [1] or their derived methods on real data. The Bayes A and Bayes B approaches have found results that are often similar or slightly superior to GBLUP in terms of accuracy of genetic value prediction for the Australian Holstein-Friesian cattle breed (+0.02 to +0.07 of correlation gain between predicted and observed values), for example [12] and New Zealand (+2% correlation gain, [31]). However, the GBLUP method required less computing time than the Bayes A method [32, 37]. Gredler et al. [38] demonstrated the superiority of the Bayes B method, in terms of the accuracy of genomic estimates, on a modified Bayes A method for integrating a polygenic effect [39]. Thus, although the Bayes B method seems slightly more efficient than the Bayes A method, numerous studies showed that the Bayes B method is not so much better in terms of accuracy of the genomic estimates than a GBLUP model [40]. Again, all researches indicate that the Bayesian approaches, which assume an a priori distribution of SNPs, increase the reliability of breeding values over traditional BLUP methods [1, 12, 14]. A common conclusion is that for most quantitative traits, the hypothesis of the traditional BLUP method, that all markers are associated with equal variances, is far from reality. By comparing the results obtained in the various populations around the world, clearly, the accuracies of GEBVs were greater than breeding values estimated from progeny test based on pedigree information. Several researches suggested combining the progeny test based on pedigree information with the breeding value from genomic to calculate the final GEBV [5, 25]. Accuracy based on modeling molecular marker and pedigree information was generally superior to that of the model including only genomic or pedigree information. Hayes et al. [12] reported that a main

advantage of using the both sources of information coming from polygenic breeding values and genomic information is that any QTL not detected by the marker effects may be detected by the progeny test based on pedigree information. A significant reduction in posterior mean of residual variance component was reported by Ref. [22] when pedigree and markers were considered jointly compared to pedigree-based model. In the same study, Spearman's rank correlation of estimated breeding value between model including marker information and pedigree-based model was close to 1.

6. Conclusion

Standard quantitative genetic model based on phenotypic and pedigree information has been very successful in term of genetic value prediction. Also, the availability of genome-wide dense markers leads researchers to be able to perform advanced genetic evaluation of quantitative traits with a high accuracy of prediction of genetic value. However, a main problem is how this information should be included into statistical genetic models. Bayesian MCMC methods appear to be convenient for genetic value prediction with a focus on the precision of the choice of prior distribution for the different parameters.

Author details

Hafedh Ben Zaabza[1]*, Abderrahmen Ben Gara[2] and Boulbaba Rekik[2]

*Address all correspondence to: hafedhbenzaabza@gmail.com

1 Institut National Agronomique, Tunis-Mahrajène, Tunisie

2 Département des productions animales, Ecole supérieure d'Agriculture de Mateur, Mateur, Tunisie

References

[1] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;**157**:1819-1829

[2] Sorenson D, Gianola D. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. 1st ed. New York: Springer-Verlag; 2002. p. 740

[3] Neumaier A, Groeneveld E. Restricted maximum likelihood estimation of covariances in sparse linear models. Genetics Selection Evolution. 1997;**30**(1):3-26

[4] Waldmann P. Easy and flexible Bayesian inference of quantitative genetic parameters. Evolution. 2009;**63**(6):1640-1643. DOI: 10.1111/j.1558-5646.2009

[5] Hallander J, Waldmann P, Chunkao W, Sillanpaa MJ. Bayesian inference of genetic parameters based on conditional decompositions of multivariate normal distributions. Genetics. 2010;**185**:645-654. DOI: 10.1534/genetics.110.114249

[6] Robert CP. Le choix bayésien Principes et pratique. 1st ed. Paris: Springer-Verlag France; 2006. p. 638

[7] Ben Zaabza H, Ben Gara A, Hammami H, Ferchichi MA, Rekik B. Estimation of variance components of milk, fat, and protein yields of Tunisian Holstein dairy cattle using Bayesian and REML methods. Archives Animal Breeding. 2016;**59**:243-248. DOI: 10.5194/aab-59-243-2016

[8] Ben Gara A, Rekik B, Bouallègue M. Genetic parameters and evaluation of the Tunisian dairy cattle population for milk yield by Bayesian and BLUP analyses. Livestock Science. 2006;**100**:142-149. DOI: 10.1016/j.livsci.2005.08.012

[9] Schenkel FS, Schaeffer LR, Boettcher PJ. Comparison between estimation of breeding values and fixed effects using Bayesian and empirical BLUP estimation under selection on parents and missing pedigree information. Genetic Selection Evolution. 2002;**34**:41-59. DOI: 10.1051/gse:2001003

[10] Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic value with semi-parametric procedures. Genetics. 2006;**173**(3):1761-1776. DOI: 10.1534/genetics.105.049510

[11] Goddard ME, Hayes BJ. Genomic selection. Journal of Animal Breeding and Genetics. 2007;**124**:323-330. DOI: 10.1111/j.1439-0388.2007

