Multiple Shrinkage Methods In GWA And GEBV Prediction: An Application To Direct Gestation Length In US Holstein And Italian Brown Populations

C. Maltecca[*], K. Weigel[†], A. Rossoni [‡]

Introduction

Model selection, or the use of regularization (Yi and Xu (2008)) are often employed in QTL mapping and GEBV prediction when models are oversaturated. Shrinkage is particularly popular since the bias introduced by forcing coefficients to 0 is offset by a reduction in mean square error. Different procedures are common. In the Bayesian-LASSO a double exponential prior is assigned to SNPs variances (de los Campos et al. (2009)). Alternatively, inverted χ^2 priors can be used (Cleveland et al. (2009)). One common limitation of these approaches is the utilization of priors with single mean and scale, while shrinkage toward multiple prior means with unknown scale would be desirable. Methods have been proposed where shrinkage is performed on a predefined set of groups. These methods though, lack in generality since they require prior knowledge of the number of groups. Recently, Maclehose and Dunson (2009) proposed a model that efficiently allows multiple locations shrinkage. In their method a Dirichlet process prior is placed on the mean and scale parameters, which induces clustering around a small set of means with different shrinkage. In this case the number of shrinkage clusters is driven by the data so that no prior assumption on the clusters number is required. The objective of this investigation was the application of multiple shrinkage models in GWA and GEBV prediction for direct gestation length (GL) in the US-Holstein and Italian Brown populations and the comparison of these with other popular models. Differences in gestation length have been reported among these two cattle breeds. While Holstein averages around 280 d, Brown GL has been reported to be around 10 d longer (Norman (2009)) . Despite longer GL though, Brown shows lower calving problems than Holsteins.

Material and methods

Sires in the CDDR (US-HOL) and in the Italian Brown (ITA-BW)populations with both EBVs and genotypes for the 50K ISelect chip were used in the analysis (N = 4743 and N=749, respectively) GWA and GEBV prediction were performed using five different approaches.
Bayes-A (B-A), Bayesian LASSO (B-L), Student-t (S-T) models:

[*]Animal Science Department NCSU. 27695 Raleigh NC, USA
[†]Dairy Science Department UW-Madison, Madison WI, USA
[‡]Italian Brown Breeders Association, Località Ferlina Bussolengo 37012, Italy

All models were considered as two level hierarchical. A flat (1) and a non informative ($\frac{1}{\sigma_e^2}$) prior were assigned to μ and σ_e^2, respectively. The remaining prior structure was: $\beta_j \sim N(0, \sigma_{gj}^2)$ for the jth SNP, $\sigma_{gj}^2 \sim Exp(\sigma_{gj}^2|2/\lambda_j^2)$ for the B-L approach and $\sigma_{gj}^2 \sim Inv-\chi^2(\sigma_{gj}^2|v, s^2)$ for the B-A and S-T approaches. Degrees of freedom ν and scale parameter s^2 for B-A were considered hyper-parameters and were assigned values as in Meuwissen (2001). The S-T model treated ν and s^2 as unknown and assigned a uniform density of $\frac{1}{\nu}$ for the interval (0,1] and a uniform distribution of s for the range $(0, A]$, with A being a large number. The λ parameter in the B-L approach was assigned a gamma prior distribution $Gamma(a, b)$. Values of a and b were set at 0.05 and respectively 1 so that the prior for λ was essentially uniform over a wide range of values. A Gibbs sampling algorithm was implemented to obtain samples from the joint posterior distribution.

Multiple shrinkage with B-LASSO (MS-B-L) or student-t (MS-S-T)specifications:

All previous models performed a single shrinkage for all markers considered. In MS models shrinkage to multiple non-null values was allowed. This was obtained expanding either B-L or S-T models specifications by including a mixture prior with separate prior location, scale (and d.f. for S-T) for each coefficient. To reduce the dimensionality of the model a Dirichlet process prior was employed allowing grouping of markers in a smaller subset of clusters. Following are the prior structure and sampling scheme for the MS-B-L model. Differences with the MS-S-T are similar to what outlined previously and are omitted.

μ, σ_{gj}^2, b_j, and σ_e^2 priors were as in previous section. The remaining prior structure was:

$$\sigma_{gj}^2 \sim Exp(\sigma_{gj}^2|2/\lambda_j^2),\ k_j \sim \sum_1^\infty \pi\delta(k_j|t),\ \pi_t = V_t\prod_{h<t}(1-V_h),\ \mathrm{V}_t \sim Beta(V_t|1,\alpha),$$
$$(\mu_t^*, \lambda_t^*) \sim \begin{cases} \delta(\mu_t^*|0)Gamma(\lambda_t^*|a_0, b_0) & \text{if t} = 1 \\ \delta(\mu_t^*|c, d)Gamma(\lambda_t^*|a_1, b_1) & \text{if t} > 1 \end{cases}$$

Where $\delta(\mu_t^*|0)$indicates a degenerate distribution with mass at 0 π_t is the probability that a coefficient will fall in a cluster of markers with (μ_t^*, λ_t^*) and is constructed through a stick-breaking process. α is a hyper-parameter that govern the clustering of markers,k_j indexes which of the bins the j^{th} coefficient falls into, and *t* represents the number of clusters at a certain iteration with * indicating *p*(number of clusters)< p(number of markers).* a_0, b_0, a_1, b_1 ,*c*, *d* are hyper-parameters set (upon simulation) at 1, 30, 30, 6.5, 6.5, 0, and 4 respectively.

