Using LASSO to estimate marker effects for Genomic Selection

Mario Graziano Usai[1], Mike E. Goddard[2], Ben J. Hayes[3]

[1]Settore Genetica e Biotecnologie - DIRPA, AGRIS-Sardegna, Olmedo, Italy

[2]Faculty of Land and Food Resources, University of Melbourne, Australia

[3]Biosciences Research Division, Department of Primary Industries Victoria, Australia

Corresponding author: Mario Graziano Usai. Settore Genetica e Biotecnologie - DIRPA, AGRIS-Sardegna. Loc. Bonassai SS291 km 18.6, 07040 Olmedo (SS), Italy - **Tel. +39 079 387318 - Fax: +39 079 389450 - Email:** graziano.usai@gmail.com

ABSTRACT - Here we suggest a least absolute shrinkage and selection operator (LASSO) approach to estimate the marker effects for genomic selection using the least angle regression (LARS) algorithm, modified to include a cross–validation step to define the best subset of markers to involve in the model. The LASSO-LARS was tested on simulated data which consisted of 5,865 individuals and 6,000 SNPs. The last generations of this dataset were the selection candidates. Using only animals from generations prior to the candidates, three approaches to splitting the population into training and validation sets for cross-validation were evaluated. Furthermore, different sizes of the validation sample were tested. Moreover, BLUP and Bayesian methods were carried out for comparison. The most reliable cross-validation method was the random splitting of overall population with a validation sample size of 50% of the reference population. The accuracy of the GEBVs (correlation with true breeding values) in the candidate population obtained by LASSO-LARS was 0.89 with 156 explanatory SNPs. This value was higher then those obtained by using BLUP and Bayesian methods, which were 0.75 and 0.84 respectively. It was concluded that LASSO-LARS approach is a good alternative way to estimate markers effects for genomic selection.

Key words: Genomic-selection, SNP, LASSO, LARS.

Introduction – Meuwissen *et al.* (2001) proposed a method to estimate breeding values by using a genome-wide dense map of markers, which they termed genomic selection. They used two ways to estimate the marker effects, BLUP and Bayesian approaches, and obtained high levels of accuracy. An alternative method to estimate the SNP effects would be to use of least absolute shrinkage and selection operator (LASSO) approach (Thibshirani, 1996). This operator includes in the model only a subset of explanatory variables, fitting to zero those which do not improve predictability. The main issue with the LASSO approach is how to best choose of the subset of variables, in this case the number of SNPs. In this paper we suggest a LASSO approach to estimate the marker effects for genomic selection using least angle regression (LARS; Efron *et al.*, 2004) algorithm, modified to include a cross–validation step to define the best size of the subset of SNPs. Different approaches for selecting cross- validation sets were compared. The LASSO approach was also compared with the BLUP and Bayesian approaches on the same simulated dataset.

Material and methods – The simulated data came from the XII QTL-MAS Workshop 2008, Uppsala (http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html). This dataset consisted of 5,865 individuals from seven generations. There were 6,000 loci evenly distributed over six chromosomes, with 0.1 cM between markers. The first four generations (4,665 individuals) and the last three generations (1,200 individuals) were the reference and candidate population respectively. There were 48 QTLs distributed on the genome of which 2 with additive effect; 2 with

epistatic effect and 44 with effects randomly sampled from a gamma distribution (Lund *et al.*, 2009). The QTLs were not among the 6,000 loci. The genotypic information consisted of bi-allelic markers (e.g., SNPs).

The allelic substitution effects of the SNPs were estimated by a LASSO approach. It minimizes the residual sum of squares constraining the sum of absolute values of the SNP effects (Tibshirani, 1996). The constraint allows some estimated SNP effects to be exactly zero.

The LASSO problem was solved by using a modified version of the LARS algorithm (Efron *et al.*, 2004). In the classical LARS procedure, the estimates of the effects are obtained in successive iterations, for each iteration the marker (currently out of the model) with the highest absolute correlation between genotypes and current residuals is added to the model. To obtain LASSO solutions, the LARS procedure was modified so that either addition or subtraction of one marker to the model might occur. One marker was subtracted when disagreed the sign of its effect with the sign of the correlation between genotypes and residuals (Efron *et al.*, 2004). Only the markers inside the model had nonzero effects.

