# SNP Selection Using Elastic Net, With Application To Genomic Selection

*B. L. Harris*[*] and D. L. Johnson[*]

## Introduction

Genomic information can be used to identify animals that have inherited chromosome segments of greater genetic merit. Single nucleotide polymorphism (SNP) markers cover the genome with high density and are now inexpensive to obtain. Statistical methods for the selection of SNPs to use in genomic selection have been outlined by Meuwissen (2001) and Moser et al. (2009). The selection of relevant SNPs within genomic evaluation experiments can be challenging. In these situations there are a large number of predictors available with a high degree of collinearity. This type of variable selection problem is widely studied in the area of statistical learning. In many statistical learning problems, a major goal is that of selecting the variables that are relevant to achieve good predictions. In the problem of variable selection the aim is to select those functions which are needed to represent the regression function, where the representation is typically given by a linear combination. In genomic selection applications, the number of features, p, in the data is usually much larger that the number of examples, n , in the training set. Zou and Hastie (2005) proposed the elastic net (EN) for variable selection with highly correlated variables. The elastic net utilises the lasso and ridge penalty to select variables while inducing a grouping effect. A regression method exhibits a grouping effect when the magnitudes of coefficients of a group of highly correlated variables tend to be similar in magnitude. When p > n, the ridge penalty further allows elastic net to select more than n variables, whereas lasso selects at most n before it saturates. The elastic net has the interpretation as a stabilised version of the lasso (Zou and Hastie, 2005). In many situations, the EN has shown improvements over lasso (Wang et al., 2006). The purpose of this study was to apply EN to identify SNPs for use in genomic selection.

## Material and methods

**Simulated data with gene dropping.** A founder population was simulated over a period of 1000 generations at an effective population size of 100. The genome consisted of 10 chromosomes each of length 100 cM. Marker loci were equally spaced across the genome with biallelic QTL placed at the midpoint of each marker interval. Starting from a segregating population, mutation rates were set at $0.5 \times 10^{-3}$ for markers and $10^{-5}$ for QTL where mutations switched to the other allele. After 1000 generations there were 20,714 SNP markers with minor allele frequency $\geq 2\%$ and 2071 polymorphic QTL. Real pedigree data was based on the 5,769 bulls born up to 2008 and some bull dams that were genotyped in the LIC genomic selection programme. There was a total of 24,017 individuals in the pedigree, inclusive of the

[*]LIC, Private Bag 3016, Hamilton, New Zealand

ancestors of genotyped individuals. The simulated founder population formed the base population for a gene dropping process through the pedigree. The QTL effect at each polymorphic locus was sampled from a Gamma distribution with shape parameter 0.4. A true breeding value (BV) was generated for each individual based on the QTL effects and a phenotype was then obtained by adding in random error with the same variance such that heritability was 0.5. The test data comprised the 970 genotyped bulls born in 2007 and 2008.

**Holstein Friesian data.** Data containing 2,192 genotyped Holstein Friesian (HF) sires based on 40,546 SNPs from the Illumina BovineSNP50 BeadChip platform was available for analysis. The training and test data sets comprised of 1625 and 567 sires, respectively. All the sires were progeny tested and each sire with at least 70 daughters in their genetic evaluation. The test data set contained the sires from the 4 most recent years of graduation from the progeny test. Protein BV was chosen as the phenotypic record.

**Statistical analyses.** Four methods of statistical analysis were undertaken on both the simulated and HF data, BLUP (random regression), Bayes A and Bayes B (Meuwissen et al., 2001), and Elastic Net (Zou and Hastie, 2005). A fifth method, Fast Bayes B (Meuwissen et al., 2009), was undertaken on the simulated data only. The EN solves the following problem

$$\frac{min}{(\beta_0, \beta) \in R^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i'\beta)^2 + \lambda P_\alpha(\beta)) \right] \tag{1}$$

where

$$P_\alpha(\beta) = (1 - \alpha)\frac{1}{2}||\beta||_{l_2}^2 + \alpha||\beta||_{l_1} \tag{2}$$

is the EN penalty. The EN penalty ($P_\alpha$) is a compromise between the ridge-regression penalty ($\alpha = 0$) and the Lasso penalty ($\alpha = 1$). The parameter $\lambda$ represents the sparsity of the solution in EN. This penalty is particularly useful in the $p \gg N$ situation, or any situation where there are many correlated predictor variables. The EN was solved using pathwise coordinate optimisation (Friedman et al., 2010) which a computationally efficient method for solving this type of convex optimisation problem. The optimal EN model was found by a grid search over the parameters $\alpha$ and $\lambda$. A grid search strategy is outlined by Friedman et al., (2010), however for the data sets used in this study a simple grid search over the intervals $(0, 1]$ and [0,1] were adequate for $\alpha$ and $\lambda$, respectively. The EN penalty is closely related to other nonparametric approaches such as kernel regression utilising linear kernels with squared error loss and support vector regression.

## Results and discussion

The results for the simulation data are given in Table 1. The test data correlations are similar for Bayes A, Bayes B, fast Bayes B and EN, all of which perform marginally better than BLUP. The optimal values for the EN were $\alpha = 0.0225$ and $\lambda = 0.675$. The largest difference among the EN and Bayes methods is the number of SNP effects in the model with the EN having considerably fewer SNP effects in the prediction model. A single EN model required under 3 minutes to compute, where as Bayes A and Bayes B required close to 36 hours for 20,000 samples each and fast Bayes B required 3 minutes. The results for the HF data are given in Table 2. The test data correlations are similar for Bayes A, Bayes B and EN, all of which out perform BLUP. As was the case with the simulated data, the EN prediction model has considerably fewer SNP effects than other methods. A large number of Bayes B estimated SNP effects were of small magnitude in both the simulated and HF data. Combining the Bayes B estimated SNP effects over a window of 10 SNPs would produce estimates closer to the EN method in both magnitude and number. The optimal values for the EN were $\alpha = 0.0088$ and $\lambda = 0.625$. The computational demands of each method were of similar magnitudes to those seen in the simulated data.