In addition to the general Gibbs sampling scheme for the previous models extra steps were included in the MS implementation:

- Sample σ_{gj}^2 from $InvGauss(\sigma_{gj}^2|\beta_j, \mu_{kj}^*, \lambda_{kj}^*)$
- Sample (μ_t^*, λ_t^*)
  λ_1 from $Gamma(\lambda_1^2|m1 + a_0/\sum_{j:kj=1}\sigma_{gj}^2 + 1/b_0)$
  λ_t from $Gamma(\lambda_t^2|mt + a_1/\sum_{j:kj=1}\sigma_{gj}^2 + 1/b_1)$
  μ_t from $N(\hat{E}_{\mu t}, \hat{V}_{\mu t})$
  with $\hat{E}_{\mu t} = \hat{V}_{\mu t}(c/d + \sum_{j:k_j=t}\beta_j/\sigma_{gj}^2)$ and $\hat{V}_{\mu t} = (1/d + \sum_{j:k_j=t}1/\sigma_{gj}^2)^{-1}$
  V_t $Beta(m+1, p - \sum_{l=1}^t m_l + \alpha)$ and generate $\pi_t = V_t\prod_{h<t}(1-V_h)$
- Update cluster configuration through metropolis step Papaspiliopoulos and Roberts (2008)

The Gibbs sampling algorithms for all methods were implemented in `R` . For each analysis a single chain of 150,000 iterations was run with a burn-in period of 20,000 iterations. Samples

were stored every 30 iterations. Convergence of each chain was assessed both by visual inspection of the trace and the use of estimates of effective sample size for variances obtained through the `coda` package. Inferences on the parameters were made on the average of the posterior samples after burn-in.
GWA step: At each iteration, an approximate LOD (B-LOD) score was calculated for each marker as $2log_{10}\frac{L(FM)}{L(RM)}$ where $L(FM)$ and $L(RM)$ are the posterior likelihoods of full and reduced models, respectively. Significances were obtained from the median of the each SNP B-LOD scores. For this step all sires were retained.
GEBV step: For genomic predictions sires were split in training and a prediction datasets as in VanRaden (2008) for US-HOL and similarly for ITA-BW. GEBVs in the prediction set were calculated from SNP solutions obtained from the training set (38410 and 35622 SNPs for US-HOL and ITA-BW respectively). GEBVs were then compared to EBVs in the prediction set to evaluate performaces of the different models.

Results and discussion

GWA: GWA Results for the most significant SNPs in either population are reported in Table 1. B-LOD scores and estimated effects and, within parentheses, the method yielding the highest score and the heritability for the SNP are reported. The most significant SNP for direct GL in the US-HOL population is on BTA18. B-LOD score for the SNP is 13.4. The same SNP was identified by Cole et al. (2009) as affecting calving ease and stillbirth. The significance of this marker, was not confirmed in the Brown population. Two significant markers were identified on BTA28 in both US-HOL and ITA-BW in proximity of the *Tubulin folding cofactor E* gene, whose mechanism of action relates to post-embryonic development. Significant SNPs for the US-HOL were also found on BTAs 15 and 4 where QTL for gestation length were previously mapped in different populations. In both population MS models produced the highest signal with MS-S-T higher in US-HOL and MS-B-L in ITA-BW, respectively

Table 1: SNPs for direct GL in US-HOL and ITA-BW populations line[a]

		B-LOD (Method)		Effect(h^2)	
Chr	Pos	US-HOL	ITA-BW	US-HOL	ITA-BW
28	6656203	4.6(d)		0.17(.01)	
28	45455730		12.1(e)		0.43(.04)
25	940039	4.4(d)	4.6 (e)	0.15(.01)	0.11(.008)
24	55438536		6.2(e)		0.67(.05)
18	57125868	13.4(d)		0.71(.04)	
15	61154802	4.3(d)		0.21(.008)	
6	4948746		14.1(e)		0.91(.05)
4	104396726	4.3(d)		0.11(.02)	

[a]Methods: *a* B-A, *b* S-T, *c* B-L, *d* MS-S-T, *e* MS-B-L.
Effects are expressed as |*days*|

GEBV prediction: Correlations between GEBV and EBV in the prediction dataset for the

different models investigated are presented in Table 2. All models performed similarly with Bayes-A ranking the worst and MS models ranking the best. A slight advantage of MS methods vs. their respective single shrinkage counterparts was offset by larger models and higher computation time. MS-S-T outperformed MS-B-L by a narrow margin in US-HOL while MS-B-L performed better in ITA-BW.

Table 2: Correlation between EBV and GEBV in the prediction datasets for all models

	US-HOL	ITA-BW
B-A	0.55	0.37
B-L	0.58	0.39
S-T	0.60	0.37
MS-B-L	0.59	0.39
MS-S-T	0.61	0.39

Conclusion

Several significant SNPs were discovered in the two populations investigated. Marginal overlapping suggests that multiple mutations might regulate the length of gestation in the two breeds. Multiple shrinkage methods can represent a viable alternative in both GWA studies and two stages GEBV prediction. This at the cost of larger models and at least presently of higher computation costs.

References

Cleveland, M., Forni, S., Deeb, N., and Maltecca, C. (2009). *BMC proceedings*, Epub ahead of print.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. (2009). *Genetics*, 182:375–385.

Maclehose, R. and Dunson, D. (2009). *Biometrics*, Epub ahead of print.

Norman, H.D., e. a. (2009). *Journal of Dairy Science*, 89:978–988.

Papaspiliopoulos, O. and Roberts, G. (2008). *Biometrika*, 95:169–186.

Yi, N. and Xu, S. (2008). *Genetics*, 179:1045–1055.