[12] Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science. 2009;**92**:433-443. DOI: 10.3168/jds.2008-1646

[13] Wittenburg D, Melzer N, Reinsch N. Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. BMC Genetics. 2011;**12**(74):14

[14] VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science. 2009;**92**:16-24. DOI: 10.3168/jds.2008-1514

[15] Colombani C, Croiseau P, Fritz S, Guillaume F, Legarra A, Ducrocq V, Robert-Granié C. A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. Journal of Dairy Science. 2012;**95**:2120-2131. DOI: 10.3168/jds.2011-4647

[16] Su G, Guldbrandtsen B, Gregersen VR, Lund MS. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. Journal of Dairy Science. 2010;**93**(3):1175-1183. DOI: 10.3168/jds.2009-2192

[17] Villumsen TM, Janss L, Lund MS. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics. 2009;**126**(1):3-13. DOI: 10.1111/j.1439-0388.2008

[18] Meuwissen THE. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genetics Selection Evolution. 2009;**41**:35. DOI: 10.1186/1297-9686-41-35

[19] Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S. Improved lasso for genomic selection. Genetics Research. 2011;**93**(1):77-87. DOI: 10.1017/S0016672310000534

[20] Tibshirani R. Regression shrinkage selection via the LASSO. Journal of the Royal Statistical Society Series B. 1996;**73**(3):273-282. DOI: 10.2307/41262671

[21] Park T, Casella G. The Bayesian lasso. Journal of the American Statistical Association. 2008;**103**(482)681-686. DOI: 10.1198/016214508000000337

[22] De los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009;**182**:375-385. DOI: 10.1534/genetics.109.101501

[23] Weigel KA, De los Campos G, González-Recio O, Naya H, Wu XL, Long N, et al. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science. 2009;**92**(10): 5248-5257. DOI: 10.3168/jds.2009-2092

[24] Kizilkaya K, Fernando RL, Garrick DJ. Genomic prediction of simulated multi-breed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of Animal Science. 2010;**88**(2):544-551. DOI: 10.2527/jas.2009-2064

[25] Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;**12**:12

[26] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics. 1953;**21**:1087-1092

[27] Gianola D, Manfredi E, Fernando RL. Additive genetic variability and the Bayesian alphabet. Genetics. 2009;**183**:347-363. DOI: 10.1534/genetics.109.103952

[28] Gianola D, Van Kam JBCHM. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;**178**(4):2289-2303. DOI: 10.1534/genetics.107.084285

[29] Su G, Madsen P, Nielsen US, Mäntysaari EA, Aamand GP, Christensen OF, et al. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science. 2012;**95**:909-917. DOI: 10.3168/jds.2011-4804

[30] Harris BL, Johnson DL, Spelman RJ. Genomic selection in New Zealand and the implications for national genetic evaluation. In: Proceeding Interbull Meeting; 2008; Canada. The 36th International Committee for Animal Recording (ICAR) Session, held June 16-20, in Niagara Falls; 2008

[31] Harris BL, Johnson DL. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science. 2009;**93**(3):1243-1252. DOI: 10.3168/jds.2009-2619

[32] Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HM. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution. 2009;**41**(56). DOI: 10.1186/1297-9686-41-56

[33] Legarra A, Misztal I. Technical note: Computing strategies in genome-wide selection. Journal of Dairy Science. 2008;**91**(1):360-366. DOI: 10.3168/jds.2007-0403

[34] González-Recio O, Gianola G, Rosa GJM, Weigel KA, Kranis A. Genome-assisted prediction of a quantitative trait measured in parents and progeny: Application to food conversion rate in chickens. Genetics Selection Evolution. 2009;**41**(3):10. DOI: 10.1186/1297-9686-41-3

[35] VanRaden P. Efficient methods to compute genomic predictions. Journal of Dairy Science. 2008;**91**(11):4414-4423. DOI: 10.3168/jds.2007-0980

[36] Boichard D, Fritz S, Rossignol MN, Bosher MY, Malafosse A, Colleau JJ. Implementation of marker-assisted selection in French dairy cattle. In: 7th World Congress on Genetics Applied to Livestock Production; 19-23 August 2002; Montpellier, France. 2002. Session 22. Exploitation of molecular information in animal breeding. Electronic communication 22-03. p. 4

[37] Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution. 2009;**41**(29):8. DOI: 10.1186/1297-9686-41-29

[38] Gredler B, Nirea KG, Solberg TR, Egger-Danner C, Meuwissen THE, Solkner J. Genomic selection in Fleckvieh/Simmental—First results. In: Proceedings of the Interbull Meeting; 21-24 August 2009; Interbull Bulletin, Barcelone, Espagne; 2009;**40**:209-213

[39] Hayes BJ. Genomic selection in the era of the $1000 genome sequence. In: Symposium Statistical Genetics of Livestock for the Post-Genomic Era; USA: Wisconsin-Madison, USA; 2009

[40] Habier DJ, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution. 2010;**42**(5). DOI: 10.1186/1297-9686-42-5