**Table 1: Results for the simulated data**

| Traits | BLUP | Bayes A | Bayes B | Fast Bayes B | Elastic Net |
|---|---|---|---|---|---|
| Number of SNP effects[a] | 20,714 | 20,524 | 20,461 | 20,435 | 1,762 |
| Test data correlation | 0.607 | 0.624 | 0.629 | 0.616 | 0.623 |

[a]estimates $> 10^{-6}$ for Bayes A , Bayes B and Fast Bayes B

For the optimal EN models for the simulated data and the HF data, the value of $\alpha$ was close to zero. This indicates that the optimal models are closer in their statistical characteristics to a ridge regression model rather than a Lasso model. The values for $\lambda$ for both data sets are moderate in magnitude. As $\lambda$ approaches 0 the EN model becomes less sparse with an increasing number of SNP effects being fitted. Figure 1 illustrates the simulated QTL effects from the simulated data and SNP estimates from Bayes A, Bayes B and EN for chromosome 1. The results displayed in Figure 1 are indicative of the results found for the remaining 9 chromosomes, as well as the estimated SNP effects from the HF data. The Bayes A and Bayes B models produce a large number of small QTL estimates with relatively few large QTL estimates in contrast to EN which produces a small number of moderate to large sized estimates.

**Table 2: Results for the Holstein Friesian data**

| Traits | BLUP | Bayes A | Bayes B | Elastic Net |
|---|---|---|---|---|
| Number of SNP effects[a] | 40,546 | 39,928 | 39,872 | 3,070 |
| Test data correlation | 0.540 | 0.638 | 0.635 | 0.603 |

[a]estimates $> 10^{-6}$ for Bayes A and Bayes B

## Conclusion

In this study the application of the Bayes A, Bayes B and EN methods for prediction of breeding values provided similar levels of performance in both the simulated and the HF data sets. The EN model used in this study is one of many which belong to a class of models known as generalized ridge-lasso estimators (Daye and Jeng 2009). It possible to include more complicated covariance terms in the second component of the penalty term of the EN model, such as fusion weights. These covariance terms may further enhance the method where large numbers of correlated variables exist in the data. With the advent of higher density SNP chips (e.g. 750k SNP chips) it will become important to identify a subset SNPs that provide useful breeding value predictions for a given phenotype. This would allow the removal of SNPs that are uninformative and just add noise to the predictions. EN provides is a useful tool for this task and has the added benefit of not being computationally demanding.

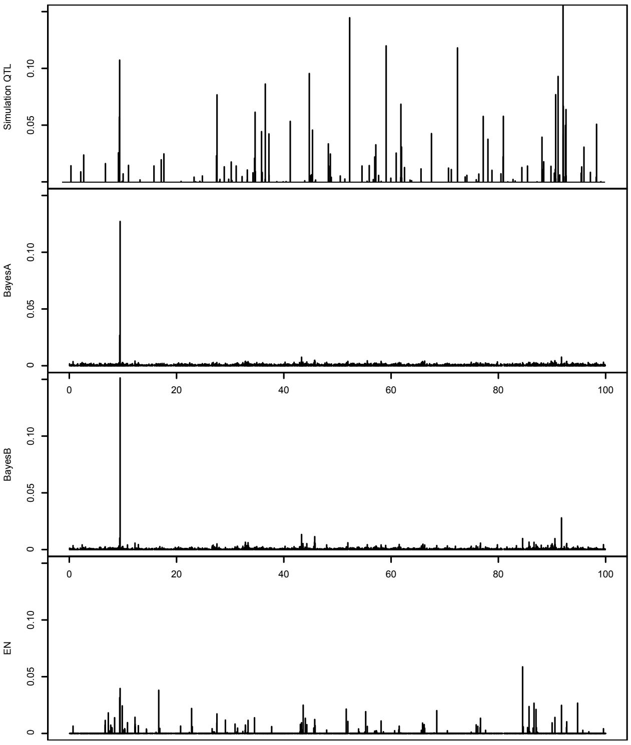

**Figure 1: QTL effects and SNP estimates from Bayes A, Bayes B and elastic net (EN) models for chromosome 1 in the simulated data.**

## References

Daye, Z. J. and Jeng, X. J. (2009). *Computational Statistics and Data Analysis*, 54:1284-1298
Friedman J., Hastie, T. and Tibshirani, R. (2010). *J. Statistical Software*, 33:1-22.
Meuwissen T. H. E., Hayes, B. J., Goddard, M. E. (2001). *Genetics*, 157:1819-1829.
Meuwissen T. H. E., Solberg T. R., Shepherd R. and Wooliams J. A. (2009). *GSE*, 41:2-11.
Moser G., Tier B., Crump R. E., Khatkar M. S. and Raadsma H. W. (2009). *GSE*, 41: 41-56.
Wang, L., J. Zhu, and Zou H. (2006). *Statistica Sinica*, 16:58915.
Zou, H and Hastie, T. (2005). *J. Statist. Soc. B*, 67:301-320.