Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Flexibility of Bayesian LASSO under different genetic structure

L.O. Duitama[1], M.M. Farah[1], D.A. Garcia[1], R.K. Ono[1], L. Cavani[1],R. da Fonseca[1]
[1] Sao Paulo state University, (UNESP), Jaboticabal, Sao Paulo, Brazil.

ABSTRACT: To evaluate the flexibility of the Bayesian LASSO six scenarios were simulated considering a trait controlled by 145 or 725 quantitative trait loci (QTL), where the QTL effects were sampled from three distributions: Normal, Gamma and Uniform. The population size was 4,500 animals and each of genetic structures were simulated 30 times. The posterior mean for additive genetic variance and heritability were close to the real value. Correlations between estimated and true genomic breeding values (GBV) were high and close in all simulated scenarios, indicating that the structure of the trait does not affect the predictive ability of the model. The Bayesian LASSO showed similar performance in terms of genetic parameters estimation and genomic breeding values predictions in spite of the underlying assumptions of genetic structure.
Keywords: genomic selection; QTLs; markers

Introduction

With the advance of genomic methodologies, new statistical methods have been developed (Meuwissen et al. (2001); de los Campos et al. (2009); Misztal et al. (2009)), which enable analyze genomic information in genetic evaluations. The Bayesian methods, proposed by Meuwissen et al. (2001), provided a better adjust of genetic structure of the trait and showed the best results in the SNP effects estimation (Single Nucleotide Polymorphism) and higher correlations between predicted and true GBV (Genomic Breeding Value), when these approaches were compared with the Ridge Regression and GBLUP. Besides these methodologies, the Bayesian LASSO (Park and Casella, (2008); de los Campos et al. (2009), Legarra et al. (2010)) has been successfully applied to genomic enable prediction as reported by some authors (Cleveland et al. (2010); Pérez et al. (2010)). In summary, the LASSO does shrinkage of regression coefficients toward zero and variable selection. Therefore, SNPs which are not in regions near to QTLs, that affects the trait, have higher probability to be regressed toward zero. On the other hand, SNPs which are in linkage disequilibrium with QTLs, that are associated to phenotype, will be selected and assigned effects different from zero.

The goal of the Genomic Selection (GS) model is predict the GBV of animals that do not have phenotypic records using the available genotype information (SNPs). The best model is the one that maximizes the correlation between the estimated and true GBV. However, in real data the true GBV are unknown. Thus, it is common use simulation studies aiming to compare and evaluate proposed models.

The objective of this study was to evaluate the flexibility of the Bayesian LASSO in terms of genetic parameters estimation and genomic breeding value predictions, when data with different genetic structures were simulated.

Materials and Methods

Simulation: The data were simulated using the QMSim software, developed by Sargolzaei and Schenkel (2009). The populations were simulated in two steps: i) A base population was defined and 50 generations were used to establish the balance between mutation and genetic drift; ii) The population structure, which was defined by 10 overlapping generations of 450 animals randomly mated. The total of animals in each simulated population was 4,500 and those from the last 4 generations were used in the analysis.

The genome was simulated considering 29 chromosomes of 100 cM of length and 29,000 biallelic SNPs, that had equal allelic frequency and randomly distributed throughout the genome. The genetic structures were simulated combining QLTs number, 145 or 725, and QTL effects, which were drawn from three distributions: Normal, Gamma (shape =0.4), or Uniform. The true heritability assumed was 0.4 and each genetic structure was simulated 30 times.

Analysis: The multiple regression model can be described as:

$$Y_i = B_0 + \sum_{j=i}^{p} x_{ij} B_j + e$$

where Y_i = phenotype of the ith genotyped animal in population, B_0 = intercept, X_{ij} = genotype marker of the jth marker for the ith animal, B_j = effect of the jth marker, e = residual effect, wich was assumed follow a normal distribution $N(0, \sigma^2_e)$. The marker genotypes X_{ij} were expressed as 0 for the first homozygous, 1 for the heterozygous, and 2 for the second homozygous.

The improved Bayesian LASSO (Legarra, et al. (2011)) was applied for variance components estimation and genomic breeding value prediction. This approach assumes a double exponential distribution as marginal

Table 1: Mean and highest probability density interval (HPD95) to addictive genetic variance and the heritability for the simulated scenarios

Scenario	G145	G725	N145	N725	U145	U725
σ^2_a	0.390 [0.274-0.517]	0.387 [0.256-0.559]	0.368 [0.239-0.520]	0.375 [0.245-0.526]	0.379 [0.262-0.541]	0.367 [0.235-0.518]
h^2	0.391 [0.295-0.483]	0.385 [0.279-0.511]	0.373 [0.265-0.481]	0.377 [0.267-0.485]	0.382 [0.286-0.492]	0.369 [0.262-0.473]

G, N e U= QTL effect provided of Gamma, Normal e Uniform distributions
145 e 725 number of QTLs that controlled the trait
σ^2_a=addictive genetic variance
h^2 = heritabilityww.