The problem of choosing the best constraint values was dealt with, here, as the selection of the best subset size of explanatory SNPs to retain in the final model. Thus a cross-validation approach using random sub-sampling replication (Kohavi, 1995) was performed. In each replication the reference population was randomly split in two samples: training sample (T; to estimate the SNP effects); validation sample (V; to validate the results obtained by T). Then the LARS procedure was carried out on T. For each LARS iteration, the genomic estimated breeding values of the validation sample ($GEBV_V$) were calculated as product between the current vector of the effects estimated on T and the matrix of the genotypes of V. When the correlation between $GEBV_V$ and the phenotypes of V reached the maximum the LARS procedure was stopped and the number of SNPs in the model was retained as best subset size for that replication. We evaluated three different approaches to splitting the data into training and validation sets. First, individuals were allocated by random splitting of overall population (RAN). Second, individuals assigned to V belonged to the last generation of the reference population only (WFAM). Finally, entire families of the last generation were assigned to V (BFAM). Furthermore, different sizes of V (T) were tested, 5% (95%), 10% (90%) and 20% (80%) for each approach. The V size of 50% was tested for RAN only. For each trial (splitting approach x V size) 1,000 replications were performed. For each trial, mean and standard deviation of the SNP subset sizes overall replications were calculated. Then the LARS procedure was performed on the whole reference population. In this case, for each iteration the current vector of effects was used to calculate the GEBVs in the candidate population. Then the accuracy of these GEBVs was calculated as the correlation between GEBVs and true breeding values (TBVs). The regression coefficient of TBVs on GEBVs for each LARS iteration was also determined. The accuracy values corresponding to the number of active SNPs equalling the mean of marker subset sizes of each cross-validation trial were used to compare them.

We compared the accuracy of GEBV using prediction of SNP effects from the LASSO-LARS with GEBV calculated from SNP effects predicted by a BLUP approach and the BayesA methods as described by Meuwissen *et al.* (2001).

Results and conclusions – Table 1 shows means and standard deviation of the markers subset sizes from 1,000 random sub-sampling replications for each cross-validation trial. For each subset size Table 1 shows the corresponding GEBV accuracy in the candidate population. The subset markers size ranged from 184 for BFAM-10% to 220 for WFAM-10%. In general the GEBV were high, with only small difference in accuracy between methods. The GEBV accuracy ranged from 0.8810 to 0.8941 for WFAM-10% and RAN-50%, respectively (Table 1). RAN-50% was the way to best select the subset size of explanatory SNPs to retain in the final model, because it provided the highest level of accuracy and the lowest standard deviation value, hence a higher reliability. Furthermore, the RAN method might be applied even in cases where the structure of the population might be unknown or not homogenous as in simulated data.

Table 1. Means and standard deviation (s.d.) of the best marker subset sizes from 1,000 replications and the corresponding GEBV accuracy {r(TBV,GEBV)} in the candidate population for the three random sampling methods and four validation sample (V) sizes.

V size	RAN		WFAM		BFAM	
	mean (s.d.)	r(TBV,GEBV)	mean (s.d.)	r(TBV,GEBV)	mean (s.d.)	r(TBV,GEBV)
5%	195 (108)	0.8921	207 (111)	0.8907	153 (112)	0.8940
10%	200 (85)	0.8915	220 (96)	0.8810	184 (100)	0.8930
20%	195 (58)	0.8921	216 (62)	0.8878	197 (73)	0.8921
50%	156 (32)	0.8941				

RAN: individual random sampling overall population; WFAM: individual random sampling in the last generation only; BFAM: across families sampling in the last generation only.

The candidate population GEBV accuracy corresponding to the average subset size of RAN-50% sampling was compared with the accuracy obtained by using BLUP and BayesA approaches (Table 2). The accuracy obtained by LASSO-LARS exceeded that of BLUP and BayesA by about 20% and 7% respectively. Table 2 also shows the regression of TBV on GEBV. The differences between estimated values of such coefficient and the target value of 1 were 14.8% higher for LASSO-LARS and 13.2% and 8.5% lower for BLUP and BayesA respectively. Therefore the LASSO-LARS GEBV underestimates the true breeding values to small extent. However, it is important to point out that some of the QTLs had simulated epistatic action, and our LASSO-LARS does not account for this.

Table 2. Accuracy of selection (r), regression coefficient (b) of TBV on GEBV.

	r(TBV,GEBV)	b(TBV,GEBV)
BLUP	0.7477	0.8676
BayesA	0.8359	0.9155
LASSO-LARS	0.8941	1.1481

Our results demonstrate that LASSO-LARS can potentially estimate QTL effects from dense SNP data more accurately than BLUP and BayesA, leading to higher accuracies of GEBV for genomic selection. Unlike BLUP and BayesA, a feature of LASSO-LARS is that some of the SNP effects are set to zero. Given the very large amount of SNP data now available, this could be desirable since it allows the selection of a small subset of the markers which are predictive for a particular trait.

Mario Graziano Usai was funded during his stay at the Department of Primary Industries by the Regione Autonoma della Sardegna program "Master and Back", D.G.R. n. 27/13 and n. 59/34.

REFERENCES – Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32:407-499. **Kohavi**, R., 1995. A study of cross-validation and bootstrap for estimation and model selection. Proc. 14th IJCAI-95, Montreal, Canada, pp. 1137-1143. **Lund,** M.S., Sahana, G., de Koning, D.J., Su, G., Carlborg, Ö., 2009. Comparison of analyses of the QTLMAS XII common dataset I: genomic selection. BMC Proc. (In press). **Meuwissen**, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829. **Tibshirani**, R., 1996. Regression shrinkage and selection via the Lasso. J. Royal. Stat. Soc. B. 58:267-288.