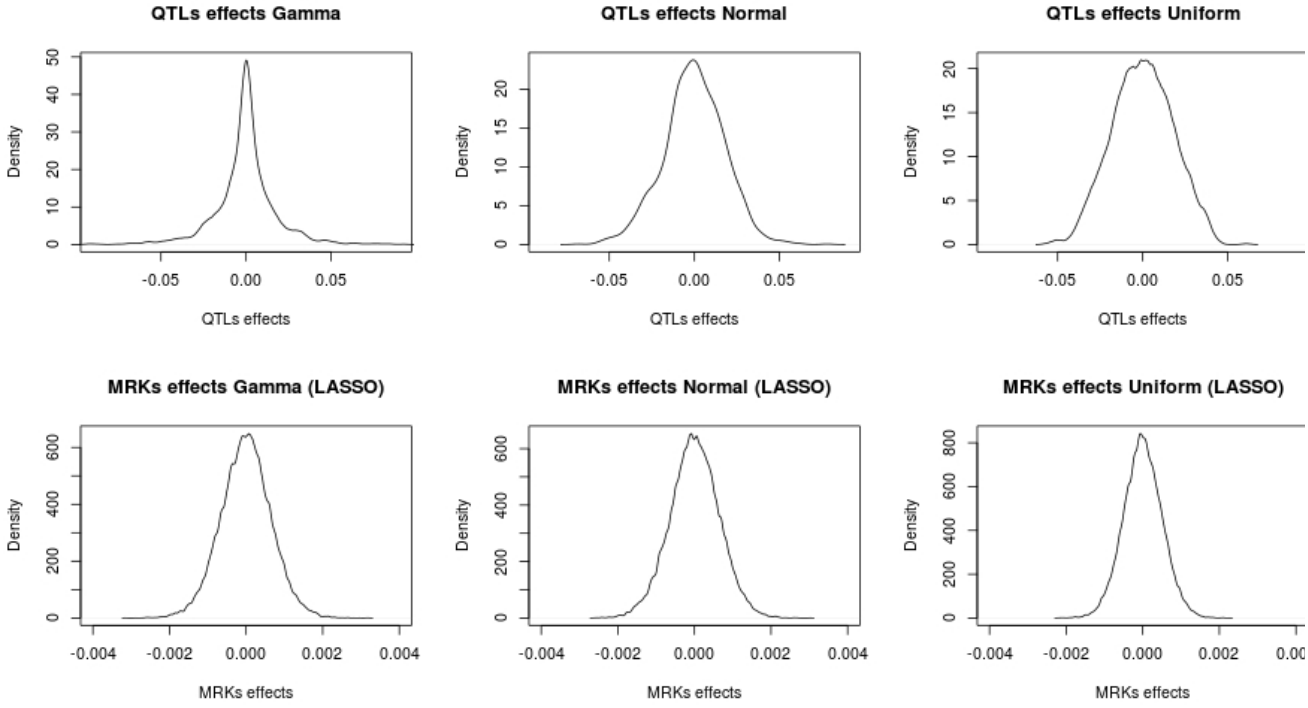

Figure 1. Distribution of the QTLs effect (from Gamma, Normal and Uniform distribution) and markers effects estimated by Bayesian LASSO.

priori for marker effects. The GS3 software was used to run a single chain of 200,000 draws and 30,000 samples were discarded as burn-in.

The comparison among simulated scenarios was based on highest probability density interval (HPD95), for the additive genetic variance (σ^2_a) and heritability (h^2), and Pearson correlations between predicted and true GBVs.

Results and Discussion

The first three graphs of Figure 1 show the distributions of the QTLs effects for Gamma, Normal, and Uniform, respectively. As expected, the simulated QTL effects using a Gamma distribution showed the greatest number of effects near to zero and highest proportion of markers with large effect, it can be seen by the thickest tails of QTLs distribution. On the other hand, the Uniform distribution showed the lowest number of markers which the effects were close to zero.

In the last three graphs of Figure 1, the estimated marker effects (29,000) are shown. It can be seen that independent from the density applied to simulate QTL effects it did not influence the marker effects estimation. Moreover, the SNP effects are smaller than the QTL effect, indicating that the effects of QTLs are distributed among the SNPs that were in linkage disequilibrium with the QTL.

The posterior means of genetic variance and heritability were very close to the true values (0.40) and these true genetic parameters belonged to HPD95 estimates in all genetic structures (Table 1). Additionally, the Bayesian LASSO prediction ability were similar, because high correlation between predicted and true GBVs were observed in the six simulations scenarios (Table 2).

Table 2: Mean and standard error of the correlation between predicted and true GBVs for the simulated scenarios

Sce.	G145	G725	N145	N725	U145	U725
$r_{u\hat{u}}$	0.7780 (0.005)	0.7743 (0.005)	0.7707 (0.005)	0.7725 (0.005)	0.7787 (0.004)	0.7739 (0.005)

G, N e U= QTL effect provided of Gamma, Normal e Uniform distributions
145 e 725 number of QTLs that controlled the trait
$r_{u\hat{u}}$ = Correlation between estimated and true GBVs

Conclusion

The Bayesian LASSO showed similar performance in terms of genetic parameters estimation and genomic breeding values predictions in spite of the underlying assumptions of genetic structure.

Literature Cited

Cleveland, M.A., Forni, S., Deeb, N., et al. (2010). BMC Proc. 4 (Suppl 1), S6.

de los Campos, G., Naya,H., Gianola, D., et al. (2009). Genetics., 182:375–385.

Legarra, A., Robert-Granié, C., Croiseau, P., et al. (2010). Proceedings of 9th WCGALP, pp. 1–8.

Legarra, A., Robert-Granié, C., Croiseau, P., et al. (2011). Genet Res. 93:77-87.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Genetics, 157:1819–1829.

Misztal, I., Legarra, A., Aguilar, I. (2009). J. Dairy. Sci., 92:4648–4655.

Park, T., Casella, G. (2008). J. Am. Stat. Assoc. 103:681–686.

Pérez, P., de los Campos, G., Crossa, J., et al. (2010). Plant Genome, 3:106–116.

Sargolzaei, M. and F. S. Schenkel. (2009). Bioinformatics, 5: 680-681.